Seeing Touch from Motion: A Unified Modality-Aware Visuo-Tactile Policy with Tactile Motion Correlation

Shengqi Xu[1,2,3], Yang Liu[1,2], Guojin Zhong[1,2], Fanjie Wang[1,2], Hu Luo[1,2], Hanyu Zhou[4], Weiyao Zhang[1,2], Ziyi Ye[1,2], Zuxuan Wu[1,2,3,✉], Yu-Gang Jiang[1,2,✉]

[1] Institute of Trustworthy Embodied AI, Fudan University, China
[2] Shanghai Key Laboratory of Multimodal Embodied AI, China [3] NeoteAI, China
[4] School of Computing, National University of Singapore, Singapore
Website: https://shengqi77.github.io/Seeing-Touch-from-Motion/

Abstract. Visuo-Tactile policies leveraging optical tactile sensors have shown great promise in contact-rich manipulation. These sensors achieve high spatial resolution and multi-dimensional force sensing by utilizing an internal camera to monitor the deformation of their elastic gel surface, thereby indirectly inferring tactile cues. Despite their advantages, extracting fine-grained contact states necessary for contact-rich manipulation remains an open challenge. Existing methods typically use either raw images or cumulative motion fields to represent tactile cues. However, both are prone to perception ambiguity. Raw tactile images mainly capture appearance changes, while cumulative motion fields only reflect the aggregate gel deformation. Consequently, distinct fine-grained contact states can exhibit highly similar patterns, making it difficult to explicitly distinguish subtle contact variations. To address this issue, we explore the dynamic priors of tactile motion and discover that the correlation between transient and cumulative motion can explicitly distinguish fine-grained contact states. Based on this insight, we propose a motion-aware tactile representation to facilitate contact-rich manipulation. Beyond tactile representation, effective fusion of tactile and visual modalities is also critical. Most existing fusion methods either directly concatenate features from each modality or train modality-specific networks separately and fuse their outputs. However, these strategies struggle to simultaneously model cross-modal interactions and preserve modality-specific characteristics. In this work, we take advantage of the Mixture-of-Transformers architecture and propose a unified modality-aware visuo-tactile policy that captures cross-modal complementarity while maintaining modality-specific properties. Extensive experiments on four challenging contact-rich manipulation tasks indicate the superior performance of our method.

Keywords: Manipulation · Tactile representation · Visuo-tactile fusion

1 Introduction

Contact-rich manipulation is crucial for robots performing real-world fine-grained tasks, such as precision insertion and object wiping, where success hinges on ac-

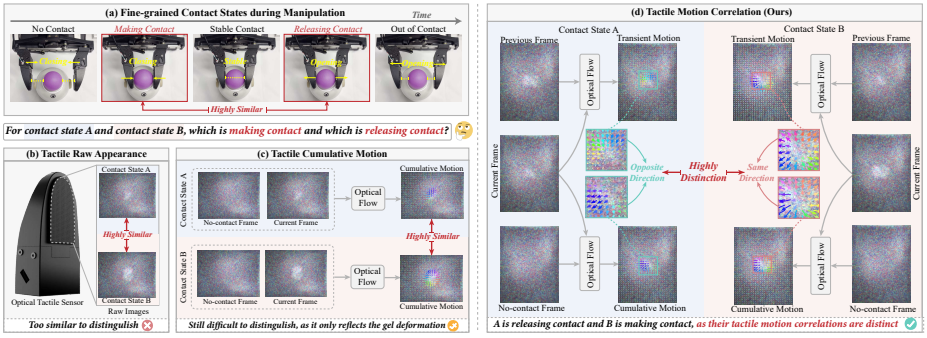


Fig. 1: Comparison of three different tactile representations for distinguishing fine-grained contact states. (a) Fine-grained contact states, such as making contact (contact state B) and releasing contact (contact state A), can be visually similar. (b) Tactile raw appearance primarily captures visual changes but suffers from ambiguity due to their deceptive similarity, making them difficult to distinguish. (c) Tactile cumulative motion (optical flow between the no-contact frame and the current frame) reflects overall deformation but still exhibits similar patterns, making it equally difficult to distinguish. (d) In contrast, our tactile motion correlation explicitly resolves this ambiguity by analyzing the dynamic relationship between transient and cumulative motion. The directional correlation (same v.s. opposite) effectively breaks the similarity, providing highly discriminative cues crucial for contact-rich manipulation.

curately perceiving subtle physical interactions. Such tasks typically rely on the complementary cooperation of global visual perception and local tactile feedback. Specifically, vision provides global contextual information to support the planning of coarse motion trajectories, while touch captures local contact feedback to enable fine-grained interaction with objects.

To capture local contact feedback, a series of tactile sensing technologies has been developed. Among these, optical tactile sensor [1–3,66]) are widely adopted due to their high spatial resolution, texture perception and multi-dimensional force sensing capabilities. These advantages arise from their imaging mechanism, where an internal camera is utilized to monitor the deformation of an elastic gel surface. The surface is typically patterned with markers to better present gel motion during external contacts.

Although optical tactile sensors offer information-rich visual measurements that indirectly reflect tactile cues, effectively extracting fine-grained contact states from these cues remains an underexplored problem. Most existing methods [8,58,62] typically rely on either raw images or cumulative motion fields (*i.e.*, optical flow between the current frame and the initial non-contact frame) to represent tactile information. However, both representations struggle to distinguish fine-grained contact states, such as the distinction between making contact and releasing contact (see Fig. 1(a)), which is crucial for contact-rich manipulation. As shown in Fig. 1(b), raw images primarily capture appearance changes caused by extern contact but suffer from perceptual ambiguity due to their deceptive similarity, making them difficult to distinguish. Similarly, Figure 1(c) illustrates

that cumulative tactile motion only reflects the direction and magnitude of overall gel deformation but still exhibits similar patterns, making it equally difficult to distinguish subtle contact variations.

To address the above issue, we thoroughly explore the underlying motion priors in tactile sensing and discover that *the correlation between transient motion and cumulative motion can more explicitly distinguish fine-grained contact states*. Specifically, transient motion (*i.e.*, optical flow between consecutive frames) captures instantaneous deformation, while cumulative motion captures the overall deformation relative to the initial non-contact state. We observe that different contact states exhibit distinct correlation patterns between these two types of motion. For instance, Figure 1(d) illustrates that when releasing contact, the motion vectors point in opposite directions, whereas when making contact, they align in the same direction (See Fig. 3(a) for more examples). These observations inspire us to propose **Tactile Motion Correlation (TMC)**, a motion-aware tactile representation that explicitly captures fine-grained contact states by modeling the correlation between transient and cumulative motions through their *dot product*. This tactile representation offers three main advantages. First, the sign and magnitude of the dot product enable precise differentiation among fine-grained contact states, while carrying a physically interpretable meaning of the underlying contact. Second, the magnitude exhibits a strong positive correlation with the applied force, thereby faithfully reflecting contact intensity. Third, it is a sensor-agnostic representation compatible with various optical sensors, allowing it to bridge heterogeneous sensors and holds promise for enabling a general tactile representation in the future.

While an informative tactile representation is essential for contact-rich manipulation, the effective integration of local touch with global vision is equally crucial. Existing fusion methods mainly fall into two categories: feature-level fusion methods [39,59,71] and decision-level fusion methods [10] . The former typically concatenates visual and tactile features or applies simple attention mechanisms directly to them. However, such strategies overlook the distinct properties of each modality and may allow one modality to dominate the other. In contrast, the latter typically train an independent policy network for each modality and fuses their outputs. Though it respects modality-specific differences, the absence of interaction between vision and touch prevents the policy from exploiting cross-modal cues, making it difficult to achieve genuine complementarity.

In this work, we take advantage of the Mixture-of-Transformers (MoT) architecture [44] and propose **ViTacMotor**, a unified yet modality-aware **Vi**suo-**Tac**tile policy that incorporates the proposed **Mot**ion-aware c**or**relation representation. The MoT adopted in ViTacMotor decouples modality-specific parameters to preserve the unique properties of each modality, while enabling effective cross-modal interaction via a shared global self-attention mechanism. This design enables ViTacMotor to capture cross-modal complementarity while respecting the individuality of visual and tactile inputs, thereby facilitating effective integration of local touch with global vision for contact-rich manipulation. Our main contributions are summarized as follows:

1. We reveal that the correlation between transient and cumulative tactile motion can more explicitly distinguish fine-grained contact states. Motivated by this finding, we propose **Tactile Motion Correlation**, a motion-aware representation that addresses the perceptual ambiguity in prior arts and thus enable fine-grained contact-state perception for contact-rich manipulation.
2. We propose **ViTacMotor**, a unified yet modality-aware visuo-tactile policy framework. Compared to existing methods, it can capture cross-modal interactions while preserving the unique properties of each modality, enabling an effective integration of global visual perception and local tactile sensing.
3. We compare our method with existing methods on a variety of challenging contact-rich manipulation tasks. Extensive experiments demonstrate the superior performance and robustness of our method.

2 Related Work

Tactile Sensing in Robotics is crucial for acquiring rich microscopic contact feedback during interactions with objects that complements macroscopic visual perception [33, 42, 53]. Tactile sensors aim to transduce external physical contact into measurable electrical signals. They can be categorized into various types based on the transduction mechanisms, including piezoelectric [27,35,68], capacitive [48, 49], magnetic [5, 32], and optical sensors [26, 45, 46, 54, 55, 57, 63, 66]. In particular, optical sensors have been widely used due to their ability to produce high-resolution tactile images and capture multi-dimensional force information.

Most existing visuo-tactile policies [7,14,20,31,43,47,58,61,65,67,69] based on optical sensors typically utilize raw tactile images as observations. For instance, Gu *et al.* [22] and Wu *et al.* [58] directly employ raw tactile images acquired from GelSight-like sensors as inputs to visuomotor policies. However, raw images mainly capture appearance changes resulting from contact and often suffer from perception ambiguity, making it challenging to reflect fine-grained contact states. Recently, several works [8, 62, 74] have attempted to employ the cumulative deformation of the gel surface as motion cues to construct tactile representation. Bogert *et al.* [8] decompose tactile motion using Helmholtz-Hodge decomposition [24], enabling policy transfer across different robotic embodiments. Similarly, Xue *et al.* [62] extract tactile motion cues captured by the GelSight Mini sensor and apply principal component analysis [4] for dimensionality reduction to obtain tactile cues. However, the cumulative motion only describe the magnitude and direction of gel deformation induced by contact, making it insufficient to distinguish subtle contact variations. In this work, we reveal that the correlation of transient and cumulative motion can explicitly distinguish fine-grained contact states, and we propose tactile motion correlation as a robust tactile representation to facilitate contact-rich manipulation.

A related line of learning-based tactile representation work [18, 19, 25] also leverages both transient and cumulative changes, but encodes the cumulative difference and temporal tactile frames *implicitly* into a latent space to learn general tactile representation. In contrast, our method *explicitly* computes the dot product between cumulative and transient motion, yielding physically interpretable

values for contact states. We appreciate their contribution and encourage readers to consult these works for complementary insights.

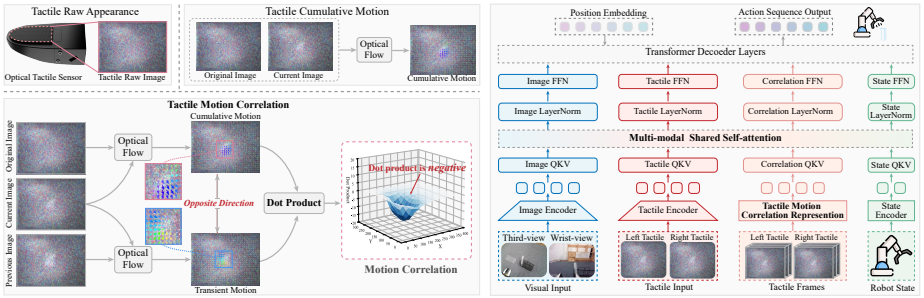
Visual-Tactile Fusion in Robotics is essential for achieving contact-rich tasks and successful interaction with the environment [11, 23, 27, 28, 51, 64, 72]. Vision typically provides global perception, whereas tactile sensing offers local contact feedback during interactions; together they complement each other to enable closed-loop control. The most common fusion methods involve simply concatenating visual and tactile features [29, 40, 50, 59] or applying straightforward attention-based fusion between the two modalities [13, 38, 39, 73]. Feng *et al.* [17] propose a stage-guided dynamic fusion that adaptively ingrates multi-modal features. Li *et al.* [39] propose a self-attention fusion mechanism to integrate visual, auditory, and tactile modalities. However, such strategies treat different modalities uniformly, neglecting their unique characteristics. To address this, chen *et al.* [10] introduce a policy-level approach in which separate policies are trained for each modality, and their outputs are combined via an adaptive weighting mechanism. Though this method respects the individuality of each modality, it overlooks the potential correlations between them. In this work, we propose a unified yet modality-specific visuo-tactile policy based on a Mixture-of-Transformers architecture, capturing cross-modal correlations while preserving the unique properties of each modality.

Mixture-of-Transformers (MoT) [44] has recently shown strong potential in multi-modal learning. It introduces a sparse multimodal transformer architecture that leverages modality-specific parameter decoupling and global self-attention to efficiently process heterogeneous modalities while preserving cross-modal interactions. Owing to this capability, MoT has been used in various tasks, including 3D understanding and reconstruction [12], multimodal generation and understanding [16, 34, 56], world model [6, 41], and Vision-Language-Action (VLA) model [9, 21, 30, 52, 60]. Closet to ours, Wu *et al.* [60] propose an MoT-based VLA model that performs functional decoupling: a vision–language model acts as the understanding expert, while a generative model serves as the action expert, together forming a general-purpose manipulation system. Distinctly, ViTacMotor is the first policy leveraging the MoT-based architecture to achieve unified yet modality-aware visuo-tactile fusion for contact-rich manipulation.

3 Visuo-Tactile Policy with Tactile Motion Correlation

3.1 Overview

In this work, we propose ViTacMotor, a unified yet modality-aware visuo-tactile policy based on a Mixture-of-Transformers (MoT) architecture, augmented with tactile motion correlation representation for contact-rich manipulation, as shown in Fig. 2. On the one hand, we discover that the correlation between transient and cumulative tactile motion can explicitly distinguish fine-grained contact states and propose tactile motion correlation (TMC) as a robust tactile representation to better capture fine-grained contact states (Section 3.2). On the other hand, we take advantage of the MoT architecture and introduce a unified yet



(a) Tactile motion correlation representation (b) Unified modality-aware Visuo-Tactile policy framework

Fig. 2: Overview of the Tactile Motion Correlation (TMC) representation and unified modality-aware visual-tactile policy. (a) TMC models the correlation between transient and cumulative tactile motion through their dot product, which can explicitly distinguish fine-grained contact states. (b) Unified yet modality-aware visuo-tactile policy based on MoT-based architecture for capturing cross-modal complementarity while preserving unique properties of each modality.

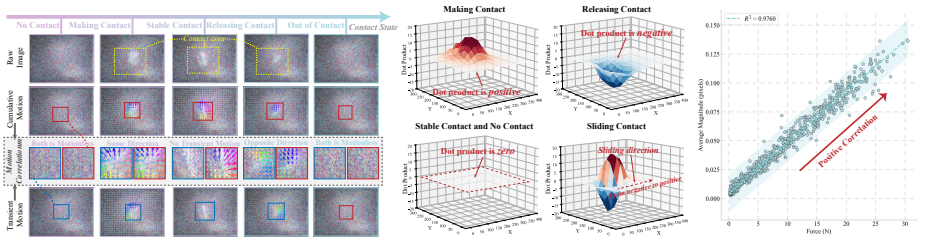
modality-aware visuo-tactile framework to effectively integrate global visual perception and local tactile feedback, achieving cross-modal complementarity while preserving the unique properties of each modality (Section 3.3).

3.2 Tactile Motion Correlation Representation

Existing visuo-tactile policies based on optical tactile sensors utilize either raw tactile appearance or cumulative motion fields as tactile information. However, both representations struggle to distinguish fine-grained contact states, which are important for precise contact-rich manipulation. To address this issue, we thoroughly explore the tactile motion priors of optical sensor underlying various contact states and propose a tactile motion correlation representation that more explicitly capture such fine-grained contact states.

Transient-Cumulative Motion Correlation. To explore tactile motion correlation priors underlying various fine-grained contact states (*e.g.*, no contact, making contact, releasing contact, etc.), we conduct an analysis experiment using a optical sensor with dense markers [2], which features an elastomer surface with markers that facilitate better capture of gel deformation. We use an efficient optical flow estimation method [37] to compute two types of tactile motion: cumulative motion and transient motion. Cumulative motion refers to the optical flow between the current frame and the original non-contact frame, while transient motion is the optical flow between the consecutive frames. Note that we conduct analysis experiments on various optical sensors [1, 3] to validate the *universality* of our findings. (See Appendix for details)

In Fig. 3(a), we visualize the raw tactile images, cumulative motion, and transient motion under different contact states. Raw images mainly reflect the appearance deformation caused by contact, making it difficult to distinguish fine-grained states. Though cumulative motion captures the magnitude and direction



(a) Visualization of tactile raw image, cumulative and transient motion (b) Dot product of two motions in various states (c) Relationship between magnitude and force

Fig. 3: Analysis of tactile motion correlation properties using dense-marker optical tactile sensor. (a)Visualization of raw tactile images, cumulative motion, transient motion, and the correlation between the two types of motion under different fine-grained contact states. (b) 3D distribution of the dot product between transient and cumulative motion under different contact states, showing that the dot product exhibits strong discriminability across contact states. (c) Relationship between the dot product magnitude during making contact and the contact force magnitude, revealing that the dot product magnitude is positively correlated with contact force.

of deformation, it provides limited discriminability, particularly between making contact, stable contact and releasing contact. In contrast, transient motion clearly captures subtle instantaneous deformations between consecutive frames. It shows clear motion differences between making contact and releasing contact, while nearly vanishing when maintaining stable contact. More importantly, we find that the correlation between transient and cumulative motion provides strong discriminative cues for contact states: (1) During no contact and out of contact, both are motionless, indicating no gel deformation; (2) During making contact, both motions exhibit the same direction, reflecting continuous accumulation of gel deformation; (3) When entering stable contact, transient motion almost disappears while cumulative motion remains, indicating the deformation has stabilized; (4) During releasing contact, the two motions show opposite directions, implying the gel is recovering. This insight reveals that *transient motion complements the temporal lag of cumulative motion, and their correlation can more explicitly distinguish fine-grained contact states*, inspiring us to propose tactile motion correlation as a tactile representation via their dot product.

Representing Correlation via Dot Product. Given the initial frame T_0, the current frame T_t, and the previous frame T_{t-1}, we estimate the transient motion $\mathbf{M}_t^{\text{tran}}$ and the cumulative motion $\mathbf{M}_t^{\text{cumu}}$ using optical flow:

$$\begin{aligned} \mathbf{M}_t^{\text{tran}} &= \mathbf{O}_t^{t-1\to t} = \text{Flow}(T_{t-1}, T_t), \\ \mathbf{M}_t^{\text{cumu}} &= \mathbf{O}_t^{0\to t} = \text{Flow}(T_0, T_t). \end{aligned} \tag{1}$$

Here, $\text{Flow}(\cdot,\cdot)$ denotes the optical flow estimation function. $\mathbf{O}_t^{t-1\to t}$ denotes the instantaneous motion from the previous frame to the current frame, whereas $\mathbf{O}_t^{0\to t}$ captures the aggregate motion relative to the initial state. To model the correlation between these two motions, we compute their pixel-wise dot product:

$$Corr_t(x) = \mathbf{M}_t^{\text{tran}}(x) \cdot \mathbf{M}_t^{\text{cumu}}(x), \quad x \in \Omega. \tag{2}$$

where x denotes a pixel location and $\Omega \subset \mathbb{Z}^2$ is the image domain. $Corr_t(x)$ refers to the motion correlation at pixel x and time t, encoding the correlation between instantaneous and aggregate gel deformation.

How does Dot Product Reflect Contact Feedback? We further discuss the relationship between the dot product and the contact state. Fig. 3(b) illustrates the 3D distribution of the dot product between two motions under different contact states. It is observed that during making contact, the dot product is positive as the two motions are in the same direction, whereas during releasing contact it becomes negative as the two motions are opposite, and during no contact and stable contact it is close to zero as the transient motion is nearly zero. Moreover, we discover that during sliding contact, the dot product exhibits clear positive-negative separation: one side is negative while the other side is positive. The negative region corresponds to the area where the gel is recovering, while the positive region corresponds to the area where the gel is deforming, and the transition from negative to positive reflects the sliding direction. This reveals that the dot product can explicitly characterize fine-grained contact states.

Furthermore, we analyze the relationship between the dot product magnitude during making contact and the contact force in Fig. 3(c). We apply different levels of force to the sensor and obtain the force magnitude using the sensor SDK. It is obvious that the dot product magnitude is positively correlated with the contact force, indicating that the dot product of two motions can not only reflect fine-grained differences in contact states but also capture the magnitude of contact force, providing richer tactile cues for contact-rich manipulation.

3.3 Unified yet Modality-aware Visuo-Tactile Fusion

Existing visuo-tactile fusion strategies typically follow two main paradigms: feature-level integration, which relies on direct concatenation or basic attention mechanisms, and decision-level aggregation, which involves training independent policies for each modality and subsequently averaging their outputs. However, neither strategy can simultaneously capture cross-modal complementarity and modality-specific properties. To address this, we take advantage of the Mixture-of-Transformers (MoT) architecture and propose a unified yet modality-aware fusion framework to effectively capture cross-modal complementarity and modality-specific properties. To our knowledge, we are the first to introduce the MoT architecture for visuo-tactile fusion.

Multimodal Observation Space. At each time step t, the robot perceives raw sensory inputs $\mathcal{O}_t = (I_t, T_0, T_t, T_{t-1}, p_t)$. Here, $I_t = \{I_t^{\text{wrist}}, I_t^{\text{third}}\}$ comprises the RGB images from the wrist and third-view cameras; T_0, T_t, and T_{t-1} represent the initial non-contact, current, and previous tactile frames, respectively; and p_t denotes the robot's proprioceptive state (*e.g.*, end-effector pose and gripper width). I_t and T_t are processed through pre-trained encoders to extract visual embeddings $\mathbf{e}_t^I$ and tactile embeddings $\mathbf{e}_t^T$. Concurrently, we compute the transient and cumulative tactile motions using (T_0, T_t, T_{t-1}). The resulting tactile motion correlation representation $Corr_t$, derived as detailed in Section 3.2, is passed through a ResNet encoder to yield the motion correlation embedding

$\mathbf{e}_t^{Corr}$. Finally, the proprioceptive state p_t is projected into a latent space via an MLP encoder to obtain the proprioceptive embedding $\mathbf{e}_t^p$.

Mixture-of-Transformers Multimodal Fusion. Given the modality-specific embeddings $\{\mathbf{e}_t^I, \mathbf{e}_t^T, \mathbf{e}_t^{Corr}, \mathbf{e}_t^p\}$, we adopt an MoT architecture to perform a unified yet modality-aware fusion. In this framework, tactile motion correlation is treated as a distinct modality from tactile appearance to reflect their heterogeneous characteristics. For each modality $m \in \{I, T, Corr, p\}$, modality-specific projection matrices are applied to compute the query, key, and value vectors:

$$\mathbf{Q}_t^m = \mathbf{e}_t^m W_Q^{(m)}, \quad \mathbf{K}_t^m = \mathbf{e}_t^m W_K^{(m)}, \quad \mathbf{V}_t^m = \mathbf{e}_t^m W_V^{(m)}, \tag{3}$$

where $\{W_Q^{(m)}, W_K^{(m)}, W_V^{(m)}\}$ denote learnable parameters unique to each modality. Then, all modality embeddings are jointly processed through a shared self-attention mechanism to facilitate cross-modal interactions:

$$\mathbf{A}_t = \text{softmax}\left(\frac{\mathbf{Q}_t \mathbf{K}_t^\top}{\sqrt{d}}\right)\mathbf{V}_t. \tag{4}$$

The resulting representations are further refined through modality-specific output projections, feed-forward networks (FFN), and layer normalization (LN):

$$\mathbf{h}_t^m = \mathbf{e}_t^m + \text{LN}_{\text{attn}}^{(m)}\left(\mathbf{A}_t^m W_O^{(m)}\right), \tag{5}$$

$$\mathbf{y}_t^m = \mathbf{h}_t^m + \text{LN}_{\text{ffn}}^{(m)}\left(\text{FFN}^{(m)}(\mathbf{h}_t^m)\right), \tag{6}$$

where $W_O^{(m)}$, $\text{FFN}^{(m)}$, and $\text{LN}^{(m)}$ are all modality-specific. This design allows the model to capture rich cross-modal interaction while preserving modality-specific properties. Finally, the fused embeddings are aggregated and fed into a Transformer-based decoder to autoregressively predict the action chunk.

4 Experimental Results

4.1 Experimental Setup

Training Details. Our model is trained using a β-VAE objective to model the stochasticity in human demonstrations. The overall loss function is defined as:

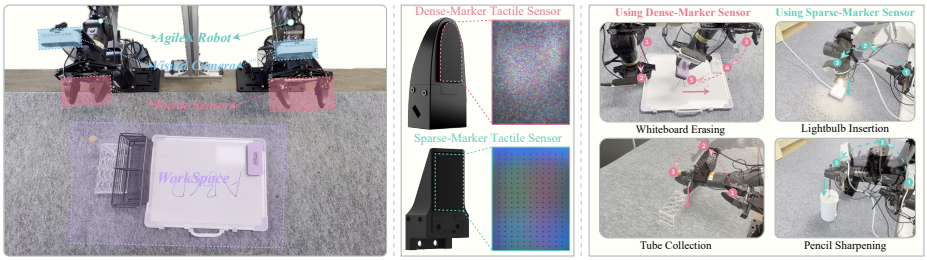
$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \beta \mathcal{L}_{\text{KL}}. \tag{7}$$

The prediction loss employs an L1 loss for precise action prediction:

$$\mathcal{L}_{\text{pred}} = \left\|\hat{a}_{t:t+k} - a_{t:t+k}\right\|_1, \tag{8}$$

where $\hat{a}_{t:t+k}$ denotes the predicted action chunk and $a_{t:t+k}$ denotes the ground-truth action sequence. The KL loss regularizes the posterior distribution

$$q_\phi(z \mid a_{t:t+k}, o_t) \tag{9}$$



(a) Real world setup (b) Optical tactile sensors (c) Contact-rich manipulation tasks

Fig. 4: Real-world experimental setup. (a) An Agilex robot equipped with visual cameras (wrist and third-view) and optical tactile sensors. (b) Two kinds of representative optical tactile sensors: one with random-distributed dense markers [1, 2] and one with regularly-arranged sparse markers [3]. (c) Four contact-rich manipulation tasks: cylinder collection, whiteboard erasing, lightbulb insertion, and pencil sharpening.

toward a standard Gaussian prior:

$$\mathcal{L}_{\mathrm{KL}} = D_{\mathrm{KL}}\left(q_\phi(z \mid a_{t:t+k}, o_t) \parallel \mathcal{N}(0, I)\right). \tag{10}$$

The hyperparameter β modulates the strength of the information bottleneck.The model is trained from scratch for each task using the Adam optimizer [36]. Training takes approximately 5 hours per task on a single RTX 3090 Ti GPU. The batch size is set to 16, and the learning rate is initialized at 2×10^{-4} with a cosine decay schedule. More implementation details are provided in the appendix.

Hardware. Our robotic system consists of two 6-DoF Agilex robotic arms, each with a 1-DoF gripper mounted on its end-effector. One third-view camera and two wrist-view cameras provide RGB observations for global visual perception. For local tactile sensing, we employ we employ two kinds of representative tactile sensors: one with random-distributed dense markers [1,2] and one with regularly-arranged sparse markers [3]. The overall system is shown in Fig. 4(a), and the tactile sensors are presented in Fig. 4(b).

Tasks and Evaluation Criteria. We evaluate our ViTacMotor on four challenging real-world contact-rich manipulation tasks. Representative execution results for each task are shown in Fig. 5. Below are the detailed descriptions and evaluation criteria for each task:

(1) Tasks Requiring Precise Contact State Information

Tube Collection *(Using Dense-Marker Sensor)*: The robot needs to pick up a transparent tube from a bucket and place it into a transparent rack. The task must be executed carefully to prevent breakage, with precise alignment to the correct slot and successful insertion through the two holes. *Evaluation Criteria*: A trial is considered successful if the tube is placed in the rack without damage.

Lightbulb Insertion *(Using Sparse-Marker Sensor)*: The robot needs to rotate a USB-type lightbulb to the correct orientation, move it toward the socket, and insert it until the light turns on. The task requires precise alignment and adjustment during contact to ensure correct insertion. *Evaluation Criteria*: A trial is successful if the lightbulb is correctly inserted into the socket and illuminates.

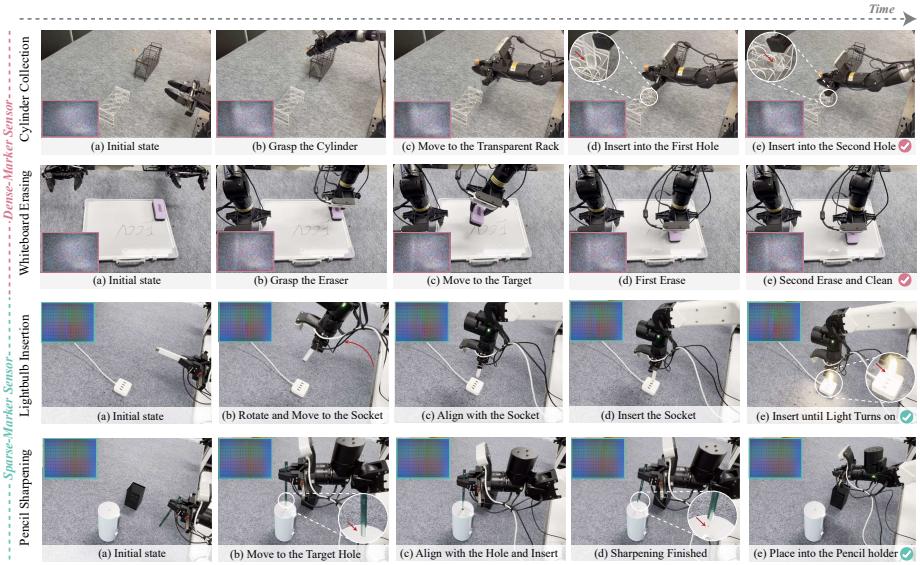


Fig. 5: Qualitative results of policy execution. We evaluate our visuo-tactile policy using two kinds of sensors with various marker patterns on four contact-rich manipulation tasks: tube collection, whiteboard erasing, lightbulb insertion, and pencil sharpening.

(2) Tasks Requiring fine-grained force Control

Whiteboard Erasing *(Using Dense-Marker Sensor)*: The robot needs to pick up a eraser and then remove the writing from a whiteboard. It must apply appropriate pressure to effectively erase the marker ink while avoiding excessive force that could potentially damage the system. The task requires stable and well-controlled contact force throughout. *Evaluation Criteria:* The task is considered successful if all visible ink on the whiteboard is completely removed.

Pencil Sharpening *(Using Sparse-Marker Sensor)*: The robot needs to insert a pencil into a sharpener to sharpen it, and then place it into a holder. The task requires precise alignment during contact, along with sufficient force to ensure effective sharpening. *Evaluation Criteria*: A trial is considered successful if the pencil is successfully sharpened and then placed into the pencil holder.

Baselines. We compare our method with the following open-source baselines:

- *Action Chunking with Transformers (ACT)* [70]: A transformer-based visual policy that relies only on visual images and robot states.
- *ACT + Tactile Image (ACT+T)*: The ACT policy augmented with tactile images as input. The tactile images are processed by a ResNet-18 encoder and then fused with visual features using a concatenation operation.
- *Diffusion Policy (DP)* [15]: A diffusion-based visual policy that relies only on visual images and robot states.
- *Policy Consensus* [10]: A diffusion-based visuo-tactile policy that fuses vision and touch at the decision level.

Table 1: Quantitative Comparison of success rates (%) with Baselines. We evaluate our policy over 15 episodes, and the best performance is highlighted in bold. The numbers in parentheses denote the number of training demonstrations.

Tasks Requiring In-Hand State Information							
Methods	**Tube Collection (35 demos)**				**Lightbulb Insertion (60 demos)**		
	Grasp Tube	**Insert 1st Hole**	**Insert 2nd Hole**	**Whole Task**	**Align with Socket**	**Insert & Light on**	**Whole Task**
ACT*	66.7	60.0	53.3	53.3	**60.0**	13.3	13.3
DP*	73.3	53.3	40.0	40.0	53.3	6.7	6.7
ACT + T	86.7	60.0	60.0	60.0	**60.0**	26.7	26.7
Policy Consensus	80.0	53.3	53.3	53.3	53.3	26.7	26.7
TactileACT	**93.3**	73.3	60.0	60.0	**60.0**	**40.0**	**40.0**
ViTacMotor	**93.3**	**80.0**	**73.3**	**73.3**	**60.0**	**40.0**	**40.0**
Tasks Requiring Fine-Grained Force Control							
Methods	**Whiteboard Erasing (50 demos)**				**Pencil Sharpening (40 demos)**		
	Grasp Eraser	**First Erase**	**Second Erase**	**Whole Task**	**Insert & Sharpening**	**Put into holder**	**Whole Task**
ACT*	**100.0**	60.0	46.7	46.7	40.0	33.3	33.3
DP*	**100.0**	66.7	40.0	40.0	33.3	33.3	33.3
ACT + T	**100.0**	80.0	60.0	60.0	40.0	40.0	40.0
Policy Consensus	93.3	73.3	66.7	66.7	33.3	33.3	33.3
TactileACT	**100.0**	**86.7**	73.3	73.3	46.7	46.7	46.7
ViTacMotor	**100.0**	**86.7**	**86.7**	**86.7**	**60.0**	**60.0**	**60.0**

• *TactileACT* [20]: A transformer-based visuo-tactile policy that fuses vision and touch at the feature level using a concatenate operation.

For each of the four tasks, we conduct 15 trials for each method and report the average success rate in Table 1.

4.2 Comparison Results

Comparison on Tasks Requiring In-Hand State Information. In Table 1, we compare ViTacMotor with several baselines on tasks that require precise contact state information. Visual-only policies perform poorly, as visual inputs are often severely occluded. For example, in the tube collection task, when the gripper holds the tube above the rack, the third-person view is blocked by the robotic arm and gripper, making it difficult to infer the current contact state or even determine whether the tube has been inserted into the hole. This ambiguity confuses vision-based policies and degrades insertion accuracy. A similar issue arises in the lightbulb insertion task. Although policies augmented with raw tactile images show some improvement, they still lack robustness, as fine-grained contact states are difficult to directly interpret from raw tactile appearance. In contrast, our method explicitly perceives fine-grained contact states and achieves superior performance on these tasks.

Table 2: Ablation study of the proposed ViTacMotor.

Base	TMC	MoT-Fusion	White. Erasing	Tube Collect.
✓			46.7	53.3
✓	✓		73.3	66.7
✓		✓	66.7	60.0
✓	✓	✓	**86.7**	**73.3**

Table 3: Effectiveness of TMC embedded into existing visual policies.

Methods	ACT [70]	DP [15]
Raw Image	60.0	53.3
Cumu. Motion	53.3	40.0
TMC (Ours)	**73.3**	**66.7**

Table 4: Effectiveness of MoT-based fusion strategy compared to others.

Fusion Strategy	White. Erasing	Tube Collect.	Avg.
Concat.	73.3	66.7	70.0
Cross Att.	73.3	60.0	66.7
Ours	**86.7**	**73.3**	**80.0**

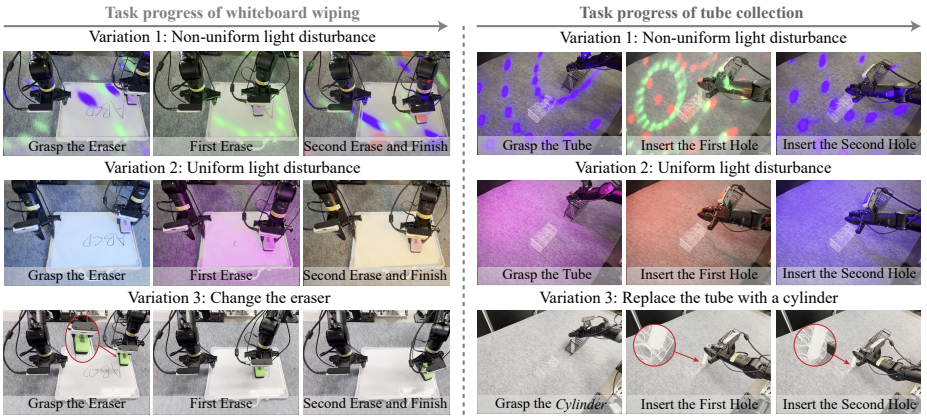


Fig. 6: Robustness to environment and object variations. We evaluate the robustness of our method under both environment and object variations using the whiteboard erasing and tube collection tasks. For environment variations, we introduce temporal varying spatially non-uniform and uniform lighting conditions. For object variations, we change the eraser and ink appearance in the whiteboard erasing task, and replace the tube with a cylinder in the tube collection task.

Comparison on Tasks Requiring Fine-Grained Force Control. We further compare on tasks that demand precise force control in Table 1. Visual-only policies struggle in such scenarios, as visual observations alone cannot provide feedback on contact forces, thereby limiting precise force modulation. For instance, in the whiteboard erasing task, the robot must apply appropriate pressure to effectively remove ink. However, vision-based methods cannot perceive contact forces and therefore fail to achieve complete erasure. In the pencil sharpening task, when the pencil is misaligned with the hole during insertion, visual inputs cannot detect the resulting shear forces, making fine corrective adjustments difficult. Although existing visuo-tactile policies achieve comparable performance, their lack of explicit force modeling limits stable force control. In contrast, our proposed TMC method not only captures fine-grained contact states but also implicitly encodes contact force magnitudes, leading to superior performance.

4.3 Ablation and Discussion

How does TMC Facilitate Contact-rich Manipulation? We further validate the importance of TMC in contact-rich manipulation tasks. As shown in

Table 2, removing TMC leads to a performance drop in whiteboard erasing and tube collection. For whiteboard erasing, the absence of TMC hinders the model's ability to implicitly infer force magnitudes, leading to inconsistent erasing pressure. For tube collection, without TMC, the model struggles to perceive fine-grained contact states leading to misalignment during insertion. These results demonstrate the effectiveness of TMC in facilitating contact-rich manipulation.

How does MoT-based Strategy Improve Visuo-Tactile Manipulation? Table 2 shows that replacing our MoT-based strategy with a simple concatenation operation leads to a performance drop, which validates the effectiveness of our MoT-based fusion framework in effectively integrating global visual perception and local tactile sensing for visuo-tactile manipulation tasks.

Universality of TMC across Different Optical Sensors. We further discuss the cross-sensor universality of TMC. Our experiments use Our experiments use two kinds of tactile sensors with different marker patterns [1–3]. Table 1 shows that our method achieves superior performance with various optical tactile sensors, validating the universality of TMC. This advantage primarily stems from our motion-aware tactile representation, which prioritizes underlying gel deformation over raw visual appearance.

Can TMC be Flexibly Embedded into Existing Policies? To further validate the flexibility and effectiveness of TMC, we embed existing representations (raw appearance and cumulative motion) and TMC into existing visual policies: ACT [70] and DP [15] on the whiteboard erasing task. Table 3 shows that existing methods obtain better performance after integrating the TMC, further demonstrating the effectiveness of TMC.

Effectiveness of MoT-based Visuo-Tactile Fusion Strategy. We further study the effectiveness of our MoT-based fusion strategy by replacing it with several commonly used fusion strategies, including simple concatenation, cross attention. As shown in Table 4, our MoT-based fusion strategy outperforms these commonly used fusion strategies, which validates the effectiveness of our MoT-based fusion framework in effectively integrating global visual perception and local tactile sensing for visuo-tactile manipulation tasks.

Robustness to Environment and Object Variations. To evaluate the robustness of our method, we introduce various disturbances to the environment, including different temporal-varying lighting disturbances (e.g., spatial uniform and non-uniform) and object variations (e.g., different colored eraser and replace the tube with a cylinder). Figure 6 shows that our method still successfully completes the tasks under these disturbances, demonstrating its robustness to environmental and object variations.

Limitations. Though our method achieves fine-grained contact state perception and effective visuo-tactile fusion for contact-rich manipulation, its generalization remains limited under extreme visual disturbances or object position variations. Failure cases under such conditions are provided in the Appendix. In future work, we will focus on improving the generalization of contact-rich manipulation policies in complex environments.

5 Conclusion

We reveal that the correlation between transient and cumulative tactile motion provides explicit discrimination of fine-grained contact states. This insight motivates us to propose a motion-aware tactile representation to facilitate contact-rich manipulation. We introduce a unified modality-specific visuo-tactile policy that captures cross-modal complementarity while preserving the unique characteristics of each modality—thereby enabling the effective integration of global visual perception and local tactile feedback. Extensive experiments on a diverse range of tasks demonstrate the superior performance of our proposed method.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62472098), the Science and Technology Commission of Shanghai Municipality (No. 24511103100) and the New Cornerstone Science Foundation through the XPLORER PRIZE.

References

1. Daimon optical tactile sensor, dm-tac w2, https://www.dmrobot.com/product/p1/dm-tacw2.html, DM-Tac W2 2, 6, 10, 14
2. Neote ai optical tactile sensor, intac s1, https://www.neoteai.com/, InTac S1 2, 6, 10, 14
3. Xense optical tactile sensor, xensesensor, https://www.xenserobotics.com/product/367/detail/9, XenseSensor 2, 6, 10, 14
4. Abdi, H., Williams, L.J.: Principal component analysis. Wiley interdisciplinary reviews: computational statistics **2**(4), 433–459 (2010) 4
5. Bhirangi, R., Hellebrekers, T., Majidi, C., Gupta, A.: Reskin: versatile, replaceable, lasting tactile skins. arXiv preprint arXiv:2111.00071 (2021) 4
6. Bi, H., Tan, H., Xie, S., Wang, Z., Huang, S., Liu, H., Zhao, R., Feng, Y., Xiang, C., Rong, Y., et al.: Motus: A unified latent action world model. arXiv preprint arXiv:2512.13030 (2025) 5
7. Bi, J., Ma, K.Y., Hao, C., Shou, M.Z., Soh, H.: Vla-touch: Enhancing vision-language-action models with dual-level tactile feedback. arXiv preprint arXiv:2507.17294 (2025) 4
8. van den Bogert, W., Iyengar, M., Fazeli, N.: Built different: Tactile perception to overcome cross-embodiment capability differences in collaborative manipulation. arXiv e-prints pp. arXiv–2409 (2024) 2, 4
9. Cai, J., Cai, Z., Cao, J., Chen, Y., He, Z., Jiang, L., Li, H., Li, H., Li, Y., Liu, Y., et al.: Internvla-a1: Unifying understanding, generation and action for robotic manipulation. arXiv preprint arXiv:2601.02456 (2026) 5
10. Chen, H., Xu, J., Chen, H., Hong, K., Huang, B., Liu, C., Mao, J., Li, Y., Du, Y., Driggs-Campbell, K.: Multi-modal manipulation via multi-modal policy consensus. arXiv preprint arXiv:2509.23468 (2025) 3, 5, 11
11. Chen, W., Xue, H., Wang, Y., Zhou, F., Lv, J., Jin, Y., Tang, S., Wen, C., Lu, C.: Implicitrdp: An end-to-end visual-force diffusion policy with structural slow-fast learning. arXiv preprint arXiv:2512.10946 (2025) 5

12. Chen, X., Chu, F.J., Gleize, P., Liang, K.J., Sax, A., Tang, H., Wang, W., Guo, M., Hardin, T., Li, X., et al.: Sam 3d: 3dfy anything in images. arXiv preprint arXiv:2511.16624 (2025) 5
13. Chen, Y., Sipos, A., Van der Merwe, M., Fazeli, N.: Visuo-tactile transformers for manipulation. arXiv preprint arXiv:2210.00121 (2022) 5
14. Cheng, Z., Zhang, Y., Zhang, W., Li, H., Wang, K., Song, L., Zhang, H.: Omnivtla: Vision-tactile-language-action model with semantic-aligned tactile sensing. arXiv preprint arXiv:2508.08706 (2025) 4
15. Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., Song, S.: Diffusion policy: Visuomotor policy learning via action diffusion. The International Journal of Robotics Research **44**(10-11), 1684–1704 (2025) 11, 13, 14
16. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025) 5
17. Feng, R., Hu, D., Ma, W., Li, X.: Play to the score: Stage-guided dynamic multi-sensory fusion for robotic manipulation. arXiv preprint arXiv:2408.01366 (2024) 5
18. Feng, R., Hu, J., Xia, W., Gao, T., Shen, A., Sun, Y., Fang, B., Hu, D.: Anytouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors. arXiv preprint arXiv:2502.12191 (2025) 4
19. Feng, R., Zhou, Y., Mei, S., Zhou, D., Wang, P., Cui, S., Fang, B., Yao, G., Hu, D.: Anytouch 2: General optical tactile representation learning for dynamic tactile perception. arXiv preprint arXiv:2602.09617 (2026) 4
20. George, A., Gano, S., Katragadda, P., Farimani, A.B.: Vital pretraining: Visuo-tactile pretraining for tactile and non-tactile manipulation policies. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 258–264. IEEE (2025) 4, 12
21. Gu, C., Liu, J., Chen, H., Huang, R., Wuwu, Q., Liu, Z., Li, X., Li, Y., Zhang, R., Jia, P., et al.: Manualvla: A unified vla model for chain-of-thought manual generation and robotic manipulation. arXiv preprint arXiv:2512.02013 (2025) 5
22. Gu, N., Kosuge, K., Hayashibe, M.: Tactilealoha: Learning bimanual manipulation with tactile sensing. IEEE Robotics and Automation Letters (2025) 4
23. He, Z., Fang, H., Chen, J., Fang, H.S., Lu, C.: Foar: Force-aware reactive policy for contact-rich robotic manipulation. IEEE Robotics and Automation Letters (2025) 5
24. Helmholtz, H.v.: Über integrale der hydrodynamischen gleichungen, welche den wirbelbewegungen entsprechen. (1858) 4
25. Higuera, C., Sharma, A., Bodduluri, C.K., Fan, T., Lancaster, P., Kalakrishnan, M., Kaess, M., Boots, B., Lambeta, M., Wu, T., et al.: Sparsh: Self-supervised touch representations for vision-based tactile sensing. arXiv preprint arXiv:2410.24090 (2024) 4
26. Hogan, F.R., Jenkin, M., Rezaei-Shoshtari, S., Girdhar, Y., Meger, D., Dudek, G.: Seeing through your skin: Recognizing objects with a novel visuotactile sensor. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1218–1227 (2021) 4
27. Huang, B., Wang, Y., Yang, X., Luo, Y., Li, Y.: 3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing. arXiv preprint arXiv:2410.24091 (2024) 4, 5
28. Huang, B., Xu, J., Akinola, I., Yang, W., Sundaralingam, B., O'Flaherty, R., Fox, D., Wang, X., Mousavian, A., Chao, Y.W., et al.: Vt-refine: Learning bimanual

assembly with visuo-tactile feedback via simulation fine-tuning. arXiv preprint arXiv:2510.14930 (2025) 5
29. Huang, J., Wang, S., Lin, F., Hu, Y., Wen, C., Gao, Y.: Tactile-vla: unlocking vision-language-action model's physical knowledge for tactile generalization. arXiv preprint arXiv:2507.09160 (2025) 5
30. Huang, W., Chen, C., Qi, H., Lv, C., Du, Y., Yang, H.: Motvla: A vision-language-action model with unified fast-slow reasoning. arXiv preprint arXiv:2510.18337 (2025) 5
31. Huang, Y., Lin, P., Li, W., Li, D., Li, J., Jiang, J., Xiao, C., Jiao, Z.: Tactile-force alignment in vision-language-action models for force-aware manipulation. arXiv preprint arXiv:2601.20321 (2026) 4
32. Jamone, L., Natale, L., Metta, G., Sandini, G.: Highly sensitive soft tactile sensors for an anthropomorphic robotic hand. IEEE sensors Journal **15**(8), 4226–4233 (2015) 4
33. Jiang, J., Zhang, X., Gomes, D.F., Do, T.T., Luo, S.: Rotipbot: Robotic handling of thin and flexible objects using rotatable tactile sensors. IEEE Transactions on Robotics (2025) 4
34. Jin, W., Niu, Y., Liao, J., Duan, C., Li, A., Gao, S., Liu, X.: Srum: Fine-grained self-rewarding for unified multimodal models. arXiv preprint arXiv:2510.12784 (2025) 5
35. Kang, X., Tian, T., Lee, S.W., Huang, B., Li, Y., Kuo, Y.L.: Learning force-regulated manipulation with a low-cost tactile-force-controlled gripper. arXiv preprint arXiv:2602.10013 (2026) 4
36. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 10
37. Kroeger, T., Timofte, R., Dai, D., Van Gool, L.: Fast optical flow using dense inverse search. In: European conference on computer vision. pp. 471–488. Springer (2016) 6
38. Lee, G., Lee, Y., Kim, K., Lee, S., Noh, S., Back, S., Lee, K.: Manipforce: Force-guided policy learning with frequency-aware representation for contact-rich manipulation. arXiv preprint arXiv:2509.19047 (2025) 5
39. Li, H., Zhang, Y., Zhu, J., Wang, S., Lee, M.A., Xu, H., Adelson, E., Fei-Fei, L., Gao, R., Wu, J.: See, hear, and feel: Smart sensory fusion for robotic manipulation. arXiv preprint arXiv:2212.03858 (2022) 3, 5
40. Li, J., Wu, T., Zhang, J., Chen, Z., Jin, H., Wu, M., Shen, Y., Yang, Y., Dong, H.: Adaptive visuo-tactile fusion with predictive force attention for dexterous manipulation. arXiv preprint arXiv:2505.13982 (2025) 5
41. Li, L., Zhang, Q., Luo, Y., Yang, S., Wang, R., Han, F., Yu, M., Gao, Z., Xue, N., Zhu, X., et al.: Causal world modeling for robot control. arXiv preprint arXiv:2601.21998 (2026) 5
42. Li, S., Wang, Z., Wu, C., Li, X., Luo, S., Fang, B., Sun, F., Zhang, X.P., Ding, W.: When vision meets touch: A contemporary review for visuotactile sensors from the signal processing perspective. IEEE Journal of Selected Topics in Signal Processing **18**(3), 267–287 (2024) 4
43. Li, Y., Chen, Y., Zhao, Z., Li, P., Liu, T., Huang, S., Zhu, Y.: Simultaneous tactile-visual perception for learning multimodal robot manipulation. arXiv preprint arXiv:2512.09851 (2025) 4
44. Liang, W., Yu, L., Luo, L., Iyer, S., Dong, N., Zhou, C., Ghosh, G., Lewis, M., Yih, W.t., Zettlemoyer, L., et al.: Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. arXiv preprint arXiv:2411.04996 (2024) 3, 5

45. Lin, C., Lin, Z., Wang, S., Xu, H.: Dtact: A vision-based tactile sensor that measures high-resolution 3d geometry directly from darkness. arXiv preprint arXiv:2209.13916 (2022) 4
46. Lin, C., Zhang, H., Xu, J., Wu, L., Xu, H.: 9dtact: A compact vision-based tactile sensor for accurate 3d shape reconstruction and generalizable 6d force estimation. IEEE Robotics and Automation Letters **9**(2), 923–930 (2023) 4
47. Liu, F., Li, C., Qin, Y., Xu, J., Abbeel, P., Chen, R.: Vitamin: Learning contact-rich tasks through robot-free visuo-tactile manipulation interface. arXiv preprint arXiv:2504.06156 (2025) 4
48. Liu, F., Deswal, S., Christou, A., Sandamirskaya, Y., Kaboli, M., Dahiya, R.: Neuro-inspired electronic skin for robots. Science robotics **7**(67), eabl7344 (2022) 4
49. Liu, F., Deswal, S., Christou, A., Shojaei Baghini, M., Chirila, R., Shakthivel, D., Chakraborty, M., Dahiya, R.: Printed synaptic transistor–based electronic skin for robots to feel and learn. Science Robotics **7**(67), eabl7286 (2022) 4
50. Liu, J.J., Li, Y., Shaw, K., Tao, T., Salakhutdinov, R., Pathak, D.: Factr: Force-attending curriculum training for contact-rich policy learning. arXiv preprint arXiv:2502.17432 (2025) 5
51. Liu, Z., Liu, J., Xu, J., Han, N., Gu, C., Chen, H., Zhou, K., Zhang, R., Hsieh, K.C., Wu, K., et al.: Mla: A multisensory language-action model for multimodal understanding and forecasting in robotic manipulation. arXiv preprint arXiv:2509.26642 (2025) 5
52. Luo, H., Wang, Y., Zhang, W., Zheng, S., Xi, Z., Xu, C., Xu, H., Yuan, H., Zhang, C., Wang, Y., et al.: Being-h0. 5: Scaling human-centric robot learning for cross-embodiment generalization. arXiv preprint arXiv:2601.12993 (2026) 5
53. Luo, S., Lepora, N.F., Yuan, W., Althoefer, K., Cheng, G., Dahiya, R.: Tactile robotics: An outlook. IEEE Transactions on Robotics (2025) 4
54. Ren, J., Zou, J., Gu, G.: Mc-tac: Modular camera-based tactile sensor for robot gripper. In: International Conference on Intelligent Robotics and Applications. pp. 169–179. Springer (2023) 4
55. Taylor, I.H., Dong, S., Rodriguez, A.: Gelslim 3.0: High-resolution measurement of shape, force and slip in a compact tactile-sensing finger. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 10781–10787. IEEE (2022) 4
56. Wang, X., Zhang, Z., Zhang, H., Lin, Z., Zhou, Y., Liu, Q., Zhang, S., Li, Y., Liu, S., Zheng, H., et al.: Hbridge: H-shape bridging of heterogeneous experts for unified multimodal understanding and generation. arXiv preprint arXiv:2511.20520 (2025) 5
57. Ward-Cherrier, B., Pestell, N., Cramphorn, L., Winstone, B., Giannaccini, M.E., Rossiter, J., Lepora, N.F.: The tactip family: Soft optical tactile sensors with 3d-printed biomimetic morphologies. Soft robotics **5**(2), 216–227 (2018) 4
58. Wu, L., Yu, C., Ren, J., Chen, L., Jiang, Y., Huang, R., Gu, G., Li, H.: Freetacman: Robot-free visuo-tactile data collection system for contact-rich manipulation. arXiv preprint arXiv:2506.01941 (2025) 2, 4
59. Wu, T., Li, J., Zhang, J., Wu, M., Dong, H.: Canonical representation and force-based pretraining of 3d tactile for dexterous visuo-tactile policy learning. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 6786–6792. IEEE (2025) 3, 5
60. Wu, W., Lu, F., Wang, Y., Yang, S., Liu, S., Wang, F., Zhu, Q., Sun, H., Wang, Y., Ma, S., et al.: A pragmatic vla foundation model. arXiv preprint arXiv:2601.18692 (2026) 5

61. Xu, Y., Wei, L., An, P., Zhang, Q., Li, Y.L.: exumi: Extensible robot teaching system with action-aware task-agnostic tactile representation. arXiv preprint arXiv:2509.14688 (2025) 4
62. Xue, H., Ren, J., Chen, W., Zhang, G., Fang, Y., Gu, G., Xu, H., Lu, C.: Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation. arXiv preprint arXiv:2503.02881 (2025) 2, 4
63. Yamaguchi, A., Atkeson, C.G.: Implementing tactile behaviors using fingervision. In: 2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids). pp. 241–248. IEEE (2017) 4
64. Yu, J., Liu, H., Yu, Q., Ren, J., Hao, C., Ding, H., Huang, G., Huang, G., Song, Y., Cai, P., et al.: Forcevla: Enhancing vla models with a force-aware moe for contact-rich manipulation. arXiv preprint arXiv:2505.22159 (2025) 5
65. Yu, K., Han, Y., Wang, Q., Saxena, V., Xu, D., Zhao, Y.: Mimictouch: Leveraging multi-modal human tactile demonstrations for contact-rich manipulation. arXiv preprint arXiv:2310.16917 (2023) 4
66. Yuan, W., Dong, S., Adelson, E.H.: Gelsight: High-resolution robot tactile sensors for estimating geometry and force. Sensors **17**(12), 2762 (2017) 2, 4
67. Zhang, C., Hao, P., Cao, X., Hao, X., Cui, S., Wang, S.: Vtla: Vision-tactile-language-action model with preference learning for insertion manipulation. arXiv preprint arXiv:2505.09577 (2025) 4
68. Zhang, J., Yao, H., Mo, J., Chen, S., Xie, Y., Ma, S., Chen, R., Luo, T., Ling, W., Qin, L., et al.: Finger-inspired rigid-soft hybrid tactile sensor with superior sensitivity at high frequency. Nature communications **13**(1), 5076 (2022) 4
69. Zhang, Z., Ma, J., Yang, X., Wen, X., Zhang, Y., Li, B., Qin, Y., Liu, J., Zhao, C., Kang, L., et al.: Touchguide: Inference-time steering of visuomotor policies via touch guidance. arXiv preprint arXiv:2601.20239 (2026) 4
70. Zhao, T.Z., Kumar, V., Levine, S., Finn, C.: Learning fine-grained bimanual manipulation with low-cost hardware. arXiv preprint arXiv:2304.13705 (2023) 11, 13, 14
71. Zhao, T.Z., Tompson, J., Driess, D., Florence, P., Ghasemipour, K., Finn, C., Wahid, A.: Aloha unleashed: A simple recipe for robot dexterity. arXiv preprint arXiv:2410.13126 (2024) 3
72. Zhao, Z., Haldar, S., Cui, J., Pinto, L., Bhirangi, R.: Touch begins where vision ends: Generalizable policies for contact-rich manipulation. arXiv preprint arXiv:2506.13762 (2025) 5
73. Zhu, X., Huang, B., Li, Y.: Touch in the wild: Learning fine-grained manipulation with a portable visuo-tactile gripper. arXiv preprint arXiv:2507.15062 (2025) 5
74. Zhu, Y., Ye, Z., Hu, B., Zhao, H., Qi, Y., Wang, D., Platt, R.: Residual rotation correction using tactile equivariance. arXiv preprint arXiv:2511.07381 (2025) 4