Simultaneous Tactile-Visual Perception for Learning Multimodal Robot Manipulation

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

https://tacthru.yuyang.li

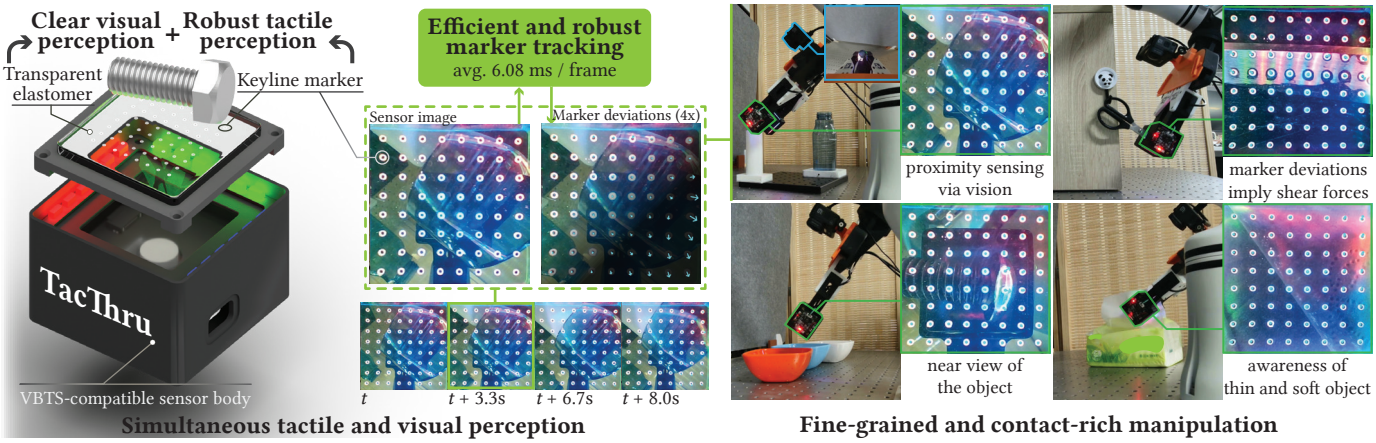

Fig. 1: **Learning multimodal robot manipulation with simultaneous tactile-visual perception.** **`TacThru`** enables clear visual perception and robust marker tracking via transparent elastomer and keyline markers (left), providing rich multimodal signals for learning manipulation policies. **`TacThru-UMI`** 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 alternating tactile-visual (66.3%) and vision-only (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**

This work is supported in part by the National Science and Technology Innovation 2030 Major Program (2025ZD0219400), the National Natural Science Foundation of China (62376009), the State Key Lab of General AI at Peking University, the PKU-BingJi Joint Laboratory for Artificial Intelligence, and the National Comprehensive Experimental Base for Governance of Intelligent Society, Wuhan East Lake High-Tech Development Zone. We thank Lei Yan (LeapZenith AI Research), Shengyu Guo (PKU), Yu Liu (THU), and Leiyao Cui (PKU) for their assistance.

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 Artificial Intelligence.

All authors are with the State Key Lab for General Artificial Intelligence.

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

Yinghan Chen is also 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 have complementary limitations. Vision provides rich global context, but frequently fails during manipulation due to occlusions from the robot's own end-effector or objects, precisely when precise control is most critical [3]. Vision-based Tactile Sensors (VBTSs) [4–7] excel in providing high-fidelity 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 precision and rich information content of vision or tactile modalities, limiting their utility for fine-grained manipulation.

See-Through-Skin (STS) sensors have emerged as a promising solution integrating tactile and visual sensing [10, 14–27]. Adapting VBTS designs, they replace opaque reflective coatings with semi-opaque or transparent alternatives, allowing embedded cameras to "see through the skin" for both tactile

signals (*e.g.*, contact, force distributions) and visual signals (*e.g.*, object appearance, proximity). Recent work has demonstrated their effectiveness in object perception [22, 25, 26], slip detection [15], grasping [17, 22, 28], in-hand manipulation [10], and articulated object manipulation [18, 24].

Despite this progress, three fundamental limitations prevent the broader adoption of STS sensors. Most current designs require *the* switch between tactile and visual modalities through illumination control [23, 24] or movable components [29, 30], preventing *simultaneous* multimodal perception and requiring additional control logic. Tactile markers essential for shear force measurement become difficult to track against *noisy*, *unpredictable* external backgrounds where contrast diminishes, limiting applications in open environments [14]. Finally, existing STS applications rely primarily on *hand-crafted* controllers, with integration into modern data-driven manipulation pipelines that remain largely unexplored.

We introduce **`TacThru`**, an STS sensor that addresses these limitations, featuring specialized design choices: (i) a fully transparent elastomer that enables clear visual perception, (ii) persistent illumination that eliminates mode switching, (iii) novel keyline markers that maintain visibility against any background, and (iv) an efficient tracking algorithm processing marker deviations at $6.08\,\mathrm{ms}$ per frame. This design enables truly *simultaneous* tactile-visual perception while remaining compatible with the standard VBTS fabrication pipeline for seamless integration into existing systems.

We further develop **`TacThru-UMI`**, integrating **`TacThru`** into an imitation learning framework based on Universal Manipulation Interface (UMI) [31]. With both visual and tactile modalities provided, a Transformer-based Diffusion Policy learns to properly attend to these rich 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 achieve an average success rate of 85.5%, representing 54.3% and 29.0% relative improvements over baselines of vision only (55.4%) and alternating tactile-visual (66.3%), 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

STS Sensors: Conventional VBTSs like GelSight [4] and 9D-Tact [5] provide high-resolution tactile sensing for contact measurement [4, 5, 7, 32], state estimation [8], and manipulation [3, 33–35], but require physical contact, limiting pre-contact perception. STS sensors address this limitation by integrating tactile and visual perception [14, 18–25], typically replacing opaque coatings with semi-transparent alternatives [18, 23, 24]. Existing implementations include alternating internal illumination for modal switching [18, 19, 22, 23], mechanically movable components [27, 29], and switchable-transparency films [20, 21]. Advanced designs incorporate fluorescent markers with UV lighting [19, 25], stereo depth perception [23, 28], and lens arrays [25]. In contrast, our design achieves truly *simultaneous* tactile-visual perception through fully transparent elastomer and persistent illumination, eliminating the need for modal *switching*. Further designs include novel keyline markers with an efficient marker tracking algorithm for robust tactile perception.

Multimodal Perception for Manipulation: The combination of tactile and visual modalities addresses the inherent limitations of individual sensing modes: tactile feedback provides precise contact information but remains local and sparse [10], while 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. However, previous research focused predominantly on hand-crafted task-specific controllers rather than leveraging modern general-purpose learning frameworks [31, 36].

Learning with Multimodal Sensing: Modern learning-based policies encode multimodal sensory streams into unified representations or tokens. Visual data processing employs Convolutional Neuron Network (CNN) or Transformer encoders [37–41], while tactile integration varies by signal characteristics: low-dimensional vectors of piezoelectric arrays utilize Multi-Layer Perceptrons (MLPs) [42–44], while high-resolution VBTS images leverage specialized encoders or foundation models like T3 [45, 46]. To reduce learning complexity, many approaches use processed tactile signals—contact depth, marker displacements—rather than raw images [47–49]. Despite these advances, the integration of STS sensors into learning frameworks remains largely unexplored [24]. We address this gap by developing **`TacThru-UMI`**, which integrates **`TacThru`** into an imitation learning framework and demonstrates that simultaneous multimodal perception enables superior performance in fine-grained and contact-rich manipulation.

III. The `TacThru` Sensor

The **`TacThru`** sensor 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 [29, 30], but compromise visual clarity and require complex switching mechanisms. We adopted a fully transparent elastomer with

Fig. 2: **Fabrication of the `TacThru` sensor and integration into the `TacThru-UMI` system.** (a) The keyline marker elastomer is fabricated by sequentially spraying inner (black) and outer (white) markers on transparent elastomer using laser-cut masks. (b) The **`TacThru`** sensor features an extended linkage that serves as gripper fingers. (c) The **`TacThru-UMI`** platform includes a robot end-effector (left) and a data collector (right) that share identical body and finger designs, with the fingers actuated by an Inspire LAS30-021D servo electric cylinder with a maximum opening width of 72 mm.

persistent LED illumination, trading traditional depth perception for continuous visual access. This design recognizes that geometric contact information, while valuable, is often less critical than multimodal coordination for manipulation tasks [18, 24, 50]. Nevertheless, contact detection remains available through two mechanisms: light reflection changes at contact interfaces [18, 22, 50] and marker divergence from elastomer deformation [24].

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.

As shown in Fig. 2ab, the markers are fabricated by sequentially painting inner circles (black, $r_{\text{in}} = 0.6\,\text{mm}$) and outer circles (white, $r_{\text{out}} = 1.0\,\text{mm}$) using laser-cut masks. We deploy $N_m = 64$ markers with 3.5 mm spacing 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 for robust marker tracking, modeling each marker's position $x_t \in \mathbb{R}^2$ at time t using its known initial position x_0 from fabrication. The state transition and measurement follow:

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

where $w_t \sim \mathcal{N}(0, \sigma_w^2 \mathbb{I}_2)$ and $v_t \sim \mathcal{N}(0, \sigma_v^2 \mathbb{I}_2)$ are process and measurement noise. We adopt the random walk model with $A_t = \mathbb{I}_2$ and direct position observation with $H_t = \mathbb{I}_2$. The filter maintains posterior estimates $\hat{x}_t$ and covariances $P_t := \mathbb{E}\left[(x_t - \hat{x}_t)(x_t - \hat{x}_t)^T\right]$ following standard prediction and update steps [51]. The final deviations of the markers are computed as $\Delta x_t := \hat{x}_t - \hat{x}_0$.

Measurement acquisition proceeds through grayscale conversion, intensity thresholding (pixels $< \tau$ set to black) to reveal keylines while removing environmental edges, and blob detection resulting in candidates $Z_t := \{\hat{z}_t\} = \text{BlobDet}(I_t)$. Since Z_t also contains environmental false detections, we filter through distance-based data association, matching each marker to its nearest detection:

$$z_t = \arg\min_{z \in Z_t} \|z - \hat{x}_{t-1}\|. \tag{2}$$

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 [31], one with keyline markers and one with solid black markers. We collect 8 trajectories (1628 frames per sensor), grasping a plastic bottle with complex black-and-white text that challenges marker detection. The **`TacThru`** sensor 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.

Fig. 3b illustrates the detection challenge: solid markers become invisible against black backgrounds, while keyline markers remain detectable, although environmental noise requires filtering. We further 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 with Kalman filtering (Sec. III-C), using $\sigma_w = 0.1, \sigma_v = 0.025$.

Fig. 3c presents quantitative results that include detected markers per frame and processing time. Solid markers suffer

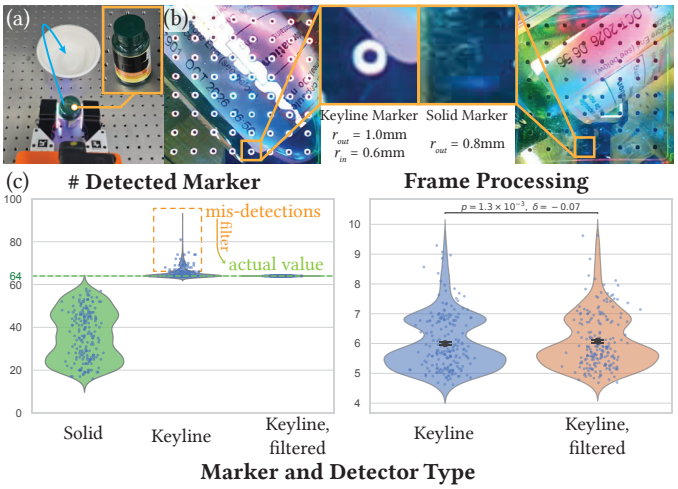

Fig. 3: **Keyline marker design and filtering enable robust tracking.** (a) Evaluation setup compares two sensor types (keyline *vs*. solid markers) during bottle grasping tasks. (b) The **TacThru** sensor view comparison shows that keyline markers (left) remain distinct against complex backgrounds, 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 per frame), while solid markers suffer missed detections and unfiltered keyline detection produces false positives (count > 64).

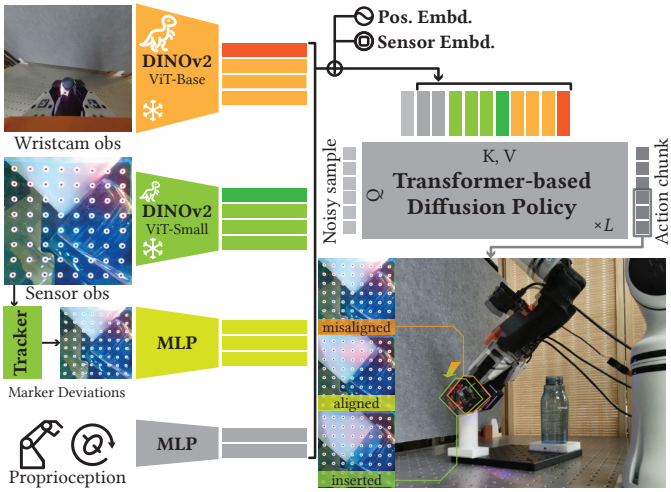

Fig. 4: **Diffusion policy architecture for TacThru-UMI.** Multimodal observations—wrist-camera RGB images, sensor RGB images, detected marker deviations, and proprioception—are encoded into tokens and augmented with positional and modality-specific embeddings. These tokens condition a Transformer-based diffusion policy that denoises Gaussian noise into action chunks for robot execution. The example shows how the policy leverages the **TacThru**'s close-up view to align the cap and mount during the **InsertCap** task.

from severe missed detections, resulting in incomplete tactile information. Unfiltered keyline detection reports marker counts exceeding the actual $N_m = 64$ due to environmental false detections (orange dashed box). Our filtering strategy achieves the design goal: persistent tracking of all markers with minimal computational overhead, eliminating false detections, while maintaining 6.08 ms processing suitable for high-frequency perception (*e.g.*, 120 Hz) and real-time operation.

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 [31] and the Diffusion Policy [36] with multimodal 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 [31] design, replacing standard fingers with STS sensors on extended linkages (Fig. 2c). The **TacThru** sensors are connected to a computer via USB to stream real-time images during demonstrations and policy inference.

Robot end-effector: Although the UMI end-effector design allows for inference with various parallel grippers such as the Panda Hand, we design a low-cost gripper that directly mirrors the data collector's design to minimize embodiment gaps (Fig. 2c). Both platforms share identical body and finger des, with fingers actuated by an Inspire LAS30-021D servo-electric cylinder (∼ $280).

B. Data Collection and Processing

Our pipeline extends UMI [31] with tactile modalities. To improve robustness during contact-rich manipulation, we replaced the original SLAM-based pose tracking with an HTC Vive Tracker, which maintains stable tracking even when SLAM fails due to visual occlusion. 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 Diffusion Policy [36] with the Transformer architecture [52] 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 $\mathbf{I}_w^t := \{I_w^i\}_{i=t-n_w^{\text{obs}}}^{t-1}$, sensor frames $\mathbf{I}_s^t := \{I_s^i\}_{i=t-n_s^{\text{obs}}}^{t-1}$, marker deviations $\Delta\mathbf{x}^t := \{\Delta x^{i,j}, j = 1, \ldots, N_m\}_{i=t-n_s^{\text{obs}}}^{t-1}$, and proprioception $\mathbf{s}^t := \{s^i\}_{i=t-n_p^{\text{obs}}}^{t-1}$ containing the pose of the end-effector and the width of the gripper in relative coordinates [31]. Visual observations are encoded using DINOv2 [53]: ViT-Base for wrist cameras and ViT-Small for **TacThru** frames (both 14×14 patch size). Despite the domain shift from markers and elastomer, DINOv2 effectively

handles tactile sensor imagery. Marker deviations and proprioception use dedicated MLPs. Each modality receives learnable embeddings (z_w, z_s, z_x, z_p) for transformer distinguishability:

$$\mathbf{z}_w = \left\{\text{DINO}_w(I) + z_w | I \in \mathbf{I}_w^t\right\}, \tag{3}$$

$$\mathbf{z}_s = \left\{\text{DINO}_s(I) + z_s | I \in \mathbf{I}_w^t\right\}, \tag{4}$$

$$\mathbf{z}_x = \left\{\text{MLP}_x(\Delta x) + z_x | \Delta x \in \Delta \mathbf{x}^t\right\}, \tag{5}$$

$$\mathbf{z}_p = \left\{\text{MLP}_p(s) + z_p | s \in \mathbf{s}^t\right\}. \tag{6}$$

Concatenated tokens with positional embeddings condition the diffusion policy π_θ, which denoises Gaussian samples into action chunks [54] $\mathbf{a} = \{a^i\}_{i=t}^{t+T_a-1} \sim \pi_\theta(\mathbf{a}|\mathbf{z}_w, \mathbf{z}_s, \mathbf{z}_x, \mathbf{z}_p)$. 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, position, 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. 5 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 (button head for A, socket head for B and C) and color (black for A and C, silver for B). 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. This task, borrowed from Liu *et al.* [42] to benchmark VBTS-based UMI systems, requires tactile feedback to distinguish successful hanging (triggering gripper release) from missed attempts (triggering retry).

`InsertCap`: A bottle cap is placed on a bottle. The robot must grasp the cap and insert it onto a white mount, which requires millimeter-level precision alignment. Although insertion tasks are typically based on tactile signals [48, 49], **`TacThru`** enables direct visual servoing for the alignment of the cap. When the view is occluded (*e.g.*, due to biased grasping), tactile signals from marker displacement provide robust fallback guidance.

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. 6) and train four policy variants that naturally form ablations and comparisons:

- **`TT-M`:** Full **`TacThru`** with image frames ($\mathbf{I}_s^t$) and marker deviations ($\Delta\mathbf{x}^t$).
- **`TT`: `TacThru`** image frames only ($\mathbf{I}_s^t$); markers are visible in the frames but not explicitly extracted and provided (ablation w/o marker deviations).
- **`GS-M`:** GelSight sensor frames (rectified by the idle image for revealing only intensity changes) and marker deviations (tactile baseline).
- **`Wrist`:** Wrist camera only ($\mathbf{I}_w^t$) (vision-only baseline).

All policies include a wrist camera ($\mathbf{I}_w^t$) and proprioception ($\mathbf{s}^t$) as the base input.

Training and evaluation: We use observation horizons $n_w^{\text{obs}} = 1, n_s^{\text{obs}} = 1, n_p^{\text{obs}} = 2$, predict $T_a = 16$ action steps, and execute a 6-step window (steps 3-8). Each policy is trained for 150 epochs using the AdamW optimizer [55] with a learning rate of 3×10^{-4} and a one-cycle scheduling. We evaluated the checkpoints after 20 runs per task (24 for `SortBolt`) with randomized initialization that matched the data-collection distribution.

C. Results

Basic manipulation: All variants achieve near-perfect success in **`PickBottle`** (Fig. 7a), with success rates exceeding 95%. This validates that **`TacThru-UMI`**'s architecture effectively learns manipulation policies across different sensing modalities. The consistent performance demonstrates that our diffusion policy framework successfully integrates diverse input types without degrading basic manipulation capabilities. This establishes a foundation for evaluating more challenging scenarios in which specific sensing modalities become critical.

Thin-and-soft object perception: **`PullTissue`** reveals the fundamental limitations of conventional tactile sensing. Traditional contact-based sensors require sufficient normal and shear forces to generate detectable signals, but tissues exert minimal pressure and deform rather than resist. **`TacThru`** overcomes this through direct visual observation of the inter-finger workspace, providing continuous feedback on tissue position and deformation. When tissue slippage (Fig. 7b) occurs—often due to insufficient grip force on the delicate

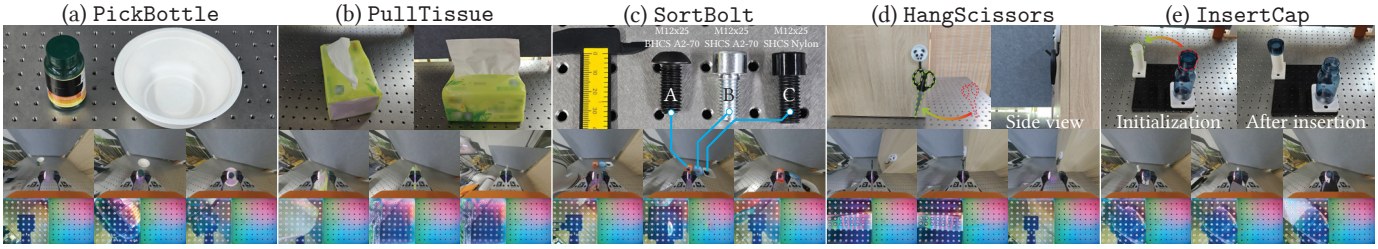

Fig. 5: **Task demonstrations across five manipulation scenarios.** (a) `PickBottle`: basic pick-and-place, (b) `PullTissue`: thin-and-soft object manipulation, (c) `SortBolt`: visual discrimination, (d) `HangScissors`: tactile discrimination, (e) `InsertCap`: multimodal fusion. **Top:** Initial object configurations. **Middle:** Wrist-camera view progression during demonstration. **Bottom:** Corresponding **`TacThru`** (top) and GelSight sensor (bottom) observations, illustrating distinct sensing modalities and information content.

material—**`TacThru`** immediately detects the displacement and triggers corrective re-grasping actions. The wrist camera fails due to insufficient resolution at its typical mounting distance (15 cm), whereas the `GS-M` shows near-zero success because soft tissues cannot generate the contact patterns needed for reliable tactile inference.

Visual discrimination: **`SortBolt`** (Fig. 7c) demonstrates **`TacThru`**'s superior visual discrimination capabilities. The small M12×25 bolts (12mm head diameter) present significant challenges: wrist cameras cannot resolve geometric details at manipulation distances, while identical bolt shapes make purely tactile discrimination impossible. **`TacThru`**'s close-proximity view (2-3mm from objects) captures fine geometric features and subtle color differences that remain invisible to distant cameras. Fig. 8 provides quantitative evidence through DINOv2 embedding analysis: **`TacThru`** produces clearly separated feature clusters with inter-cluster distances exceeding 0.8, while GelSight embeddings for bolts B and C overlap with similarity scores above 0.9. This embedding separation directly correlates with the 85% *vs.* 45% success rates observed between `TT-M` and `GS-M` policies.

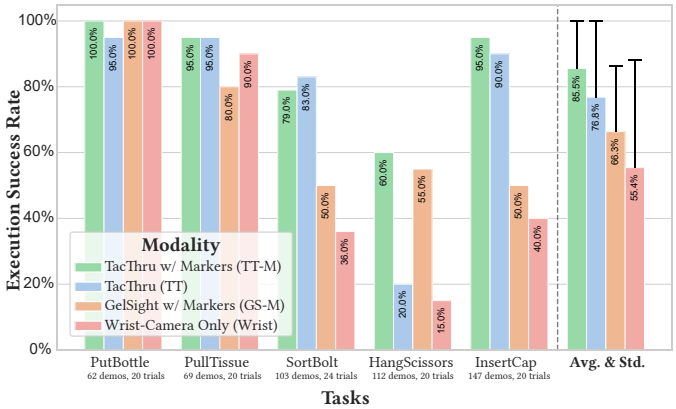

Fig. 6: **Quantitative results across manipulation tasks and sensing modalities.** Success rates for four policy variants: `TT-M` (**`TacThru`** with markers), `TT` (**`TacThru`** only), `GS-M` (GelSight with markers), and `Wrist` (vision-only). Each task evaluates specific sensing capabilities: basic manipulation (`PickBottle`), thin-and-soft object manipulation (`PullTissue`), visual discrimination (`SortBolt`), tactile discrimination (`HangScissors`), and multimodal fusion (`InsertCap`). Error bars show standard deviation across evaluation runs. The rightmost column presents overall performance averages.

Tactile discrimination: **`HangScissors`** (Fig. 7d) exemplifies scenarios where visual observation alone cannot determine task completion. The 2D wrist camera cannot reliably detect whether scissor handles have successfully engaged the hook, due to depth perception and occlusion. Physical contact patterns captured through marker displacement provide unambiguous confirmation: successful engagement creates characteristic force patterns as the scissors settle into position, while failed attempts show continued downward forces. Both `TT-M` and `GS-M` leverage these tactile signals effectively, achieving success rates of 80%+ compared to 35% for vision-only baselines. The explicit marker-based feedback enables precise timing of the gripper release—a critical decision point that determines task success.

Multimodal fusion: **`InsertCap`** (Fig. 7e-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, the policy employs vision-based servoing, directly aligning visual features for precise insertion (Fig. 7e). However, when grasping occludes the view or lighting conditions degrade visual signals, the policy seamlessly returns to tactile-based insertion, using marker displacement patterns to detect contact and guide alignment (Fig. 7f). This adaptive behavior emerges naturally from the training process without explicit strategy programming, demonstrating the policy's ability to weight modalities based on their reliability in different contexts. The 90% success rate reflects this robust dual-strategy approach, significantly outperforming single-modality baselines that lack this adaptive capability.

D. Discussions

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

Adaptive multimodal strategies: Policies trained with **`TacThru`** naturally learn to weight sensing modalities based on context reliability, 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 approaches.

Overcoming conventional tactile limitations: **`TacThru`** uniquely handles scenarios where traditional

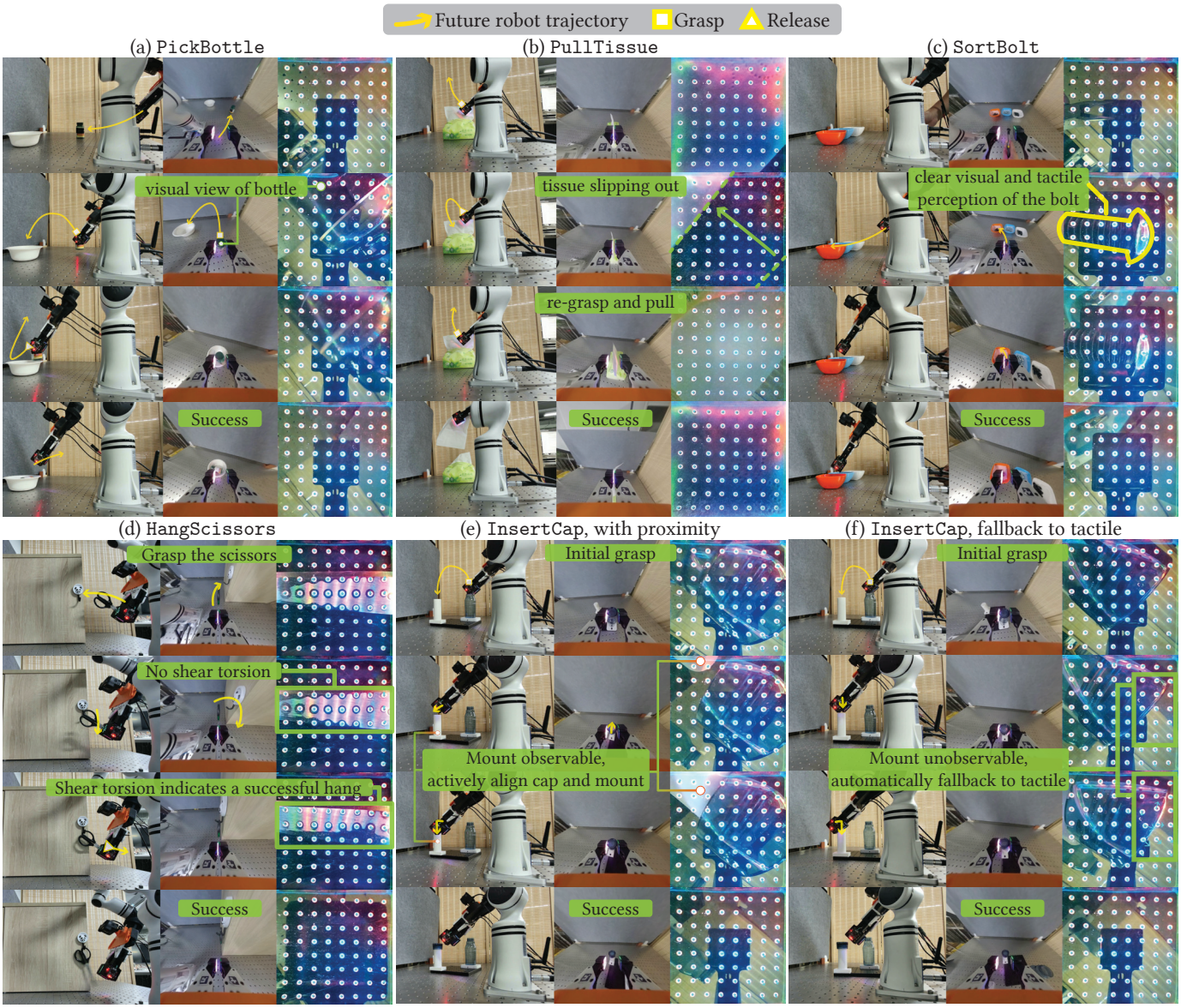

Fig. 7: **Qualitative policy rollouts demonstrating manipulation execution.** Each column (a-f) shows the temporal progression of a task execution from top to bottom, with three synchronized views: third-person perspective (left), wrist camera (center), and **TacThru** close-up (right) with tracked marker deviations overlaid (4× magnified for visibility). Colored annotations highlight key manipulation phases and sensing feedback. See supplementary video for additional rollouts across all policy variants.

contact-based sensors fail. Thin objects such as tissue generate insufficient forces for conventional tactile detection, but remain clearly observable through direct optical monitoring. This expands the range of manipulable objects beyond current tactile sensing capabilities.

Practical deployment viability: Despite significant domain differences—transparent elastomer, marker overlays, 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.

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 of **TacThru** and **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 [35, 48, 49] 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.

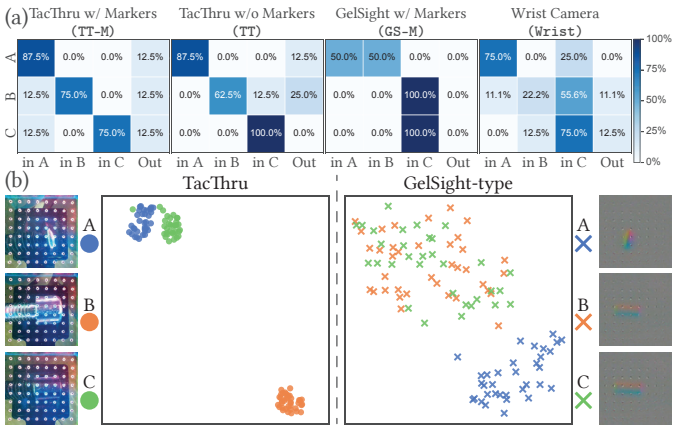

Fig. 8: **Bolt sorting performance analysis across sensing modalities.** (a) Confusion matrices showing placement accuracy for each policy variant. Rows indicate ground-truth bolt type, columns show predicted placement (bowls A-C or "Out" for misses). **`TacThru`**-based policies (`TT-M`, `TT`) successfully distinguish all bolt types, while `GS-M` confuses geometrically identical bolts B and C. (b) t-SNE visualization of DINOv2 CLS token embeddings from sensor images. **`TacThru`** produces clearly separated clusters for all bolts, while GelSight embeddings for bolts B and C overlap, explaining the observed classification failures.

[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, "TacMan: Tactile-informed prior-free manipulation of articulated objects," *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," *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," *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 *IROS*, 2022.

[11] L.-T. Jiang and J. R. Smith, "Seashell effect pretouch sensing for robotic grasping.," in *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," *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 *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 *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," *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 *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," *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.

[23] E. Roberge, G. Fornes, and J.-P. Roberge, "Stereotac: A novel visuotactile sensor that combines tactile sensing with 3d vision," *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," *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 *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 *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 *ICRA*, 2016.

[29] J. Xu, L. Wu, C. Lin, D. Zhao, and H. Xu, "Dtactive: A vision-based tactile sensor with active surface," *arXiv preprint arXiv:2410.08337*, 2024.

[30] D. Yueshi, 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.

[31] 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 *RSS*, 2024.

[32] 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.

[33] R. Li, R. Platt, W. Yuan, A. Ten Pas, N. Roscup, M. A. Srinivasan, and E. Adelson, "Localization and manipulation of small parts using gelsight tactile sensing," in *IROS*, 2014.

[34] J. Lloyd and N. F. Lepora, "Pose-and-shear-based tactile servoing," *IJRR*, vol. 43, no. 7, pp. 1024–1055, 2024.

[35] 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 *NeurIPS*, 2025.

[36] 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," *IJRR*, vol. 44, no. 10-11, pp. 1684–1704, 2025.

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

[38] 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 *CoRL*, 2018.

[39] Y. Wu, W. Yan, T. Kurutach, L. Pinto, and P. Abbeel, "Learning to manipulate deformable objects without demonstrations," in *RSS*, 2020.

[40] M. Shridhar, L. Manuelli, and D. Fox, "Perceiver-actor: A multi-task transformer for robotic manipulation," in *CoRL*, 2023.

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

[42] 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.

[43] 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.

[44] 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.

[45] 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 *ICML*, 2021.

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

[47] N. Funk, C. Chen, T. Schneider, G. Chalvatzaki, R. Calandra, and J. Peters, "On the importance of tactile sensing for imitation learning: A case study on robotic match lighting," *arXiv preprint arXiv:2504.13618*, 2025.
[48] 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," *T-RO*, vol. 40, pp. 1509–1526, 2024.
[49] 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 *CoRL*, 2023.
[50] 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," *RA-L*, vol. 8, no. 10, pp. 6307–6314, 2023.
[51] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.
[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.
[53] 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.
[54] T. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *RSS*, 2023.
[55] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.