Simultaneous Tactile-Visual Perception for Learning Multimodal Robot Manipulation

Yuyang Li, Yinghan Chen, Zihang Zhao, Puhao Li, Tengyu Liu, Siyuan Huang, and Yixin Zhu

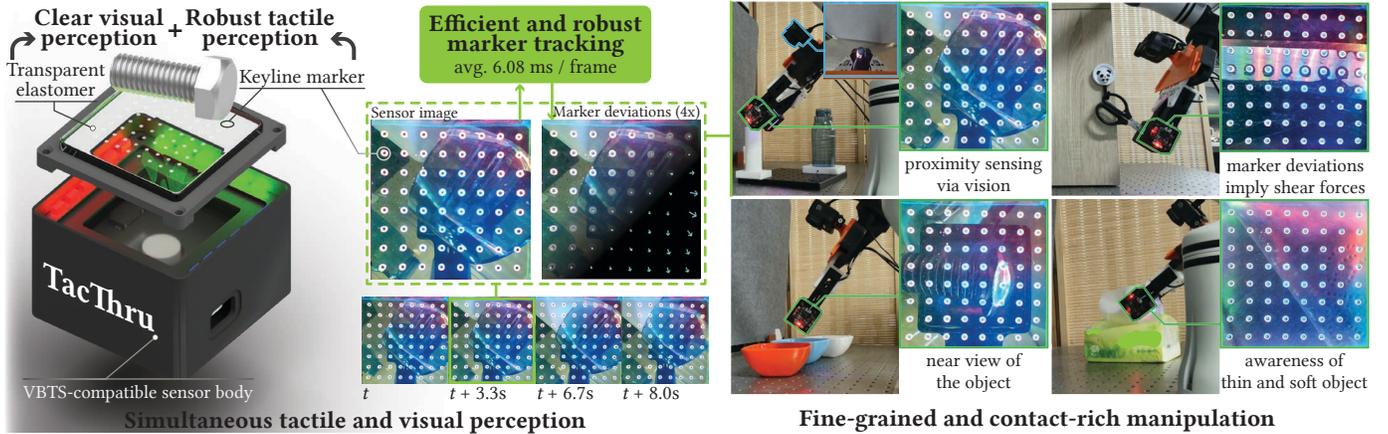**https://tacthru.yuyang.li**

Fig. 1: **Learning multimodal robot manipulation with simultaneous tactile-visual perception. `TacThru`** enables clear visual perception and robust marker tracking via transparent elastomer and keyline markers (left), providing rich multimodal signals for learning manipulation policies. **`TacThru-UMI`** demonstrates efficacy across fine-grained and contact-rich tasks requiring precise multimodal coordination (right).

Abstract—**Robotic manipulation requires both rich multimodal perception and effective learning frameworks to handle complex real-world tasks. See-Through-Skin (STS) sensors, which combine tactile and visual perception, offer promising sensing capabilities, while modern imitation learning provides powerful tools for policy acquisition. However, existing STS designs lack simultaneous multimodal perception and suffer from unreliable tactile tracking. Furthermore, integrating these rich multimodal signals into learning-based manipulation pipelines remains an open challenge. We introduce `TacThru`, an STS sensor enabling *simultaneous* visual perception and *robust* tactile signal extraction, and `TacThru-UMI`, an imitation learning framework that leverages these multimodal signals for manipulation. Our sensor features a fully transparent elastomer, persistent illumination, novel keyline markers, and efficient tracking, while our learning system integrates these signals through a Transformer-based Diffusion Policy. Experiments on five challenging real-world tasks show that `TacThru-UMI` achieves an average success rate of 85.5%, significantly outperforming the baselines of tactile policy (66.3%) and vision-only policy (55.4%). The system excels in critical scenarios, including contact detection with thin and soft objects and precision manipulation requiring multimodal coordination. This work demonstrates that combining simultaneous multimodal perception with modern learning frameworks enables more precise, adaptable robotic manipulation.**

Index Terms—**Manipulation, tactile sensing, proximity sensing, imitation learning**

Yuyang Li and Yinghan Chen contributed equally. Corresponding emails: `yixin.zhu@pku.edu.cn`, `huangsiyuan@ucla.edu`, and `liutengyu@bigai.ai`.

Yuyang Li, Yinghan Chen, Zihang Zhao, and Yixin Zhu are with the Institute for Artificial Intelligence, Peking University.

Yixin Zhu is also with the School of Psychological and Cognitive Sciences, Peking University.

Yuyang Li, Yinghan Chen, Zihang Zhao, and Yixin Zhu are also with the Beijing Key Lab of Behavior and Mental Health, Peking University.

Yuyang Li, Tengyu Liu, and Siyuan Huang are with the Beijing Institute for General AI.

All authors are with the State Key Lab for General AI.

Yixin Zhu is with the Embodied Intelligence Lab, PKU-Wuhan Institute for AI.

Yinghan Chen is with the Department of Computer Science and Technology, University of Cambridge.

I. INTRODUCTION

MANIPULATION requires a comprehensive environmental perception that spans the pre-contact to post-contact phases [1, 2]. Current sensing modalities each address different aspects but exhibit complementary limitations. Vision provides rich global context, but often downgrades during manipulation due to visual occlusions, particularly when precise control is required [3]. Vision-based Tactile Sensors (VBTSs) [4–7] excel in providing high-resolution contact information [3, 4,

8, 9], but offer no information during the crucial pre-contact approach phase and provide only local, sometimes sparse signals [10]. Dedicated proximity sensors partially bridge this pre-contact gap [11–13], but lack the spatial richness and precision needed for fine-grained manipulation.

See-Through-Skin (STS) sensors integrate tactile and visual sensing to address these limitations [10, 14–27]. Adapting VBTS designs, they replace opaque reflective coatings with semi-opaque or transparent alternatives, allowing the embedded camera to "see through the skin" for visual signals (*e.g.*, object appearance, proximity) beside the tactile signals (*e.g.*, contact, force). Recent work has validated their utility in object perception [22, 25, 26], grasping [15, 17, 22, 28], in-hand translation [10], and articulated object manipulation [18, 24].

Despite this progress, three limitations prevent the broader adoption of STS sensors. First, most current designs explicitly switch between tactile and visual modalities through illumination control [23, 24] or movable components [27], preventing *simultaneous* multimodal perception and introducing additional control complexity. Second, tactile markers essential for shear force measurement are difficult to track against *noisy*, *unpredictable* external backgrounds, limiting applications in open environments [14]. Finally, existing STS applications rely primarily on *hand-crafted* controllers, with integration into modern data-driven manipulation pipelines largely unexplored.

To address these issues, we introduce **TacThru**, an STS sensor that enables simultaneous tactile-visual perception through (i) a fully transparent elastomer for clear visual perception, (ii) persistent illumination that eliminates mode switching, (iii) novel keyline markers robust to background variation, and (iv) an efficient tracking algorithm that enables 120Hz operations. This design enables truly *simultaneous* tactile-visual perception while remaining compatible with standard VBTS fabrication [29] with minimal adoption.

We further develop **TacThru-UMI**, integrating **TacThru** into an imitation learning framework based on Universal Manipulation Interface (UMI) [30]. A Transformer-based Diffusion Policy learns to properly attend to the *simultaneously provided* multimodal signals for manipulation control. We evaluate the system on five diverse real-world tasks, including pick-and-place, sorting, and insertion (Fig. 1), where **TacThru** provides persistent environment, object, and contact perception. Our policies reach an average success rate of 85.5%, achieving $1.54\times$ and $1.29\times$ performance compared to our vision (55.4%) and tactile (66.3%) baselines, respectively.

Our contributions include: (i) **TacThru**, a novel STS sensor that enables efficient, robust, simultaneous tactile-visual perception; (ii) **TacThru-UMI**, an imitation learning system with a design compatible with UMI for data collection, processing, and policy deployment; and (iii) a comprehensive experimental validation demonstrating how **TacThru**'s simultaneous multimodal perception enables superior fine-grained and contact-rich manipulation.

II. RELATED WORK

Tactile and STS Sensors: Tactile sensing enables robots to perceive fine-grained contacts during manipulation.

Piezoresistive [31, 32], piezoelectric [33, 34], and conductive-fluid [35] sensors measure dynamic contact forces with compact forms, while Vision-based Tactile Sensor (VBTS) [4, 5, 7, 8, 36, 37] provide high spatial resolution at relatively low cost. However, traditional tactile sensors are inherently limited to in-contact perception. STS sensors address this limitation by integrating tactile and visual perception [14, 18–25], typically replacing opaque coatings of VBTSs with semi-opaque or transparent alternatives [18, 23, 24] for continuous perception before, during, and after contacts. Existing implementations commonly *alternate* between tactile and visual modes using illumination control [18, 19, 22, 23], mechanically actuated elastomers [27], or PDLC films [20, 21]. Other extensions incorporate UV-excited fluorescent markers [19, 25], and stereo vision [23, 25, 28]. Such mode switching often relies on task-specific heuristics, introducing control complexity and potential errors (*e.g.*, millimeter-level error in sensor-object distance estimation [18, 27]; second-level delays during actuation [24, 27]). **TacThru** enables simultaneous tactile-visual perception with a fully transparent elastomer, persistent illumination, and robust keyline markers with efficient tracking.

Multimodal Manipulation with STS Sensors: STS sensors' visual-tactile sensing mitigates the inherent limitations of individual sensing modes: tactile feedback provides precise contact information but remains local and sparse [10]; vision offers a wider spatial context but fails during occlusion or contact. This multimodal approach improves object pose estimation [22, 26, 28], slip detection [15], material recognition [13], in-hand manipulation [10], and visual servoing [18, 22, 24, 27], allowing unified sensor solutions for object localization, approach, and interaction. We further demonstrate STS sensors' utilities in a series of diverse real-world tasks via **TacThru**.

Learning with Multimodal Sensing: Learning-based manipulation policies encode multimodal sensory streams into unified representations or tokens. Visual data processing employs Convolutional Neuron Network (CNN) or Transformer encoders [38–42], while tactile integration strategies vary by signal types: low-dimensional inputs (*e.g.*, piezoelectric contacts) are processed via Multi-Layer Perceptrons (MLPs) [43–45], while VBTSs utilize specialized encoders for high-resolution images [46–49] or extracted features (*e.g.*, contact depth, marker displacement) [50, 51]. Despite these advances, STS sensors' integration into learning frameworks remains largely unexplored [24]. We integrate **TacThru** into an imitation learning framework, **TacThru-UMI**, and demonstrate that simultaneous multimodal perception enables superior performance in fine-grained and contact-rich manipulation.

III. TACTHRU

TacThru achieves simultaneous tactile-visual perception through three design principles: fully transparent elastomer for clear visual access, persistent illumination to eliminate modal switching, and robust keyline markers for reliable tactile tracking. **TacThru** maintains compatibility with the standard VBTS fabrication pipeline, differing primarily in elastomer material to allow easy adoption.

A. *Transparent Elastomer with Persistent Illumination*

Existing STS designs preserve traditional VBTS depth estimation by semi-transparent coatings [20, 21, 23], switchable illumination [23], or movable components [27], but compromise visual clarity and require explicit, task-specific switching mechanisms. We adopt a fully transparent elastomer with persistent LED illumination, trading depth perception for continuous visual perception. The illumination includes a 24-bulb LED array (RGBW for VBTS-compatibility, or white-only for visual clarity) surrounding the elastomer. Diffusion films are applied to both the LEDs and the underlying acrylic plate surfaces to produce uniform, diffuse lighting and suppress interference from specular reflections.

Under this design, contact detection remains available through two mechanisms: light reflection changes at contact interfaces [18, 22, 52] and marker divergence from elastomer deformation [24]. This design explicitly prioritizes global contact state and visual context at the fingertips over fine-grained surface geometry reconstruction, aligning with recent literature identifying these modalities as critical for robust contact-rich manipulation [1, 18, 22, 24, 44, 52–54].

B. *Keyline Markers*

Elastomer transparency creates two marker detection challenges: (i) degraded detectability: conventional solid markers become invisible against matching backgrounds; (ii) noisy detections: environmental objects with blob-like appearances generate false detections. To address the detectability issue, we introduce *keyline markers*: two concentric circles with contrasting colors, ensuring the inner edge remains visible as a detectable "keyline" regardless of background.
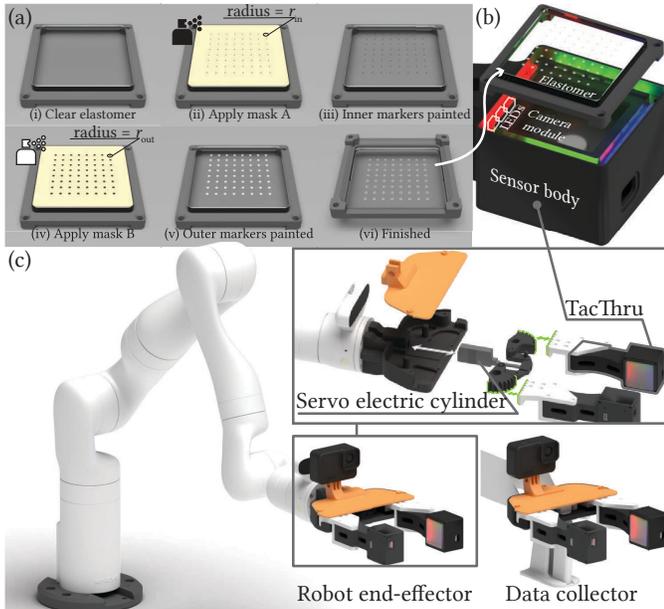

Fig. 2: **TacThru's fabrication and integration in TacThru-UMI.** (a) The keyline markers are fabricated by sequentially spraying inner (black) and outer (white) markers on a transparent elastomer. (b-c) **TacThru-UMI** includes a robot end-effector (left) and a data collector (middle) with identical body and **TacThru** fingers. The end-effector's fingers are actuated by a servo electric cylinder (right).

As shown in Fig. 2ab, the markers are fabricated by sequentially painting inner circles (black, $r_{in} = 0.6\,mm$) and outer circles (white, $r_{out} = 1.0\,mm$) using laser-cut masks. We deploy $N_m = 64$ markers with spacing $d_{marker} = 3.5\,mm$ ($\tilde{d}_{marker} = 51\,px$ in pixel coordinate) on our $40\,mm \times 40\,mm$ elastomer. Design constraints include: camera focus distance that enables both marker detection and visual perception, marker size that balances detectability with minimal visual occlusion, and spacing that exceeds the maximum marker deviation to prevent tracking ambiguity.

C. *Robust and Efficient Marker Tracking*

Although keyline markers solve the fundamental detectability problem, environmental noise and large contact deformations still cause tracking failures. We employ Kalman filtering [55] for tracking each marker's state (position in image) $x_t \in \mathbb{R}^2$ at time t based on its measurement (detected position) $z_t \in \mathbb{R}^2$ following:

$$x_t = A_t x_{t-1} + w_t, \quad z_t = H_t x_t + v_t. \quad (1)$$

We adopt a Random Walk model ($A_t = \mathbb{I}_2$) with direct position observation ($H_t = \mathbb{I}_2$). Process and measurement noises $w_t \sim \mathcal{N}(0, Q), v_t \sim \mathcal{N}(0, R)$ are parameterized by $Q = \sigma_w^2 \mathbb{I}_2, R = \sigma_v^2 \mathbb{I}_2$. Computationally, at timestep t, we perform the standard Kalman Filter steps: **(i) Predict** the prior state estimate $\tilde{x}_t$ and covariance $\tilde{P}_t$ (Eqs. (2) and (3)). **(ii) Update** the posterior state $\hat{x}_t$ and covariance $\hat{P}_t$ based on measurement z_t and computed Kalman Gain K_t (Eqs. (4) to (6)). $\hat{x}_0, \hat{P}_0 = \epsilon \mathbb{I}_2$ are initialized based on sensor model with low uncertainty (*e.g.*, $\epsilon = 10^{-3}$).

$$\tilde{x}_t = A_t \hat{x}_{t-1} \qquad\qquad = \hat{x}_{t-1} \qquad (2)$$
$$\tilde{P}_t = A_t \hat{P}_{t-1} A_t^T + Q \qquad = \hat{P}_{t-1} + Q \qquad (3)$$
$$K_t = \tilde{P}_t H_t^T (H_t \tilde{P}_t H_t^T + R)^{-1} \quad = \tilde{P}_t (\tilde{P}_t + R)^{-1} \quad (4)$$
$$\hat{x}_t = \tilde{x}_t + K_t (z_t - H_t \tilde{x}_t) \quad = \tilde{x}_t + K_t (z_t - \tilde{x}_t) \quad (5)$$
$$\hat{P}_t = (\mathbb{I}_2 - K_t H_t) \tilde{P}_t \qquad = (\mathbb{I}_2 - K_t) \tilde{P}_t \qquad (6)$$

Measurement acquisition z_t (marker's detected position) from sensor image I_t involves:

a) Grayscale: : Conver the image to a grayscale iamge: $I_{gs,t} \leftarrow \texttt{Grayscale}(I)$.

b) Thresholding: : Set pixels to binary based on a threshold τ_i to reveal markers edges and suppress low-contrast background edges: $I_{bin,t}(u,v) \leftarrow \mathbb{1}(I_{gs,t}^{(u,v)} > \tau_i)$, where (u,v) is pixel coordinate and $\mathbb{1}(\cdot)$ is an indicator function .

c) Blob Detection: : Extract blob detections $Z_t : \{z\} = \texttt{Blob}(I_{bin,t})$ with OpenCV *SimpleBlobDetector*.

d) Data association:: Match the nearest blob for each marker as measurement: $z_t = \arg\min_{z \in Z_t} \|z - \hat{x}_{t-1}\|$, which rejects potential environmental noises.

In extreme cases (*e.g.*, catastrophic elastomer deformation and ambient lights), a few markers may become temporarily undetectable, and z_t might be mis-matched to surrounding blobs. We enforce a marker motion continuity constraints: if the inter-frame displacement $\|z_t - \hat{x}_{t-1}\| > \tau_z$, we reject z_t by setting $K_t = 0$ instead of Eq. (4), then the state estimate holds $\hat{x}_t = \hat{x}_{t-1}$ and the covariance accumulates ($\hat{P}_t = \hat{P}_{t-1} + Q \geq \hat{P}_{t-1}$). τ_z should be selected to upper bound both the largest

possible inter-frame marker displacement and half the marker spacing $\frac{\tilde{d}_{\mathrm{marker}}}{2}$.

To determine the parameters, we first calibrate $\sigma_v = 0.42$ px using sensor recordings with no marker motions, and empirically tune the noise ratio $\eta = \frac{\sigma_v}{\sigma_w} \approx 4$ (thus $\sigma_w = 0.11$ px) to balance tracking responsiveness with stability. $\tau_i = 0.784$ (200 in `UINT8`) and $\tau_z = 10$px are determined empirically.

D. Evaluation on Marker Tracking

We evaluate the proposed marker tracking algorithm by comparing keyline and solid marker designs. Fig. 3a shows two `TacThru` sensors mounted on the UMI data collector [30], one with keyline markers and one with solid black markers. Fig. 3b illustrates the challenge: solid markers become invisible against black backgrounds; keyline markers remain detectable but are subject to environmental noise.

We collect 8 trajectories (1628 frames per sensor), grasping a plastic bottle with complex black-and-white text. `TacThru` positions are swapped for half of the trajectories to ensure fair exposure to environmental conditions on both sides, and the bottle is rotated $90°$ between collections, creating marker overlays with various areas around the bottle. We quantitatively compare three tracking approaches:

- **Solid:** Solid markers with standard blob detection.
- **Keyline:** Keyline markers with standard blob detection, matching detected blobs to nearest initial positions.
- **Keyline, filtered:** Keyline markers as above, with our Kalman filtering strategy (Sec. III-C).

Fig. 3c presents quantitative results. Solid markers exhibit frequent missed detections, leading to incomplete tactile information. Keyline marker detection reports more than $N_m = 64$ markers due to environmental noises (orange dashed box). Our filtering strategy rejects these false positives, maintaining persistent tracking of all markers with an average latency of
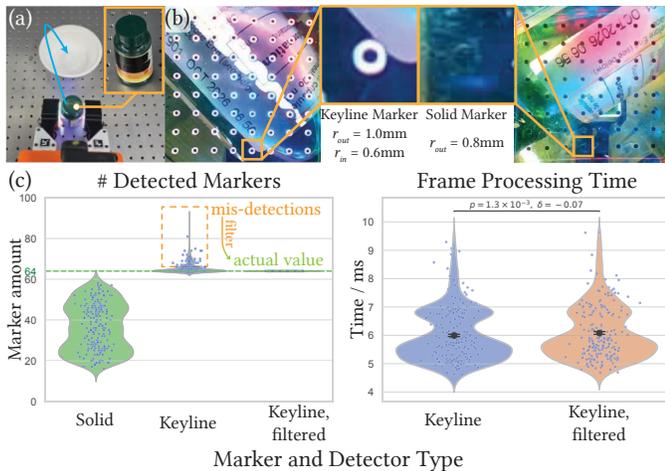
6.08 ms with low CPU overhead. This efficiency supports timely updates on the latest sensor observations, reducing temporal misalignment with other sensory modalities (*e.g.*, vision and proprioception in Sec. IV) and saving time budget to enable high-frequency (*e.g.*, 120 Hz) operation.
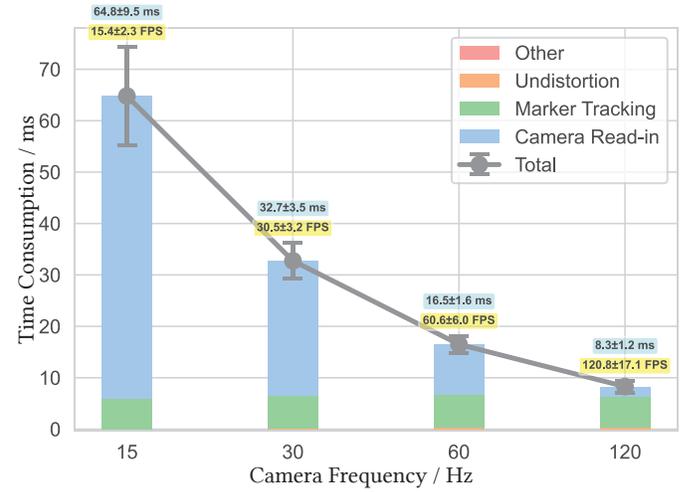

Fig. 4: Time consumptions of `TacThru`'s operation under various frequencies. `TacThru` is tested in 15, 30, 60, and 120Hz, and the time consumption of each computational step is reported. Texts above error bars: iteration time consumption (blue) and equivalent FPS (yellow). The small overshoots in FPS are due to the actual on-board oscillator frequency and the UVC camera's buffering strategy.

We further measure the time consumption of each computation step during the sensor loop in Fig. 4, evaluated on a workstation with an AMD Ryzen 9 5950X CPU. The computational load of undistortion and marker tracking remains consistently low, and the process is mainly blocked at the image read-in step (`VideoCapture.read()` or `.retrieve()`) due to the UVC protocol and OpenCV behaviors. This confirms `TacThru`'s practical capability to operate at a high frequency.

IV. LEARNING MANIPULATION WITH `TACTHRU-UMI`

Although previous studies show invaluable applications of STS sensors with task-specific controllers [10, 14–20, 22, 23, 25–27], we apply `TacThru` to robotic manipulation based on imitation-learning, using its simultaneous tactile-visual perception for fine-grained and contact-rich tasks. We develop the `TacThru-UMI` imitation learning framework, extending the UMI [30] and the Diffusion Policy [56] with multi-modal tactile-visual observations. Evaluations across multiple tasks highlight the sensor's multimodal perception capabilities, providing close-up visual observation of environments and objects along with tactile feedback, enabling more precise manipulation than traditional single-modality approaches.

A. System Setup

Data collector: The `TacThru-UMI` data collector adapts the UMI [30] design, replacing standard fingers with STS sensors on extended linkages (Fig. 2c). `TacThru`s are connected to a computer via USB to stream real-time images during demonstrations and policy inference.

Fig. 3: **Keyline marker design and filtering enable robust tracking.** (a) Evaluation setup compares two marker types (keyline *vs.* solid markers) during bottle grasping. (b) `TacThru` view comparison shows that keyline markers (left) remain distinct against complex backgrounds, while solid markers (right) become invisible. (c) Quantitative results demonstrate our filtered keyline method achieves stable tracking of all 64 markers while keeping efficiency (6.08 ms processing time), and the filtering step rejects false positives.

Robot end-effector: Although the UMI end-effector setup is compatible with multiple parallel grippers (*e.g.* the Panda Hand), we design a low-cost gripper that directly mirrors the data collector's body and **TacThru** fingers (Fig. 2c), with the finger width controlled by an Inspire LAS30-021D servo-electric cylinder ($\sim \$280$).

B. Data Collection and Processing

Our pipeline extends UMI [30] with **TacThru** (Fig. 2) and replaces the SLAM-based pose tracking with an HTC Vive Tracker for higher tracking success rates. While **TacThru** supports 120Hz operation with marker tracking, we use a sufficient and more widely-adopted [37, 43] frequency of 30FPS. All data streams (wrist camera, tactile sensors, and proprioception) are synchronized to wrist-camera timestamps and stored in Zarr format for efficient training access.

C. Policy Learning and Inference

We employ a Transformer-based Diffusion Policy [56, 57] to learn mappings from multimodal observations to robot actions, allowing dynamic attention across simultaneously provided visual, tactile, and proprioceptive signals.

At timestep t, the observations include wrist-camera frames $\mathbf{I}_w^t := \{I_w^i\}_{i=t-n_w^{\mathrm{obs}}}^{t-1}$, sensor frames $\mathbf{I}_s^t := \{I_s^i\}_{i=t-n_s^{\mathrm{obs}}}^{t-1}$, marker deviations $\Delta \mathbf{x}^t := \{\Delta x^{i,j}, j=1,\ldots,N_m\}_{i=t-n_s^{\mathrm{obs}}}^{t-1}$, and proprioception $\mathbf{s}^t := \{s^i\}_{i=t-n_p^{\mathrm{obs}}}^{t-1}$ containing the pose of the end-effector and the width of the gripper in relative coordinates [30]. Visual observations are encoded using DINOv2 [58]: ViT-B for wrist cameras and ViT-S for **TacThru** frames (14×14 patch size). Marker deviations and proprioception use dedicated MLPs. Each modality receives learnable

embeddings (z_w, z_s, z_x, z_p) for transformer distinguishability:

$$\mathbf{z}_w = \left\{ \mathrm{DINO}_w(I) + z_w | I \in \mathbf{I}_w^t \right\}, \tag{7}$$

$$\mathbf{z}_s = \left\{ \mathrm{DINO}_s(I) + z_s | I \in \mathbf{I}_w^t \right\}, \tag{8}$$

$$\mathbf{z}_x = \left\{ \mathrm{MLP}_x(\Delta x) + z_x | \Delta x \in \Delta \mathbf{x}^t \right\}, \tag{9}$$

$$\mathbf{z}_p = \left\{ \mathrm{MLP}_p(s) + z_p | s \in \mathbf{s}^t \right\}. \tag{10}$$

Concatenated tokens with positional embeddings condition the diffusion policy π_θ, which denoises Gaussian samples into action chunks [59] $\mathbf{a} = \{a^i\}_{i=t}^{t+T_a-1} \sim \pi_\theta(\mathbf{a}|\mathbf{z}_w, \mathbf{z}_s, \mathbf{z}_x, \mathbf{z}_p)$. Each action a^i includes relative end-effector pose and gripper width targets. During execution, the first L_a actions ($L_a \leqslant T_a$) are sent to the robot controller for Cartesian space servoing.

V. EXPERIMENTS

We evaluated **TacThru-UMI** across five manipulation tasks spanning pick-and-place, sorting, and insertion scenarios. These tasks systematically assess different sensing modalities: tactile information (contact events and shear forces), visual perception (object and environment observation), and their simultaneous operation. Our key finding is that **TacThru**'s multimodal feedback enriches perception of object appearance, state, and the environment throughout the entire manipulation process, even before contact or when handling extremely thin and soft objects. This allows policies to leverage detailed environmental cues for enhanced fine-grained and contact-rich manipulation while maintaining inference efficiency.

A. Task Settings

Fig. 6 illustrates the experimental setups. Each column shows a task with the object to manipulate (first row), wrist-camera view during demonstration (second row), and the corresponding **TacThru** and VBTS images (third row).

PickBottle: A bottle and a bowl are placed randomly. The robot must grasp the bottle and place it in the bowl. This basic pick-and-place task validates **TacThru-UMI**'s effectiveness of imitation learning and real-world inference.

PullTissue: A tissue pack is placed randomly, and the robot must grasp and fully extract a single tissue. Standard tactile sensors struggle to detect contact with thin, soft paper, making the visual modality of STS crucial for this task.

SortBolt: One of three M12×25 bolts is placed between fingertips. The robot must grasp the bolt and place it in the corresponding bowl. Bolts vary in head shape (A: button head; B, C: socket head) and color (A, C: black; B: silver). The global wrist-camera view cannot distinguish these small bolts, while traditional tactile sensing cannot separate geometrically identical but differently colored bolts. **TacThru** allows for distinguishing both color and shape through STS perception.

HangScissors: The scissors are placed between the fingertips; the robot must grasp and hang them on a hook [43]. This requires tactile feedback to distinguish successful hanging (triggers gripper release) from missed attempts (triggers retry).

InsertCap: The robot must grasp a bottle cap and insert it onto a mount, which requires millimeter-level precision alignment. Although insertion tasks are typically based on tactile signals [50, 51], **TacThru** enables direct visual servoing for the cap-mount alignment.
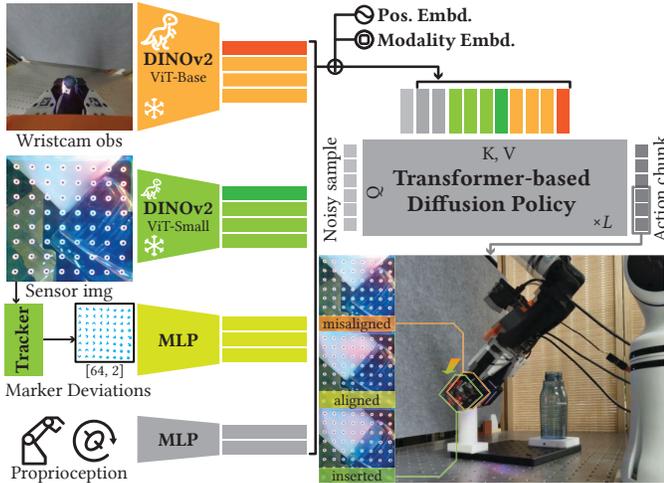

Fig. 5: **Policy architecture.** Multimodal observations (wrist camera images, sensor images, marker deviations, and proprioception) are encoded as tokens, and used to condition a Transformer-based diffusion policy that denoises Gaussian noises into robot action chunks.
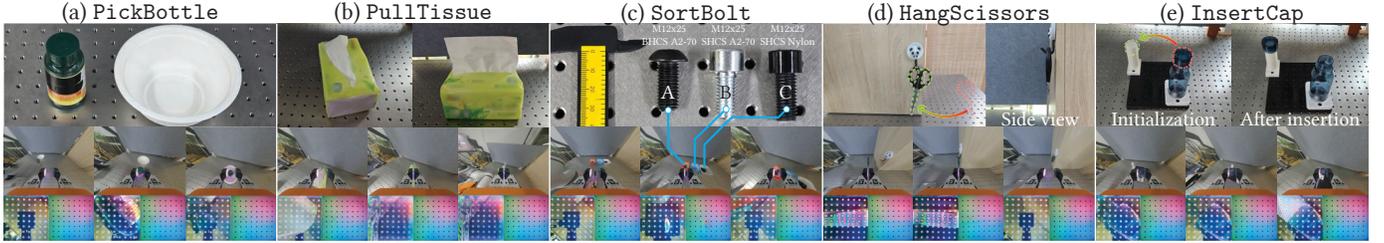
Fig. 6: **Task demonstrations across five manipulation scenarios.** (a) `PickBottle`: basic pick-and-place, (b) `PullTissue`: thin-and-soft object manipulation, (c) `SortBolt`: visual discrimination, (d) `HangScissors`: tactile discrimination, (e) `InsertCap`: multimodal fusion. **Top:** Initial object configurations. **Middle:** Corresponding wrist-camera views. **Bottom:** Corresponding **TacThru** and GelSight images.

B. Experimental Setup

Setup and policy variants: For fair comparison, we equip the gripper with **TacThru** on one finger and a GelSight-type sensor on the other, ensuring identical training trajectories across modality variants. We collect 62–147 demonstrations per task (Fig. 7) and train four policy variants that form ablations and comparisons. Of note, all policies include a wrist camera ($\mathbf{I}_w^t$) and proprioception ($\mathbf{s}^t$) as the base input.

- **TT–M**: **TacThru** images ($\mathbf{I}_s^t$) and marker deviations ($\Delta \mathbf{x}^t$).
- **TT**: **TacThru** images only ($\mathbf{I}_s^t$); markers are visible but not explicitly tracked and provided (as an ablation).
- **GS–M**: GelSight images (rectified by the idle image for isolating contact) and marker deviations (tactile baseline).
- **Wrist**: Wrist camera only ($\mathbf{I}_w^t$) (vision-only baseline).

Training and evaluation: We use observation horizons $n_w^{\mathrm{obs}} = 1, n_s^{\mathrm{obs}} = 1, n_p^{\mathrm{obs}} = 2$, predict $T_a = 16$ action steps, and execute the steps 3-8. Each policy is trained for 150 epochs using the AdamW optimizer [60] (learning rate = 3×10^{-4}, one-cycle scheduler). We evaluate each checkpoint with 20 rollouts (24 for **SortBolt**; 8 for each bolt) with randomized initialization that matched the data-collection distribution.

C. Results

Pick-and-place: In **PickBottle** (Fig. 8a), all policies achieve near-perfect success ($\geqslant 95\%$), validating that **TacThru–UMI** effectively integrates multimodal signals without compromising basic manipulation. This establishes a baseline for specialized sensing in more challenging tasks.
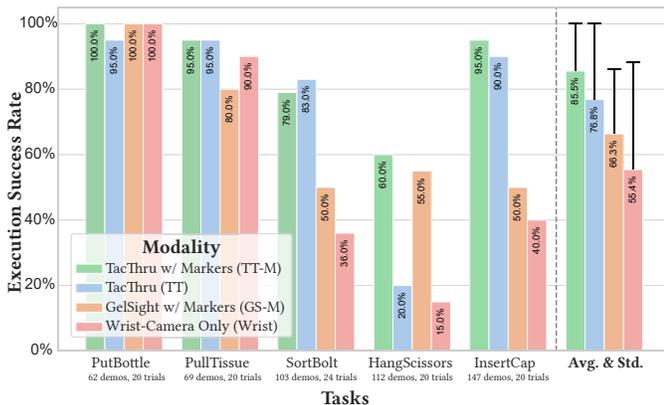

Fig. 7: **Quantitative results across manipulation tasks and sensing modalities. TT–M**: **TacThru** with markers, **TT**: **TacThru** images, **GS–M**: GelSight with markers, and **Wrist**: vision-only.

Thin-and-soft object perception: **PullTissue** reveals the limitations of conventional tactile sensing: thin-and-soft exert minor pressure that is hardly detected. **TacThru** overcomes this through direct visual observation of the inter-finger workspace, providing continuous feedback on tissue position. When tissue slippage (Fig. 8b) occurs due to insufficient grip force, while VBTS (**GS–M**) and vision-only (**Wrist**) policies fail to notice this accident, **TacThru** immediately detects the displacement and triggers corrective retry.

Visual discrimination: **SortBolt** (Fig. 8c) demonstrates **TacThru**'s fine-grained visual discrimination of the object. The small M12×25 bolts present significant challenges: wrist cameras cannot resolve geometric details at mounted distances, while identical bolt shapes make purely tactile discrimination impossible. **TacThru**'s close-proximity view captures fine geometric features and subtle color differences. Fig. 9b provides quantitative evidence through t-SNE visualization of DINOv2 **CLS** embeddings: **TacThru** tokens are clearly separated, while VBTS embeddings are more overlapped, which are aligned with **TacThru**'s $\approx 80\%$ accuracy (**TT–M**, **TT**) and **GS–M**'s obfuscation of bolts B and C (Fig. 9a).

Tactile discrimination: **HangScissors** (Fig. 8d) exemplifies scenarios where visual observation alone cannot determine task completion: the wrist camera cannot reliably identify scissors-handle engagement due to 2D perception and occlusion. Marker displacements (**TT–M**, **GS–M**) provide tactile evidence of successful hang, vital to the precise timing of the gripper release, preventing premature release or continued downward forces. This justifies the rationale of explicitly extracting the marker deviations of tactile signals.

Multimodal fusion: **InsertCap** (Fig. 8e-f) presents the unique advantage of **TacThru**: simultaneous access to visual and tactile information enables the selection of adaptive strategies. When the cap-mount interface remains visible (15 / 20), the policy employs visual servoing, directly aligning visual features for precise insertion (Fig. 8e). However, when the grasp degrades this visual perception (5 / 20), the policy seamlessly returns to tactile-based insertion, using marker displacement patterns to detect contact and guide alignment (Fig. 8f). This adaptive behavior is enabled by (i) the **TacThru**'s *simultaneous* visual-tactile perception, and (ii) Transformer-based policy's capability to properly attend to multimodal signals, acquired from the demonstration data without explicit rule-based programming.
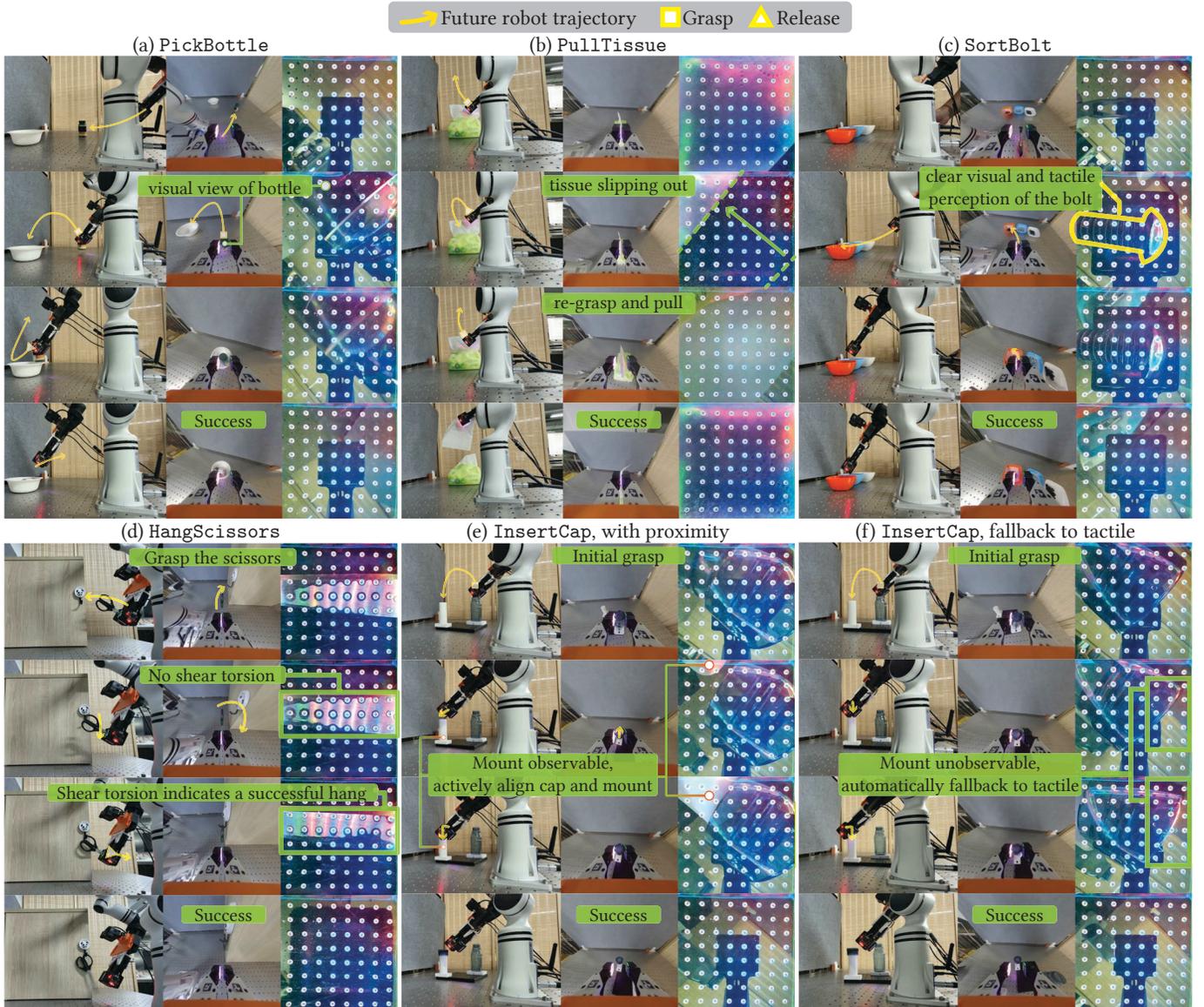
Fig. 8: **Representative policy rollouts.** Each column (a-f) shows the temporal progression of a task execution (top down), with synchronized views: third-person view (left), wrist camera (center), and **TacThru** with marker deviations overlaid (right; 4× magnified for visibility). Annotations highlight key manipulation phases and sensing feedback. See supplementary video for additional rollouts across all settings.

D. Discussions

Our experimental results reveal three key insights on tactile-visual perception for learning multimodal robot manipulation.

Encoding TacThru images with DINOv2: Despite the domain gap introduced by markers and the elastomer, we find DINOv2 effective in capturing meaningful features from **TacThru** images for downstream policies. This performance stems from DINOv2's strong generalization capability, thanks to its discriminative self-supervision training on a large, diverse, and curated dataset (LVD-142M) [58], facilitating zero-shot transfer to unseen image domains [58, 61–63] in practice. We further validate this effectiveness by performing Principal Component Analysis (PCA) on DINOv2 patch tokens extracted from our demonstration data (Fig. 10), covering both contact and non-contact scenarios. By mapping the first three principal components' coefficients to RGB channels [58], we visualize the token similarities of the patch tokens. The result-

ing heatmaps demonstrate that DINOv2 distinctively separates markers, manipulated objects, and background elements (*e.g.*, the opposing finger), despite minor noises. This confirms that DINOv2 is capable of providing clear, structured cues for downstream policy learning.

Adaptive multimodal strategies: Policies trained with **TacThru** naturally learn to weight sensing modalities based on the current context, as demonstrated in **InsertCap**, where the same policy employs vision-based alignment when visible and tactile-based insertion when occluded. This adaptive behavior emerges without explicit programming, suggesting fundamental advantages of simultaneous over sequential sensing.

Overcoming conventional tactile limitations: **TacThru** uniquely handles scenarios where traditional tactile sensors fail. *E.g.*, thin objects such as tissue generate insufficient forces for tactile sensors but remain visually observable through **TacThru**. This expands the range of

(a)

TacThru w/ Markers (TT-M)	in A	in B	in C	Out
A	87.5%	0.0%	0.0%	12.5%
B	12.5%	75.0%	0.0%	12.5%
C	12.5%	0.0%	75.0%	12.5%

TacThru w/o Markers (TT)	in A	in B	in C	Out
A	87.5%	0.0%	0.0%	12.5%
B	0.0%	62.5%	12.5%	25.0%
C	0.0%	0.0%	100.0%	0.0%

GelSight w/ Markers (GS-M)	in A	in B	in C	Out
A	50.0%	50.0%	0.0%	0.0%
B	0.0%	0.0%	100.0%	0.0%
C	0.0%	0.0%	100.0%	0.0%

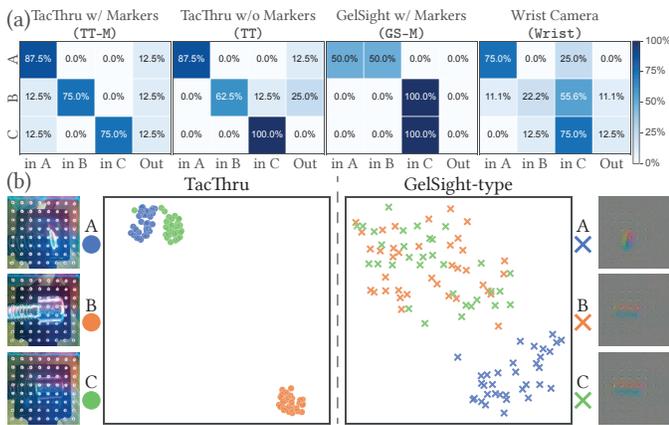| Wrist Camera (Wrist)	in A	in B	in C	Out
A	75.0%	0.0%	25.0%	0.0%
B	11.1%	22.2%	55.6%	11.1%
C	0.0%	12.5%	75.0%	12.5%

(b) TacThru | GelSight-type

Fig. 9: **Bolt sorting performance analysis.** (a) Confusion matrices of placement accuracies. Rows: ground-truth bolt type; columns: executed placement ("Out": dropping outside). **TacThru** policies (**TT−M**, **TT**) achieve high performance, while **GS−M** confuses geometrically identical bolts B and C. (b) t-SNE visualization of sensor images' DINOv2 CLS token. **TacThru** produces clearly separated clusters, while GelSight embeddings for bolts B and C overlap.

manipulable objects for tactile robotics.

Practical deployment viability: Despite significant domain differences due to transparent elastomer, marker overlays, and contact deformations, standard pre-trained visual encoders prove sufficient for robust policy learning. This finding substantially reduces implementation barriers and suggests that **TacThru** can be integrated into existing vision-based manipulation pipelines with minimal modification.

Sensor Failure Modes: While **TacThru** robustly handles nominal contact forces, it shares specific failure modes with other elastomer-based sensors, being susceptible to elastomer delamination under extreme loads or sharp indentation. Integrating reinforced surface materials (*e.g.*, stretchable protective layers [64]) represents a viable solution for enhancing durability. Besides, extremely high-intensity ambient illumination can interfere with the camera's auto-exposure, reducing marker contrast and causing missed detection. While it can be mitigated by manually decreasing the thresholding parameter τ, we identify the adaptive exposure control as a valuable feature for future research.

VI. Conclusions

We introduce **TacThru**, an STS tactile sensor that provides simultaneous tactile and visual perception through transparent elastomer, persistent illumination, and keyline marker tracking. Integrated within our **TacThru−UMI** imitation learning platform, it demonstrates superior performance across various manipulation tasks, addressing the fundamental limitations of existing tactile sensing while maintaining compatibility with standard vision pipelines.

The accessible design and open-source **TacThru−UMI** platform position this work as a practical enhancement for the manipulation research community. Future directions include large-scale data collection combined with synthetic tactile simulation [50, 51, 65] to support pre-training of specialized encoders, and exploration of complex dexterous tasks that fully leverage **TacThru**'s simultaneous sensing capabilities.

References

[1] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, p. eaat8414, 2019.

[2] Y. Zhu, T. Gao, L. Fan, S. Huang, M. Edmonds, H. Liu, F. Gao, C. Zhang, S. Qi, Y. N. Wu, *et al.*, "Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense," *Engineering*, vol. 6, no. 3, pp. 310–345, 2020.

[3] Z. Zhao, Y. Li, W. Li, Z. Qi, L. Ruan, Y. Zhu, and K. Althoefer, "Tac-Man: Tactile-informed prior-free manipulation of articulated objects," *Transactions on Robotics (T-RO)*, vol. 41, pp. 538–557, 2025.

[4] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[5] C. Lin, H. Zhang, J. Xu, L. Wu, and H. Xu, "9dtact: A compact vision-based tactile sensor for accurate 3d shape reconstruction and generalizable 6d force estimation," *IEEE Robotics and Automation Letters (RA-L)*, vol. 9, no. 2, pp. 923–930, 2023.

[6] B. Ward-Cherrier, N. Pestell, L. Cramphorn, B. Winstone, M. E. Giannaccini, J. Rossiter, and N. F. Lepora, "The tactip family: Soft optical tactile sensors with 3d-printed biomimetic morphologies," *Soft Robotics*, vol. 5, no. 2, pp. 216–227, 2018.

[7] W. Li, Z. Zhao, L. Cui, W. Zhang, H. Liu, L.-A. Li, and Y. Zhu, "Minitac: An ultra-compact 8 mm vision-based tactile sensor for enhanced palpation in robot-assisted minimally invasive surgery," *IEEE Robotics and Automation Letters (RA-L)*, vol. 9, no. 12, pp. 11170–11177, 2024.

[8] Z. Zhao, W. Li, Y. Li, T. Liu, B. Li, M. Wang, K. Du, H. Liu, Y. Zhu, Q. Wang, *et al.*, "Embedding high-resolution touch across robotic hands enables adaptive human-like grasping," *Nature Machine Intelligence*, pp. 1–12, 2025.

[9] Z. Zhao, Z. Qi, Y. Li, L. Cui, Z. Han, L. Ruan, and Y. Zhu, "Tacman-turbo: Proactive tactile control for robust and efficient articulated object manipulation," *arXiv preprint arXiv:2508.02204*, 2025.

[10] P. Lancaster, P. Gyawali, C. Mavrogiannis, S. S. Srinivasa, and J. R. Smith, "Optical proximity sensing for pose estimation during in-hand manipulation," in *International Conference on Intelligent Robots and Systems (IROS)*, 2022.

[11] L.-T. Jiang and J. R. Smith, "Seashell effect pretouch sensing for robotic grasping.," in *International Conference on Robotics and Automation (ICRA)*, 2012.

[12] S. E. Navarro, S. Mühlbacher-Karrer, H. Alagi, H. Zangl, K. Koyama, B. Hein, C. Duriez, and J. R. Smith, "Proximity perception in human-centered robotics: A survey on sensing systems and applications," *Transactions on Robotics (T-RO)*, vol. 38, no. 3, pp. 1599–1620, 2021.

[13] C. Fang, D. Wang, D. Song, and J. Zou, "Toward fingertip non-contact material recognition and near-distance ranging for robotic grasping," in *International Conference on Robotics and Automation (ICRA)*, 2019.

[14] A. Yamaguchi and C. G. Atkeson, "Combining finger vision and optical tactile sensing: Reducing and handling errors while cutting vegetables," in *International Conference on Humanoid Robots (Humanoids)*, 2016.

[15] A. Yamaguchi and C. G. Atkeson, "Implementing tactile behaviors using fingervision," in *International Conference on Humanoid Robotics (Humanoids)*, 2017.

[16] R. Patel, R. Cox, and N. Correll, "Integrated proximity, contact and force sensing using elastomer-embedded commodity proximity sensors," *Autonomous Robots*, vol. 42, no. 7, pp. 1443–1458, 2018.

[17] P. E. Lancaster, J. R. Smith, and S. S. Srinivasa, "Improved proximity, contact, and force sensing via optimization of elastomer-air interface geometry," in *International Conference on Robotics and Automation (ICRA)*, 2019.

[18] F. R. Hogan, J.-F. Tremblay, B. H. Baghi, M. Jenkin, K. Siddiqi, and G. Dudek, "Finger-sts: Combined proximity and tactile sensing for robotic manipulation," *IEEE Robotics and Automation Letters (RA-L)*, vol. 7, no. 4, pp. 10865–10872, 2022.

[19] Q. Wang, Y. Du, and M. Y. Wang, "Spectac: A visual-tactile dual-modality sensor using uv illumination," in *International Conference on Robotics and Automation (ICRA)*, 2022.

[20] Q. K. Luu, D. Q. Nguyen, N. H. Nguyen, and V. A. Ho, "Soft robotic link with controllable transparency for vision-based tactile and proximity sensing," in *International Conference on Soft Robotics (RoboSoft)*, 2023.

[21] Q. K. Luu, D. Q. Nguyen, N. H. Nguyen, N. P. Dam, and V. A. Ho, "Vision-based proximity and tactile sensing for robot arms: Design, perception, and control," *Transactions on Robotics (T-RO)*, vol. 41, pp. 5000–5019, 2025.

[22] S. Athar, G. Patel, Z. Xu, Q. Qiu, and Y. She, "Vistac: Toward a unified multimodal sensing finger for robotic manipulation," *IEEE Sensors Journal*, vol. 23, no. 20, pp. 25440–25450, 2023.
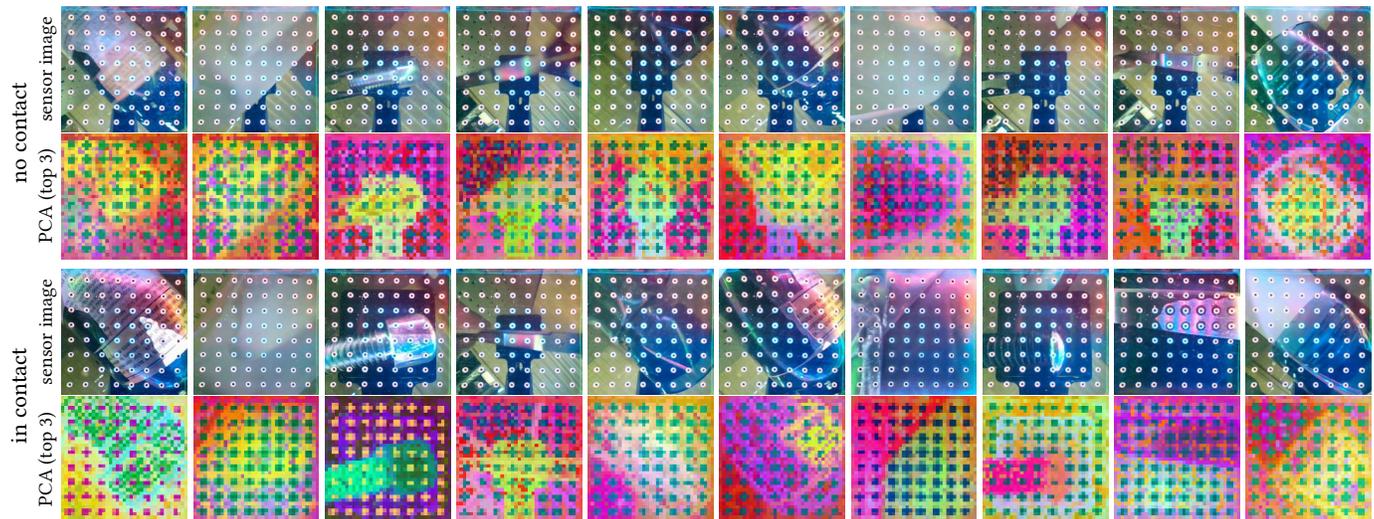
Fig. 10: **PCA heatmaps of DINOv2 image tokens.** Colors are computed by normalizing patch tokens' coefficients of the first three principal components to RGB space.

[23] E. Roberge, G. Fornes, and J.-P. Roberge, "Stereotac: A novel visuotactile sensor that combines tactile sensing with 3d vision," *IEEE Robotics and Automation Letters (RA-L)*, vol. 8, no. 10, pp. 6291–6298, 2023.

[24] T. Ablett, O. Limoyo, A. Sigal, A. Jilani, J. Kelly, K. Siddiqi, F. Hogan, and G. Dudek, "Multimodal and force-matched imitation learning with a see-through visuotactile sensor," *Transactions on Robotics (T-RO)*, vol. 41, pp. 946–959, 2025.

[25] L. Luo, B. Zhang, Z. Peng, Y. K. Cheung, G. Zhang, Z. Li, M. Y. Wang, and H. Yu, "Compdvision: Combining near-field 3d visual and tactile sensing using a compact compound-eye imaging system," in *International Conference on Intelligent Robots and Systems (IROS)*, 2024.

[26] W. Fan, H. Li, W. Si, S. Luo, N. Lepora, and D. Zhang, "Vitactip: Design and verification of a novel biomimetic physical vision-tactile fusion sensor," in *International Conference on Robotics and Automation (ICRA)*, 2024.

[27] Y. Dong, J. Ren, Z. Liu, Z. Peng, Z. Yuan, N. Zhang, and G. Gu, "Look-to-touch: A vision-enhanced proximity and tactile sensor for distance and geometry perception in robotic manipulation," in *IROS Workshop*, 2025.

[28] K. Shimonomura, H. Nakashima, and K. Nozu, "Robotic grasp control with high-resolution combined tactile and proximity sensing," in *International Conference on Robotics and Automation (ICRA)*, 2016.

[29] L. Cui, Z. Zhao, S. Xie, W. Zhang, Z. Han, and Y. Zhu, "Vi-tacman: Articulated object manipulation via vision and touch," *arXiv preprint arXiv:2510.06339*, 2025.

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

[31] Y. Yang, X. Wei, N. Zhang, J. Zheng, X. Chen, Q. Wen, X. Luo, C.-Y. Lee, X. Liu, X. Zhang, *et al.*, "A non-printed integrated-circuit textile for wireless theranostics," *Nature communications*, vol. 12, no. 1, p. 4876, 2021.

[32] Y. Yu, J. Li, S. A. Solomon, J. Min, J. Tu, W. Guo, C. Xu, Y. Song, and W. Gao, "All-printed soft human-machine interface for robotic physicochemical sensing," *Science Robotics*, vol. 7, no. 67, p. eabn0495, 2022.

[33] G. Tian, W. Deng, Y. Gao, D. Xiong, C. Yan, X. He, T. Yang, L. Jin, X. Chu, H. Zhang, *et al.*, "Rich lamellar crystal baklava-structured pzt/pvdf piezoelectric sensor toward individual table tennis training," *Nano Energy*, vol. 59, pp. 574–581, 2019.

[34] P. Lv, J. Qian, C. Yang, T. Liu, Y. Wang, D. Wang, S. Huang, X. Cheng, and Z. Cheng, "Flexible all-inorganic sm-doped pmn-pt film with ultrahigh piezoelectric coefficient for mechanical energy harvesting, motion sensing, and human-machine interaction," *Nano Energy*, vol. 97, p. 107182, 2022.

[35] J. A. Fishel and G. E. Loeb, "Sensing tactile microvibrations with the biotac—comparison with human sensitivity," in *International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, 2012.

[36] S. Suresh, H. Qi, T. Wu, T. Fan, L. Pineda, M. Lambeta, J. Malik, M. Kalakrishnan, R. Calandra, M. Kaess, *et al.*, "Neuralfeels with neural fields: Visuotactile perception for in-hand manipulation," *Science Robotics*, vol. 9, no. 96, p. eadl0628, 2024.

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

[38] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *Journal of Machine Learning Research (JMLR)*, vol. 17, no. 39, pp. 1–40, 2016.

[39] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, *et al.*, "Scalable deep reinforcement learning for vision-based robotic manipulation," in *Conference on Robot Learning (CoRL)*, 2018.

[40] Y. Wu, W. Yan, T. Kurutach, L. Pinto, and P. Abbeel, "Learning to manipulate deformable objects without demonstrations," in *Robotics: Science and Systems (RSS)*, 2020.

[41] M. Shridhar, L. Manuelli, and D. Fox, "Perceiver-actor: A multi-task transformer for robotic manipulation," in *Conference on Robot Learning (CoRL)*, 2023.

[42] A. Goyal, J. Xu, Y. Guo, V. Blukis, Y.-W. Chao, and D. Fox, "Rvt: Robotic view transformer for 3d object manipulation," in *Conference on Robot Learning (CoRL)*, 2023.

[43] F. Liu, C. Li, Y. Qin, A. Shaw, J. Xu, P. Abbeel, and R. Chen, "Vitamin: Learning contact-rich tasks through robot-free visuo-tactile manipulation interface," *arXiv preprint arXiv:2504.06156*, 2025.

[44] L. Heng, H. Geng, K. Zhang, P. Abbeel, and J. Malik, "Vitacformer: Learning cross-modal representation for visuo-tactile dexterous manipulation," *arXiv preprint arXiv:2506.15953*, 2025.

[45] Z. Zhao, S. Haldar, J. Cui, L. Pinto, and R. Bhirangi, "Touch begins where vision ends: Generalizable policies for contact-rich manipulation," in *RSS Workshop*, 2025.

[46] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, 2021.

[47] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, "Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks," *IEEE Robotics and Automation Letters (RA-L)*, vol. 7, no. 3, pp. 7327–7334, 2022.

[48] Z. Xu, R. Uppuluri, X. Zhang, C. Fitch, P. G. Crandall, W. Shou, D. Wang, and Y. She, "Unit: Data efficient tactile representation with generalization to unseen objects," *IEEE Robotics and Automation Letters (RA-L)*, 2025.

[49] J. Zhao, Y. Ma, L. Wang, and E. Adelson, "Transferable tactile transformers for representation learning across diverse sensors and tasks," in *Conference on Robot Learning (CoRL)*, 2024.

[50] W. Chen, J. Xu, F. Xiang, X. Yuan, H. Su, and R. Chen, "General-purpose sim2real protocol for learning contact-rich manipulation with marker-based visuotactile sensors," *Transactions on Robotics (T-RO)*, vol. 40, pp. 1509–1526, 2024.

[51] J. Xu, S. Kim, T. Chen, A. R. Garcia, P. Agrawal, W. Matusik, and S. Sueda, "Efficient tactile simulation with differentiability for robotic manipulation," in *Conference on Robot Learning (CoRL)*, 2023.

[52] S. Zhang, Y. Sun, J. Shan, Z. Chen, F. Sun, Y. Yang, and B. Fang, "Tirgel: A visuo-tactile sensor with total internal reflection mechanism for external observation and contact detection," *IEEE Robotics and Automation Letters (RA-L)*, vol. 8, no. 10, pp. 6307–6314, 2023.

[53] J. Yin, H. Qi, J. Malik, J. Pikul, M. Yim, and T. Hellebrekers, "Learning in-hand translation using tactile skin with shear and normal force sensing," in *International Conference on Robotics and Automation (ICRA)*, 2025.

[54] H. Qi, B. Yi, S. Suresh, M. Lambeta, Y. Ma, R. Calandra, and J. Malik, "General in-hand object rotation with vision and touch," in *Conference on Robot Learning (CoRL)*, 2023.

[55] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[56] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *International Journal of Robotics Research (IJRR)*, vol. 44, no. 10-11, pp. 1684–1704, 2025.

[57] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[58] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, *et al.*, "Dinov2: Learning robust visual features without supervision," *Transactions on Machine Learning Research (TMLR)*, 2023.

[59] T. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Robotics: Science and Systems (RSS)*, 2023.

[60] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, 2019.

[61] X. Song, X. Xu, J. Zhang, D. M. Reyes, and P. Yan, "Dino-reg: Efficient multimodal image registration with distilled features," *IEEE Transactions on Medical Imaging*, 2025.

[62] A. Caraffa, D. Boscaini, A. Hamza, and F. Poiesi, "Freeze: Training-free zero-shot 6d pose estimation with geometric and vision foundation models," in *European Conference on Computer Vision (ECCV)*, 2024.

[63] L. Waldmann, A. Shah, Y. Wang, N. Lehmann, A. Stewart, Z. Xiong, X. X. Zhu, S. Bauer, and J. Chuang, "Panopticon: Advancing any-sensor foundation models for earth observation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[64] J. Zhao, N. Kuppuswamy, S. Feng, B. Burchfiel, and E. Adelson, "Poly-touch: A robust multi-modal tactile sensor for contact-rich manipulation using tactile-diffusion policies," *International Conference on Robotics and Automation (ICRA)*, 2025.

[65] Y. Li, W. Du, C. Yu, P. Li, Z. Zhao, T. Liu, C. Jiang, Y. Zhu, and S. Huang, "Taccel: Scaling up vision-based tactile robotics via high-performance gpu simulation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.