

TierTac: Physically-Grounded Hierarchical Reasoning for Generalizable Visuo-Tactile Perception

¹Shipeng Lyu, ²Lijie Sheng, ³Wenyao Zhang, ⁴Chen Yao, ⁵Zheng Li, ²Zhenzhong Jia
^{1,*}David Navarro-Alarcon and ^{6,*}Isabel Guan

¹The Hong Kong Polytechnic University, Hong Kong

²The Southern University of Science and Technology, China

³The Shanghai Jiao Tong University, China

⁴The Harbin Institute of Technology, China

⁵The Chinese University of Hong Kong, Hong Kong

⁶The Hong Kong University of Science and Technology, Hong Kong

*Corresponding author.

Keywords: *Robotics, visuo-tactile sensors, sensor-agnostic, contact reasoning.*

Visuo-tactile perception is essential for intelligent robotic systems interacting with the physical world, yet achieving reliable understanding across heterogeneous sensors remains challenging. Existing approaches rely on representation learning or independent attribute prediction, often neglecting the physically grounded hierarchical dependencies underlying contact interactions. Here, we propose TierTac, a physically grounded hierarchical reasoning framework for generalizable visuo-tactile perception. TierTac organizes tactile understanding into a physically grounded causal hierarchy of three levels: intrinsic physical properties, compliance responses, and contact states. It combines contact-aware sensor-agnostic representation learning with hierarchical neuro-symbolic reasoning to enable physically consistent multi-attribute inference. Experiments across multiple datasets demonstrate improved cross-sensor alignment, physically consistent prediction, and generalization to unseen sensors. Real-world robotic evaluations further show more stable and physically consistent interaction under both classical control and learned policies. These results highlight physically grounded hierarchical reasoning as an effective principle for building generalizable tactile intelligence in robotic systems.

1 Introduction

Visuo-tactile perception is a fundamental capability for intelligent robotic systems that interact with the physical world. By converting contact-induced deformations into high-resolution images, recent visuo-tactile sensors, such as DIGIT, provide rich tactile cues about object properties, compliance, and contact states, enabling robots to perform contact-rich manipulation more effectively [1, 2]. However, achieving reliable tactile understanding across heterogeneous sensors remains challenging. Visuo-tactile sensors often differ in sensing mechanisms, elastomer materials, geometries, illumination conditions, and imaging characteristics, causing the same physical contact to produce substantially different observations across devices [3]. Such sensor-dependent variations introduce a perceptual gap that hinders the deployment of tactile perception models across robotic platforms with various visuo-tactile sensors. This reveals a broader challenge: generalizable visuo-tactile perception requires models that capture

essential contact semantics while maintaining physical consistency beyond sensor-specific appearance.

To address cross-sensor discrepancies, a widely adopted paradigm is to learn shared representations for multi-sensor tactile data [4–6]. These approaches aim to align tactile observations across sensors into a common embedding space, improving cross-sensor correspondence. However, such representation learning approaches remain fundamentally limited, as they primarily focus on feature alignment while overlooking the physically grounded causal structure of tactile interactions. In particular, tactile perception is often treated as either embedding similarity or independent attribute prediction, without explicitly modeling how different attributes are physically coupled through contact. As a result, even well-aligned representations may produce fragmented or physically inconsistent predictions, such as contradictory inferences between deformation and force, since they lack mechanisms to enforce coherence across attributes. Moreover, the learned embeddings may still retain substantial sensor-specific artifacts, and prior work often emphasizes cross-modal alignment or text-anchored semantics [7] with limited modeling of fine-grained contact information [8]. This limitation suggests that achieving generalizable visuo-tactile perception requires moving beyond statistical representation alignment toward modeling the intrinsic physical dependencies of contact.

More fundamentally, tactile perception is not a collection of decoupled attribute predictions, but a structured physical reasoning process governed by contact interactions. Tactile attributes exhibit hierarchical dependencies: physical properties (e.g., material, hardness) determine an object’s response under contact; these responses manifest as compliance-related behaviors (e.g., local deformation), which in turn govern the resulting contact states (e.g., slippage, rotation, and force). This hierarchy provides a physically grounded organization of tactile knowledge, reflecting causal relationships from object-level properties to interaction-level states. However, most existing tactile inference methods model these attributes separately, focusing on isolated static properties or task-specific contact cues [6, 9–11]. Ignoring such underlying physical constraints leads to fragmented and inconsistent predictions, limiting both interpretability and transferability. Therefore, achieving generalizable visuo-tactile perception requires explicitly modeling these hierarchical dependencies as a key inductive bias for coherent and generalizable tactile perception.

Building on this insight, we propose TierTac, a physically grounded hierarchical reasoning framework for generalizable visuo-tactile perception. Rather than treating tactile understanding as decoupled attribute prediction or representation alignment, TierTac explicitly models tactile perception as a principled reasoning process over hierarchical

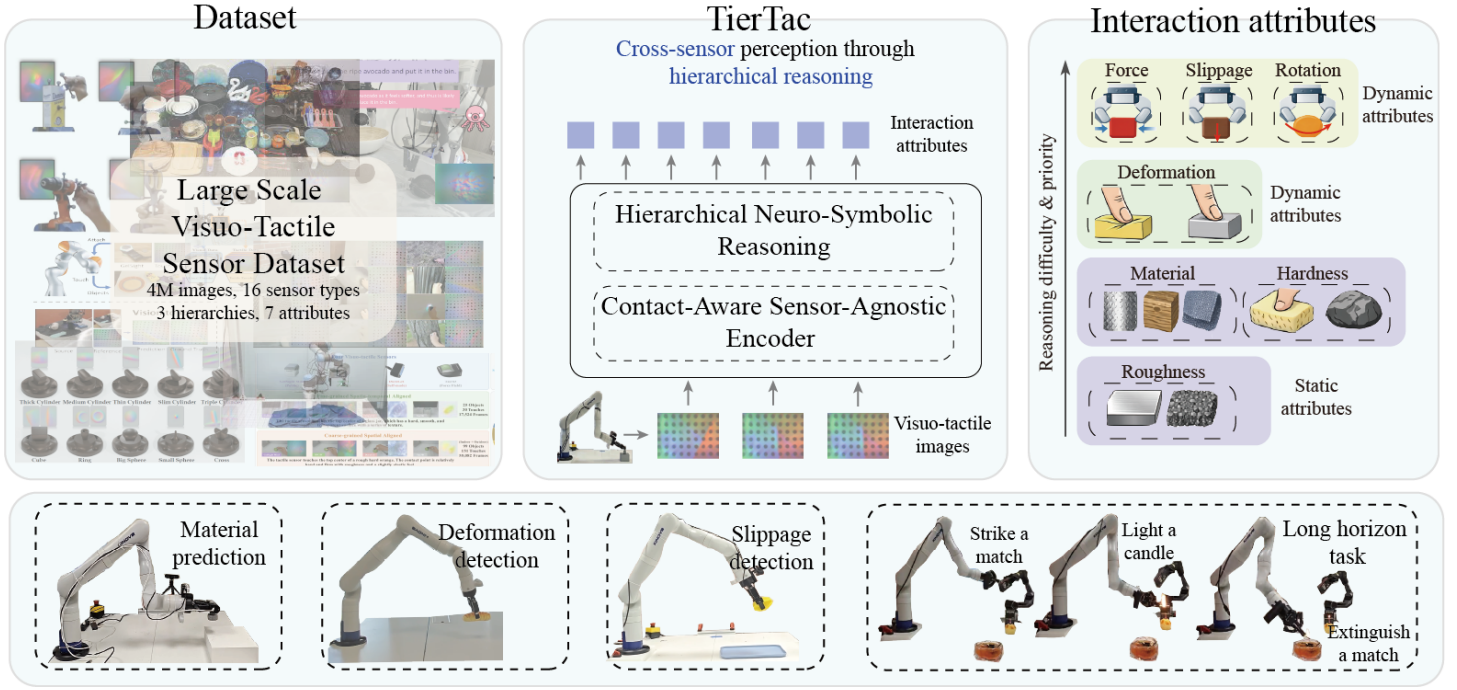


Figure 1: Overview of TierTac. The framework models visuo-tactile perception as a physically grounded hierarchical reasoning process, learning contact-aware sensor-agnostic representations and performing hierarchical neuro-symbolic reasoning over tactile attributes organized from physical properties to compliance responses and contact states. This structured formulation enables physically consistent multi-attribute inference across heterogeneous sensors, supporting complex robotic interaction.

physical dependencies. Specifically, TierTac organizes tactile attributes into three interconnected levels, i.e., physical properties, compliance responses, and contact states, and enables structured information propagation across these levels to enforce physically consistent inference. To realize this, TierTac integrates contact-aware sensor-agnostic representation learning with hierarchical neuro-symbolic reasoning. The representation module captures sensor-agnostic contact semantics while mitigating sensor-specific artifacts, whereas the reasoning module explicitly models structured dependencies among tactile attributes to ensure coherence across predictions. Through this design, TierTac provides a unified framework that bridges physical interaction principles with perceptual inference, enabling consistent and transferable visuo-tactile understanding across heterogeneous sensors. Our main contributions are summarized as follows:

- We identify physically grounded hierarchical dependencies as a missing inductive bias for physically consistent visuo-tactile perception, shifting tactile understanding from decoupled attribute prediction toward structured contact reasoning.
- We propose TierTac, a physically grounded hierarchical reasoning framework that integrates contact-aware sensor-agnostic representation learning with hierarchical neuro-symbolic reasoning to explicitly model dependencies among physical properties, compliance responses, and contact states.
- We demonstrate that TierTac improves cross-sensor alignment, physically consistent multi-attribute inference, and generalization to unseen hardware across multiple visuo-

tactile datasets, while enabling more stable and force-efficient robotic interaction in real-world manipulation tasks.

2 Related work

Recent advances in visuo-tactile perception have improved robotic interaction through transferable representation learning, cross-sensor adaptation, and tactile attribute inference. In this section, we review prior work from three perspectives and discuss their limitations in supporting physically grounded hierarchical reasoning.

Visuo-tactile perception. Visuo-tactile sensors convert contact-induced surface deformations into high-resolution images, providing rich cues for robot-object interaction [12]. By complementing visual perception with local contact feedback, visuo-tactile sensing has significantly advanced dexterous manipulation and contact-rich robotic tasks [13–16]. Existing methods mainly focus on local contact property estimation from tactile observations or global object-state inference through visuo-tactile fusion. Although effective in specific settings, many approaches are developed for particular sensors, objects, or manipulation tasks, making the learned representations tightly coupled with sensor-specific characteristics. As a result, they remain limited in capturing transferable contact semantics across heterogeneous tactile sensing platforms.

Sensor-agnostic tactile representation. To improve cross-sensor transfer, recent work has explored sensor-agnostic tactile representation learning. A common strategy is to learn a shared latent embedding space across tactile sensors to support few-shot or zero-shot transfer [5,6,17]. These representation-

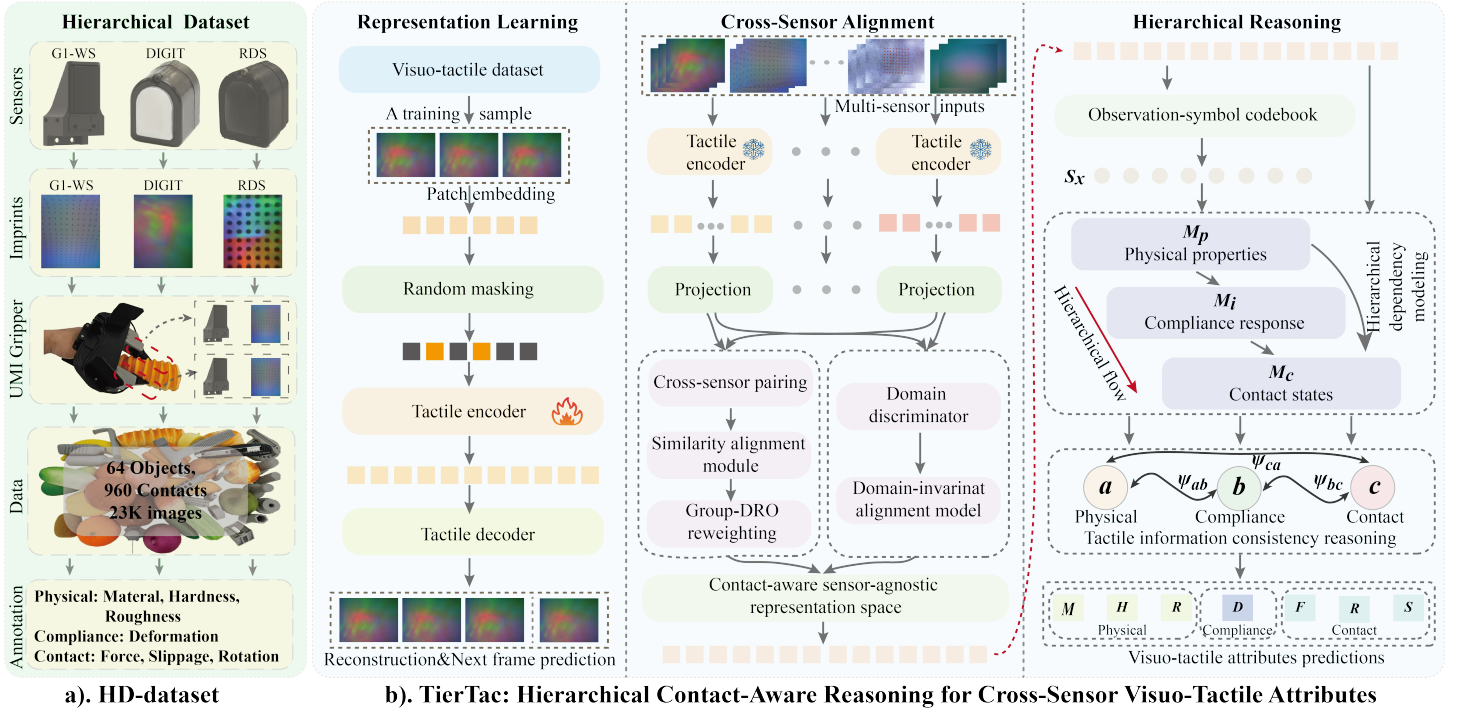


Figure 2: Overview of **TierTac**. The framework learns physically meaningful visuo-tactile representations and performs hierarchical reasoning for consistent multi-attribute inference across heterogeneous sensors. From left to right: (1) a hierarchical visuo-tactile dataset with multi-level attribute annotations; (2) spatio-temporal representation learning via self-supervised reconstruction and prediction; (3) cross-sensor alignment that learns contact-aware sensor-agnostic embeddings through multi-sensor pairing and robust reweighting; and (4) physically grounded hierarchical neuro-symbolic reasoning, which models dependencies from physical properties to compliance responses and contact states, enabling coherent visuo-tactile attribute prediction.

based approaches improve cross-sensor correspondence, but they primarily address tactile transfer as a feature alignment problem and may still retain sensor-specific artifacts. Another line of work aligns tactile features with high-level semantic anchors, such as language or multimodal representations, to enhance open-vocabulary or zero-shot generalization [3,18,19]. However, such semantic spaces are often weakly grounded in physically measurable contact properties, resulting in limited correspondence with fine-grained contact states. Recent work introduces intermediate physical domains for cross-sensor adaptation [20], but explicit modeling of physically grounded hierarchical dependencies among tactile attributes remains underexplored.

Tactile attribute inference and datasets. Tactile datasets have evolved from single-sensor and task-specific collections toward multimodal [21–23] and object-centric resources [3,4]. Some datasets emphasize static properties such as material and texture [9,10], while others include dynamic interaction cues such as sliding and contact force [11,24]. Despite this progress, existing datasets and inference methods often treat tactile attributes as separate prediction targets. Static physical properties, compliance responses, and contact states are rarely organized under a unified hierarchy that reflects the physical process of contact formation. This limits their ability to support structured multi-attribute reasoning and may lead to fragmented or physically inconsistent tactile predictions. In contrast, TierTac explicitly models visuo-tactile perception as a physically grounded hierarchical reasoning process,

linking physical properties, compliance responses, and contact states to enable coherent and transferable tactile understanding across sensors.

3 Methodology

TierTac models visuo-tactile perception as a physically grounded hierarchical reasoning process for generalizable tactile understanding. It first learns contact-aware sensor-agnostic representations that capture contact semantics while suppressing sensor-specific variations, and then performs structured inference over tactile attributes to enforce physically consistent predictions. As illustrated in Fig. 2, TierTac comprises three core components: (i) spatio-temporal tactile foundation modeling, (ii) contact-aware representation learning, and (iii) hierarchical neuro-symbolic reasoning.

3.1 Data and Hierarchical Supervision

TierTac utilizes three complementary forms of supervision to achieve generalizable visuo-tactile perception: large-scale tactile sequences for learning temporal contact dynamics, paired cross-sensor observations for sensor-agnostic alignment, and structured attribute annotations for physically grounded hierarchical reasoning.

For foundation modeling and cross-sensor alignment, we leverage a curated collection of public visuo-tactile datasets, including Octopi [10], OF-Real [25], YCB-Slide [26], TAG [22],

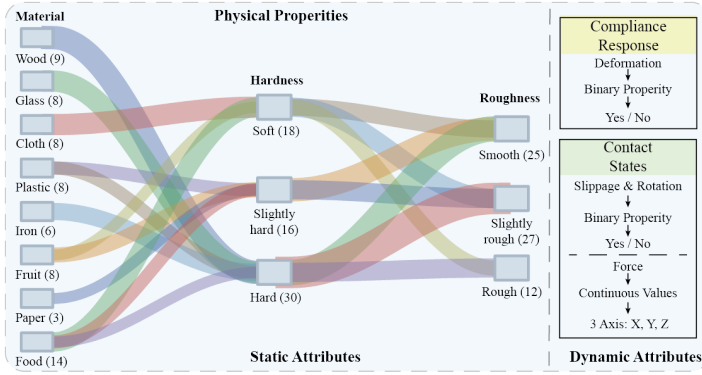


Figure 3: Physically grounded hierarchical organization of tactile attributes in the HD-dataset. Attributes are structured into three levels, i.e., physical properties, compliance responses, and contact states, reflecting the physical process of contact formation. This hierarchy encodes dependencies among attributes and provides structured supervision for physically consistent multi-attribute reasoning.

SSVTP [27], VisGel [21], TVL [28], Cloth [9], and FAF [24]. These datasets provide diverse sensor modalities, object properties, and interaction scenarios. In addition, we use TacQuad [3], which provides paired observations of identical contact events across heterogeneous sensors. These paired observations enable TierTac to align tactile semantics across sensors while mitigating sensor-specific appearance variations.

To support structured reasoning, we introduce a hierarchical visuo-tactile dataset, denoted as the HD-dataset, with multi-level attribute annotations. As illustrated in Fig. 3, tactile attributes are organized according to the physical process of contact formation, spanning intrinsic physical properties, compliance responses, and dynamic contact states. Intrinsic physical properties, such as material, hardness, and roughness, provide object-level priors; compliance responses, such as local deformation, describe intermediate physical reactions under contact; and dynamic contact states, such as force, slippage, and rotation, characterize the resulting interaction state. This organization encourages the model to treat tactile attributes not as isolated labels but as physically coupled variables.

The HD-dataset further incorporates an RDS sensor, designed as a variant of the DIGIT sensor with modified illumination and elastomer properties, to produce distinct tactile observations of the same contact interactions. While preserving the underlying contact states observed by DIGIT, RDS introduces sensor-specific variations and is reserved exclusively for zero-shot evaluation on unseen hardware, and is not used during training or validation. Contact locations are aligned across heterogeneous platforms, allowing observations from different sensors to correspond to the same physical interaction. This design provides structured supervision and a rigorous benchmark for evaluating physically consistent hierarchical reasoning across tactile sensors.

3.2 Tactile Foundation Model

To support physically grounded hierarchical reasoning for generalizable visuo-tactile perception, TierTac first learns

spatio-temporal tactile representations from large-scale visuo-tactile data. This component is designed to capture both static contact geometry and temporal interaction dynamics, as tactile attributes—such as deformation and force—are inherently governed by the evolution of contact and cannot be reliably inferred from a single frame.

Given a short tactile clip $X = \{x_t\}_{t=1}^4$, where $x_t \in \mathbb{R}^{3 \times H \times W}$, the first three frames are treated as observations and the fourth frame serves as a future prediction target. A patchify operator $P(\cdot)$ divides each frame into non-overlapping patches of size $p \times p$ and projects the clip into patch tokens:

$$T = P(X) \in \mathbb{R}^{L \times 4 \times d_p}, \quad L = (H/p) \times (W/p), \quad (1)$$

where $d_p = p^2 C$ denotes the flattened patch dimensionality. To encourage the model to capture temporal contact dynamics rather than relying on local spatial interpolation, we apply a tube-style mask $m \in \{0, 1\}^L$ that masks identical spatial locations across the observed frames. Let $\mathcal{M} = \{i \mid m_i = 1\}$ denote the set of masked patch indices. The masked tokens are encoded by a spatio-temporal encoder f_θ and reconstructed by a lightweight decoder g_ϕ , producing predicted tokens:

$$\hat{T} = g_\phi(f_\theta(T, m)), \quad \hat{T} \in \mathbb{R}^{L \times 4 \times d_p}. \quad (2)$$

We optimize objectives that jointly capture spatial structure and temporal dynamics. The static reconstruction loss anchors the representation to fine-grained contact geometry, while the dynamic objective combines observed-frame reconstruction and future-frame prediction:

$$\begin{cases} \mathcal{L}_{\text{static}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \|\hat{T}_{i,1} - T_{i,1}\|_2^2, \\ \mathcal{L}_{\text{recon}} = \frac{1}{3|\mathcal{M}|} \sum_{i \in \mathcal{M}} \sum_{t=1}^3 \|\hat{T}_{i,t} - T_{i,t}\|_2^2, \\ \mathcal{L}_{\text{pred}} = \frac{1}{L} \sum_{i=1}^L \|\hat{T}_{i,4} - T_{i,4}\|_2^2. \end{cases} \quad (3)$$

Here, $\mathcal{L}_{\text{recon}}$ enforces short-term consistency of deformation patterns across observed frames, while $\mathcal{L}_{\text{pred}}$ encourages modeling of future contact dynamics. The dynamic loss and overall objective are defined as:

$$\mathcal{L}_{\text{dynamic}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{pred}}. \quad (4)$$

$$\mathcal{L}_{\text{s1}} = \alpha \mathcal{L}_{\text{static}} + \beta \mathcal{L}_{\text{dynamic}}. \quad (5)$$

with $\alpha > \beta$ to prioritize accurate modeling of spatial contact geometry, which serves as the basis for higher-level reasoning over contact states.

The learned representation encodes transferable contact semantics that preserve physically meaningful deformation patterns while abstracting away instance-specific variations. These features provide the foundation for subsequent contact-aware sensor-agnostic representation learning and physically grounded hierarchical reasoning.

3.3 Contact-Aware Sensor-Agnostic Representation Learning

Although the foundation model can capture rich spatio-temporal patterns, its embeddings may still encode sensor-

dependent appearance cues, such as non-uniform illumination, elastomer textures, and imaging distortions, which can dominate cross-sensor comparisons and obscure shared contact semantics. To address this, we formulate a representation learning objective that preserves contact-relevant semantics while suppressing sensor-specific artifacts by maximizing shared information across sensors and minimizing dependence on sensor identity.

Let $x_i^{(s)}$ be a tactile observation of contact instance i from sensor $s \in \mathcal{S}$. The encoder f_θ and projection head h_ψ map each observation to an ℓ_2 -normalized embedding:

$$Z_i^{(s)} = \frac{h_\psi(f_\theta(x_i^{(s)}))}{\|h_\psi(f_\theta(x_i^{(s)}))\|_2}. \quad (6)$$

We formulate the learning goal as an information-theoretic objective:

$$\max_{\theta, \psi} \mathbb{E}[I(Z^a; Z^b)] - \beta I(Z; S), \quad (7)$$

where Z^a and Z^b denote embeddings of the same physical contact captured by heterogeneous sensors. Here, $I(Z^a; Z^b)$ encourages consistency between cross-sensor views, while $I(Z; S)$ penalizes dependence on sensor identity. This formulation encourages the embedding space to preserve sensor-agnostic contact semantics for downstream hierarchical reasoning.

We implement this objective using three complementary terms: multi-positive cross-sensor alignment, sensor-agnostic regularization, and robust group reweighting.

Cross-sensor alignment. We enforce cross-sensor consistency by aligning representations of the same physical contact observed across different sensors. This is implemented using a multi-positive InfoNCE objective. In each training batch, observations are organized by contact instance. For an anchor $z_i^{(s)}$, let $\mathcal{P}(i, s)$ denote the set of positives (the same physical contact captured by other sensors) and $\mathcal{N}(i, s)$ denote the set of negatives (different contact instances). We first define the scaled pairwise similarity as:

$$\ell_{ij} = \exp(\langle z_i^{(s)}, z_j \rangle / \tau). \quad (8)$$

where τ is a temperature parameter. The alignment loss for the anchor $z_i^{(s)}$ is then formulated as:

$$\ell_i^{mp} = -\log \frac{\sum_{p \in \mathcal{P}(i, s)} \ell_{ip}}{\sum_{p \in \mathcal{P}(i, s)} \ell_{ip} + \sum_{n \in \mathcal{N}(i, s)} \ell_{in}}. \quad (9)$$

This objective pulls together observations corresponding to the same physical contact across sensors while separating embeddings from distinct interactions. Consequently, the learned representation is encouraged to preserve physically meaningful contact semantics that are consistent across sensors and reflect the underlying physical interaction.

Sensor-agnostic regularization. Cross-sensor alignment alone may still retain residual sensor-identifying cues, especially when sensor-specific appearance is correlated with the input distribution. To further suppress such cues, we adopt an *InfoMax-Share* regularization strategy by introducing a sensor discriminator d_ω that predicts the sensor identity from the embedding z . The encoder is optimized to make the discriminator posterior approach a uniform distribution over sensors.

In practice, this is achieved via alternating adversarial optimization: the discriminator d_ω is trained to classify sensor identity from detached embeddings, while the encoder is optimized to minimize $\mathcal{L}_{im}$ to reduce the predictability of sensor identity:

$$\mathcal{L}_{im} = \mathbb{E}_z [\text{KL}(p_\omega(s | z) \| \mathcal{U})], \quad (10)$$

where $\mathcal{U}$ denotes the uniform distribution over sensor identities. This regularization discourages the encoding of sensor-specific information in z while preserving physically meaningful contact semantics shared across sensors, which are essential for robust downstream hierarchical reasoning.

Robust group reweighting. Given that TierTac is trained on heterogeneous sources with varying sensor characteristics and environmental noise, the alignment difficulty fluctuates across data domains. To ensure consistent performance across scenarios, we apply Group Distributionally Robust Optimization (Group-DRO) to the alignment objective. This strategy prioritizes groups with larger alignment errors, preventing the model from over-optimizing for easily aligned domains while neglecting more challenging ones.

Let $\mathcal{G} = \{\text{Fine}, \text{Indoor}, \text{Outdoor}\}$ denote the set of data-source groups. We first compute the group-wise alignment loss $\mathcal{L}_g$ for each group $g \in \mathcal{G}$ present in the current mini-batch $\mathcal{B}_g$:

$$\mathcal{L}_g = \frac{1}{|\mathcal{B}_g|} \sum_{i \in \mathcal{B}_g} \ell_i^{mp}. \quad (11)$$

The reweighted alignment objective is then formulated as:

$$\mathcal{L}_{align}^{dro} = \sum_{g \in \mathcal{G}_t} q_g \mathcal{L}_g, \quad (12)$$

where $\mathcal{G}_t$ represents the active groups in the mini-batch and q_g are adaptive weights. These weights are updated online using an exponential gradient step:

$$q_g \leftarrow q_g \exp(\eta \mathcal{L}_g), \quad g \in \mathcal{G}_t, \quad (13)$$

followed by normalization over $\mathcal{G}_t$ to maintain $\sum q_g = 1$, where η is the step size. This mechanism dynamically shifts optimization focus toward high-loss groups, promoting consistent cross-sensor alignment across heterogeneous data sources. The total objective for this stage is:

$$\mathcal{L}_{s2} = \mathcal{L}_{align}^{dro} + \lambda_{im} \mathcal{L}_{im}. \quad (14)$$

This process yields a sensor-agnostic embedding space that preserves physically meaningful contact semantics and provides a robust input space for subsequent physically grounded hierarchical reasoning.

3.4 Physically-Grounded Hierarchical Neuro-Symbolic Reasoning

The sensor-agnostic representation learned above preserves contact semantics across sensors, but it does not explicitly encode the physical constraints that govern tactile attributes. To achieve physically consistent multi-attribute inference, TierTac introduces a hierarchical neuro-symbolic reasoning framework. The hierarchy follows the physical process of contact formation: intrinsic physical properties provide priors for compliance responses, which in turn constrain

interaction-level contact states. By propagating information across this hierarchy, TierTac refines local predictions and resolves attribute-level conflicts.

Symbolic abstraction. To bridge continuous tactile embeddings with structured attribute reasoning, we introduce a latent symbolic space. The symbols are learned latent prototypes rather than manually specified semantic labels, allowing the model to capture recurring contact patterns shared across attributes. Given the sensor-agnostic embedding e , the model projects it onto an observation-symbol codebook $\{c_k^o\}_{k=1}^{K_o}$:

$$\pi^o = \text{Softmax}(W_o e / \tau), \quad s_x = \sum_{k=1}^{K_o} \pi_k^o c_k^o. \quad (15)$$

The resulting symbolic representation s_x provides a compact abstraction of the interaction and serves as a structured interface for reasoning. TierTac performs hierarchical prediction across three physically grounded levels: physical properties (p), compliance responses (i), and contact states (c):

$$\begin{cases} (s_p, \hat{y}_p, b_p) = M_p(e, s_x), \\ (s_i, \hat{y}_i, b_i) = M_i(e, s_x, s_p, b_p), \\ (s_c, \hat{y}_c) = M_c(e, s_x, s_p, b_p, s_i, b_i). \end{cases} \quad (16)$$

Here, $\hat{y}_p$, $\hat{y}_i$, and $\hat{y}_c$ denote predictions at each level; s_p , s_i , and s_c are latent symbolic representations; and b_p and b_i are intermediate features for cross-level information flow. This design allows contact-state inference to be conditioned on physical-property and compliance-level context rather than relying on independent task-specific heads.

Compatibility-based structured reasoning. To enforce cross-level coherence and resolve local inconsistencies, TierTac refines unary predictions using learned pairwise compatibility matrices Ψ over related attributes. For a triplet of related attributes (a, b, c), let u_a , u_b , and u_c denote unary logits, and let $p_a^{(t)}$, $p_b^{(t)}$, and $p_c^{(t)}$ denote their class distributions at reasoning iteration t . Starting from $p_a^{(0)} = \text{Softmax}(u_a)$, $p_b^{(0)} = \text{Softmax}(u_b)$, and $p_c^{(0)} = \text{Softmax}(u_c)$, TierTac performs iterative symbolic message passing:

$$\begin{cases} p_a^{(t+1)} = \text{Softmax}(u_a + \Psi_{ab} p_b^{(t)} + \Psi_{ac} p_c^{(t)}), \\ p_b^{(t+1)} = \text{Softmax}(u_b + \Psi_{ba} p_a^{(t)} + \Psi_{bc} p_c^{(t)}), \\ p_c^{(t+1)} = \text{Softmax}(u_c + \Psi_{ca} p_a^{(t)} + \Psi_{cb} p_b^{(t)}). \end{cases} \quad (17)$$

This refinement allows attribute predictions to exchange information across hierarchy levels, progressively resolving conflicts among related attributes and producing more physically consistent multi-attribute predictions.

Continuous force prediction. To maintain numerical precision for force estimation, we regress the continuous components (f_x, f_y, f_z) using a head conditioned on symbolic features from the contact-state layer. This allows force prediction to benefit from hierarchical context while preserving the resolution required for fine-grained manipulation.

Training objective. Training combines task-specific supervision with symbolic and structural regularization. The task loss includes classification losses for discrete attributes and a regression loss for continuous force prediction:

$$\mathcal{L}_{task} = \sum_{r \in \{p, i, c\}} \mathcal{L}_{cls}^r + \lambda_f \mathcal{L}_{force}. \quad (18)$$

The overall objective is:

$$\mathcal{L}_{s3} = \mathcal{L}_{task} + \lambda_{conf} \mathcal{L}_{conf} + \lambda_{bal} \mathcal{L}_{bal} + \lambda_{sp} \mathcal{L}_{sp} + \lambda_{it} \mathcal{L}_{it}. \quad (19)$$

Here, $\mathcal{L}_{conf}$ encourages confident symbol assignment, $\mathcal{L}_{bal}$ prevents codebook collapse, $\mathcal{L}_{sp}$ imposes an ℓ_1 penalty on compatibility matrices to encourage sparse attribute dependencies, and $\mathcal{L}_{it}$ regularizes consecutive reasoning iterations. By combining symbolic abstraction with compatibility-based message passing, TierTac enables interpretable and physically consistent multi-attribute inference.

4 Experimental study

We evaluate TierTac from two complementary perspectives: (i) perception-Level evaluation to judge sensor-agnostic representation learning and physically grounded hierarchical reasoning across heterogeneous hardware; and (ii) robotic evaluation to examine how these perception capabilities translate into improved interaction stability and force efficiency under both classical control and learned policies. Experiments are conducted on multiple large-scale benchmarks (e.g., TouchD Bench) and real-world robotic platforms, including a Kinova Gen3 robot equipped with diverse tactile sensors (e.g., DIGIT and G1-WS), to assess generalization.

Table 1: Ablation study on sensor-agnostic representations over the selected cross-sensor contact subset.

Method	P90↓	CS ↑	R@1↑	RS Acc.↓
Foundation model	0.315	-0.377	0.067	1.000
+InfoNCE & Group-DRO	1.020	-0.322	0.189	1.000
+InfoMax-Share (our)	0.004	0.990	0.981	0.656

4.1 Perception-Level Evaluation

We evaluate TierTac on multiple visuo-tactile datasets from three complementary perspectives: (i) sensor-agnostic contact representation quality; (ii) physical consistency in multi-attribute reasoning; and (iii) generalization across heterogeneous sensors and datasets.

4.1.1 Sensor-Agnostic Contact Representation

We first evaluate whether TierTac extracts contact-aware, sensor-agnostic representations that preserve underlying physical semantics while suppressing sensor-specific bias. Ideally, a robust embedding space should be structured by "contact identity" rather than imaging hardware. This analysis is conducted on a held-out subset of TacQuad [3], featuring aligned cross-sensor observations of identical contact events.

Quantitative analysis. Table 1 reports a progressive ablation study using four metrics to characterize the structure of the embedding space:

- Pair Dist. (P90): the 90th percentile of Euclidean distances between pairs of the same contact observed across sensors, measuring cross-sensor compactness;

- Contact Silhouette (CS): the silhouette score calculated using contact identity as labels, reflecting inter-contact separability;
- R@1: the top-1 accuracy of cross-sensor contact retrieval, assessing semantic alignment;
- RS Acc.: the accuracy of a linear classifier trained to identify the sensor type from embeddings, quantifying the amount of residual sensor-specific information.

The foundation model exhibits a fragmented representation space dominated by sensor-specific characteristics, evidenced by large within-contact distances and near-perfect **RS Acc.** (1.000). This indicates that contact semantics are strongly dominated by **hardware-dependent artifacts**. While introducing InfoNCE and Group-DRO improves retrieval and separability, it fails to sufficiently suppress sensor bias; the increased P90 (1.020) further suggests that naive alignment without explicit sensor-agnostic regularization may distort the embedding manifold by attempting to retain conflicting **modality-specific signatures**.

In contrast, incorporating **InfoMax-Share** yields consistent gains across all metrics. TierTac achieves near-perfect retrieval accuracy (0.981) and a substantial reduction in RS Acc. (0.656), indicating that the encoder reduces sensor-dependent cues—such as illumination and gel-texture variations—while preserving physically meaningful contact semantics. The consistent improvement across all four metrics confirms that TierTac constructs an embedding space that is both discriminative for contact states and less dependent on sensor hardware.

Qualitative visualization. We further analyze the embedding topology using t-SNE (Fig. 4). Baseline methods show distinct clustering by sensor type (DIGIT, DuraGel, and GelSight Mini), confirming that their representations remain dominated by sensor-specific characteristics. Although methods such as AnyTouch partially improve cross-sensor alignment, noticeable gaps remain between samples of the same contact across sensors, and different contact states are not well separated. In contrast, TierTac organizes the embedding space by contact identity, where samples from these heterogeneous sensors are tightly clustered for the same contact while remaining well separated from others.

This structure confirms that TierTac captures physically meaningful contact states rather than pixel-level appearance. Such a sensor-agnostic foundation is critical for the subsequent hierarchical reasoning stage, ensuring that the neuro-symbolic module operates on robust physical abstractions across different sensing platforms.

4.1.2 Physically Consistent Multi-Attribute Reasoning

We evaluate whether modeling hierarchical dependencies among tactile attributes improves both prediction accuracy and cross-attribute physical consistency. Unlike independent attribute prediction, tactile attributes are physically coupled through contact interactions, requiring coherent inference across levels. Experiments are conducted on the HR-dataset using a held-out test set of unseen contact

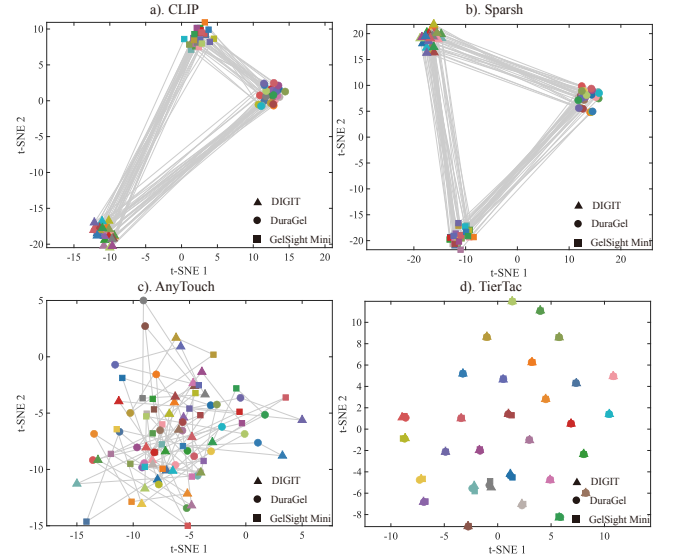


Figure 4: t-SNE visualization of the learned representation spaces for CLIP, Sparsh, AnyTouch, and TierTac on the selected cross-sensor contact subset. Triangles, circles, and squares denote DIGIT, DuraGel, and GelSight Mini, respectively. Each color represents one contact instance, and samples of the same contact across sensors are connected by dashed lines.

instances. The dataset provides structured annotations spanning physical properties, compliance responses, and dynamic contact states, enabling evaluation of both predictive performance and physical consistency.

Prediction performance. Table 2 compares TierTac with state-of-the-art baselines across classification and regression tasks. TierTac achieves the best overall performance across physical properties, compliance responses, and contact states. For physical and compliance attributes, it obtains the highest accuracy in material, hardness, roughness, and deformation, indicating that the learned sensor-agnostic representations preserve intrinsic object-level and compliance-related information. More importantly, the gains are most pronounced for contact-state attributes, including rotation, slippage, and force. These attributes are strongly coupled with lower-level physical properties and compliance responses, making them difficult to infer using independent prediction heads. The improvement therefore suggests that hierarchical reasoning is particularly beneficial for complex contact dynamics, where physically coupled attributes should be inferred coherently.

Consistency across attributes. Beyond individual task accuracy, TierTac exhibits consistently lower Cross-task Std. compared to prior methods (Table 2), reflecting a more balanced and coherent performance profile. Such consistency is critical in tactile perception, as attributes are inherently interdependent and should be inferred in a coordinated manner. In contrast, independent predictions may lead to physically inconsistent interpretations. For example, a baseline may predict high hardness together with large local deformation, or report a strong contact force without corresponding slip or rotation, which is inconsistent with typical contact behavior. In contrast, TierTac encourages cross-attribute consistency through hierarchical dependencies, producing predictions that

Table 2: Multi-attribute prediction on the HR-dataset across physical properties, compliance responses, and contact states.

Method	Physical			Compliance	Contact			Std. ↓
	Material ↑	Hardness ↑	Roughness ↑	Deformation ↑	Rotation ↑	Slippage ↑	Force ↓	
CLIP	50.74	73.62	76.61	71.84	32.64	34.48	935.63	30.64
UniTouch	56.82	74.86	77.45	72.94	35.43	35.73	1116.68	19.43
T3	53.35	78.43	76.94	75.37	53.31	53.89	1108.11	12.85
Octopi	–	76.57	78.72	–	–	–	–	1.52
Sparsh	55.93	76.53	74.92	72.05	78.92	76.25	713.14	8.39
AnyTouch1	<u>71.15</u>	79.36	<u>83.52</u>	74.26	43.74	46.38	881.71	17.09
AnyTouch2	69.82	<u>79.98</u>	83.21	<u>75.14</u>	<u>83.63</u>	<u>84.52</u>	<u>433.26</u>	5.97
TierTac	80.72	81.08	86.43	81.97	85.34	85.81	421.68	<u>2.58</u>

Footnote: Physical attributes (material, hardness, roughness) and compliance (deformation) are evaluated by accuracy. Rotation and slippage are evaluated using F1 score, and force is evaluated using RMSE (mN). Cross-task Std. is computed over classification tasks only (i.e., except Force). Rotation and slippage are evaluated using F1 score due to class imbalance, while the remaining classification tasks are evaluated using accuracy. ↑/↓ indicate higher/lower is better. Underlined values denote the second-best results, and "–" indicates unsupported tasks.

Table 3: Ablation study on the effects of symbolization and hierarchical reasoning in visuo-tactile attribute prediction.

Method	Physical			Compliance	Contact			Std. ↓
	Material ↑	Hardness ↑	Roughness ↑	Deformation ↑	Rotation ↑	Slippage ↑	Force ↓	
Multi-Head	<u>73.72</u>	<u>77.05</u>	<u>80.68</u>	76.36	<u>72.47</u>	<u>63.71</u>	<u>825.54</u>	<u>5.79</u>
+Symbolization	71.25	71.61	74.93	<u>78.75</u>	69.46	60.83	884.81	6.03
+Reasoning (Whole)	80.72	81.08	86.43	81.97	85.34	85.81	421.68	2.58

Footnote: Metrics follow Table 2. Multi-Head uses a shared backbone with independent task-specific heads. +Symbolization introduces discrete symbolic representations without structured reasoning. +Reasoning (Whole) incorporates the proposed hierarchical neuro-symbolic reasoning module. Bold and underline indicate the best and second-best results, respectively.

are not only accurate but also more coherent across tasks.

Effect of structured reasoning. To isolate the contribution of structured reasoning, we compare TierTac with variants that remove or simplify the reasoning component (Table 3).

The Multi-Head baseline achieves reasonable performance but treats each attribute independently, resulting in higher cross-task variation and less coherent predictions. This indicates that independent prediction fails to capture dependencies among tactile attributes. Introducing symbolization alone does not yield consistent improvements and can even degrade performance, suggesting that discrete representations without explicit reasoning are insufficient for modeling attribute relationships.

In contrast, incorporating hierarchical structured reasoning consistently improves all tasks. For example, slippage F1 increases from 63.71 to 85.81, and force RMSE is reduced from 825.54 to 421.68. Meanwhile, cross-task variation is reduced, indicating more coherent and physically consistent multi-attribute inference.

Importantly, since the backbone architecture and training data remain unchanged, these improvements cannot be attributed to increased model capacity, but instead highlight the role of structured reasoning in modeling dependencies among tactile attributes.

Overall, these results demonstrate that hierarchical reasoning is essential for accurate and consistent multi-attribute perception.

4.1.3 Generalization Across Datasets and Sensors

We evaluate whether TierTac generalizes across datasets and sensor types by testing its ability to transfer learned contact semantics and hierarchical reasoning to unseen data distributions and hardware configurations.

Cross-dataset generalization. Table 4 reports performance across multiple visuo-tactile benchmarks, covering both within-sensor and cross-sensor evaluation settings. TierTac consistently outperforms existing methods across these configurations. Compared with CLIP, which relies on generic visual features, TierTac achieves better performance by capturing contact-specific semantics beyond raw pixel appearance. Furthermore, TierTac demonstrates stronger robustness than tactile representation models such as Sparsh and AnyTouch, indicating that its representations are less dependent on dataset-specific statistics.

Notably, the gains remain consistent in cross-sensor scenarios, a more challenging setting due to variations in sensing characteristics. This suggests that TierTac learns transferable contact-aware features rather than overfitting to specific sensing patterns.

Generalization to unseen sensors. We further assess TierTac’s transferability to sensors not observed during training, including the real-world RDS sensor in the HR-dataset, simulated sensors in ObjectFolder, and the DM-Tac W real sensor in TouchD Bench. We compare the zero-shot setting of TierTac with TierTac*, which uses only 1% target-sensor data to fine-tune the stage-3 reasoning module.

In the zero-shot setting, TierTac consistently outperforms prior methods across unseen sensors. On RDS, TierTac im-

Table 4: Cross-dataset and cross-sensor transfer evaluation across tactile benchmarks with diverse sensing modalities. Results are reported on static object properties (ObjTac) and dynamic interaction tasks (Sparsh and TouchD Bench), including both classification and regression settings.

Method	ObjTac (Static)				Sparsh Bench (Slip)			Sparsh Bench (Force)			TouchD Bench (Force)		
	Mat. $\uparrow$	Hard. $\uparrow$	Rough. $\uparrow$	Std. ^t $\downarrow$	DG $\uparrow$	GM $\uparrow$	Std. ^s $\downarrow$	DG $\downarrow$	GM $\downarrow$	Std. ^s $\downarrow$	DG $\downarrow$	GM $\downarrow$	Std. ^s $\downarrow$
AnyTouch2	42.52	67.38	89.46	23.49	86.66	87.80	0.81	624.26	202.14	298.48	894.32	1051.03	110.81
SA-ViT [8]	<u>70.44</u>	<u>82.21</u>	93.91	<u>11.74</u>	—	—	—	—	—	—	—	—	—
TierTac	85.82	84.58	89.13	2.35	89.53	89.92	0.28	357.67	<u>316.38</u>	29.20	723.82	767.39	30.81

Footnote: Mat., Hard., and Rough. denote material, hardness, and roughness accuracy. FTS, GM, and DG denote force-based tactile sensing, GelSight Mini, and DIGIT, respectively. For FTS data, raw signals are converted into pseudo-image representations [8]. Std.^t denotes cross-task standard deviation on ObjTac, while Std.^s denotes cross-sensor standard deviation on Sparsh Bench and TouchD Bench. $\uparrow/\downarrow$ indicate higher/lower is better. Underlined values denote the second-best results, and "—" indicates unavailable results due to the lack of publicly released models.

Table 5: Generalization to unseen sensors across datasets and sensing modalities. Results are reported on the HR-dataset with the unseen RDS sensor, ObjectFolder with simulated sensors (TACTO, Taxim), and TouchD Bench with an unseen real sensor (DM-Tac W).

Method	HR-dataset (RDS)							ObjectFolder (Material)			TouchD (Force)
	Mat. $\uparrow$	Hard. $\uparrow$	Rough. $\uparrow$	Def. $\uparrow$	Rot. $\uparrow$	Slip. $\uparrow$	Std. ^t $\downarrow$	TACTO $\uparrow$	Taxim $\uparrow$	Std. ^s $\downarrow$	DM-Tac W $\downarrow$
UniTouch	33.62	42.96	43.17	46.15	29.52	31.76	7.04	47.25	75.29	19.83	6142.16
AnyTouch2	46.89	56.56	62.73	53.25	46.92	48.15	6.38	49.62	76.02	18.67	(Seen)
TierTac	<u>52.14</u>	<u>61.83</u>	<u>64.74</u>	<u>60.15</u>	<u>54.73</u>	<u>53.72</u>	<u>5.06</u>	<u>61.83</u>	<u>79.42</u>	<u>12.44</u>	<u>4637.42</u>
TierTac*	71.26	73.91	76.53	78.42	73.04	72.64	2.67	81.47	84.37	2.05	1945.73

Footnote: Mat., Hard., Rough., Def., Rot., and Slip. denote material, hardness, roughness, deformation, rotation, and slippage. Std.^t denotes cross-task standard deviation, while Std.^s denotes cross-sensor standard deviation. "TierTac" denotes direct evaluation on unseen sensors, while "TierTac*" uses 1% target-sensor data to fine-tune only the stage-3 reasoning procedure. "(Seen)" indicates the sensor is included in training. $\uparrow/\downarrow$ indicate higher/lower is better.

proves prediction across all attributes compared with UniTouch and AnyTouch2, while reducing cross-task variation from 7.04 and 6.38 to 5.06. On ObjectFolder, TierTac achieves higher material recognition accuracy on both TACTO and Taxim and reduces cross-sensor variation by over 30%. On TouchD Bench, TierTac reduces the force prediction error on the unseen DM-Tac W sensor from 6142.16 to 4637.42. These results indicate that TierTac can transfer to sensors with different tactile rendering and sensing characteristics without target-sensor data during training.

With only 1% target-sensor data, TierTac* further improves performance across all unseen-sensor settings. On RDS, cross-task variation decreases to 2.67, indicating more stable multi-attribute inference. On ObjectFolder, cross-sensor variation decreases to 2.05, while force RMSE on TouchD Bench is reduced to 1945.73.

These findings suggest that TierTac provides a transferable contact-semantic foundation. Limited target-sensor data mainly helps calibrate the reasoning module to sensor-specific observation patterns. Overall, the combination of contact-level semantic grounding and structured hierarchical reasoning is important for consistent and transferable visuo-tactile perception across heterogeneous platforms.

4.2 Robotic Evaluation

We evaluate whether physically consistent visuo-tactile perception improves downstream robotic interaction. Specifically, we examine how TierTac consistently benefits downstream interaction across diverse control paradigms.

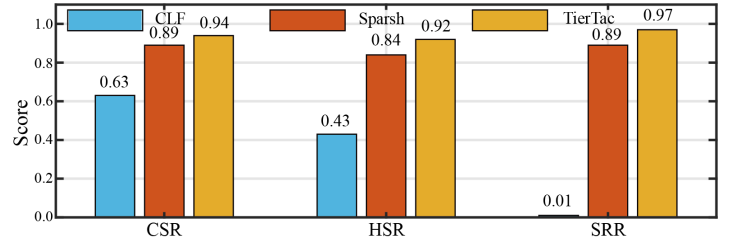


Figure 5: Grasp stability evaluation in the force-efficient stable grasping task. Comparison of CLF, Multi-Head baseline (Sparsh), and TierTac across grasp success rate (GSR), hold success rate (HSR), and slip recovery rate (SRR).

4.2.1 Classical Closed-Loop Control

We evaluate whether physically consistent tactile perception translates into improved closed-loop control performance in a force-efficient grasping task.

Setup. Experiments are conducted on a Kinova Gen3 manipulator equipped with a parallel gripper instrumented with DIGIT sensors. We select 10 objects with diverse material properties (e.g., compliance and friction) and geometries, performing 10 independent trials per object. Each trial follows a four-stage protocol: initial contact, force ramp-up, lifting, and holding.

To isolate the impact of tactile perception, all methods share the same hardware setup, object set, control policy, and force update rule; only the tactile perception module is replaced. A fixed closed-loop control scheme is used: the controller increases the normal grip force by a constant in-

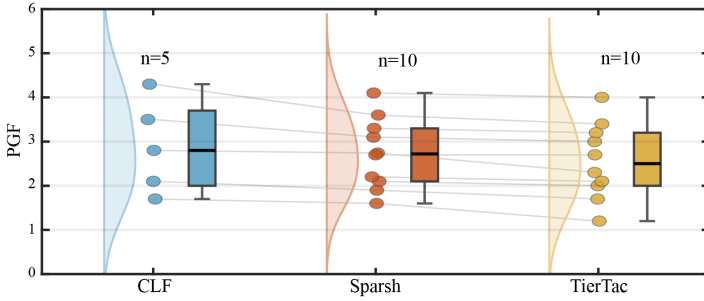


Figure 6: Force efficiency comparison in the force-efficient stable grasping task. Peak grip force (PGF) measured on successful trials for CLF, Multi-Head baseline (Sparsh), and TierTac. Each point corresponds to one object, with lines indicating paired comparisons across methods.

crement of 0.2 N when contact instability (e.g., slip or rotation) is detected, and maintains the force once a stable state is inferred. We compare three methods: (i) Constant-Force Lifting (CLF), which applies a fixed grip force without tactile feedback; (ii) a Multi-Head baseline (Sparsh), which predicts tactile attributes independently for force adjustment; and (iii) TierTac, which performs force regulation based on hierarchical tactile reasoning. We evaluate grasp success rate (GSR), hold success rate (HSR), slip recovery rate (SRR), and peak grip force (PGF). GSR, HSR, and SRR measure interaction stability, while PGF evaluates force efficiency.

Results. As shown in Fig. 5, TierTac achieves the best performance across all metrics. Improvements are particularly pronounced in SRR, indicating a stronger ability to recover from contact instability. In addition, TierTac consistently yields lower PGF than both CLF and the Sparsh baseline (Fig. 6), showing that improved stability is achieved without excessive force. These results suggest that more accurate and consistent tactile perception enables better control decisions under the same policy.

Behavior analysis. Fig. 7 provides qualitative comparisons of closed-loop grasping behaviors. CLF relies on a fixed-force strategy and cannot adapt to object-specific contact conditions, often leading to either insufficient contact or excessive force. The Sparsh baseline leverages tactile feedback but exhibits inconsistent responses, with noticeable slip in the final imprints, indicating unreliable interpretation of contact states.

In contrast, TierTac maintains more stable and structured visuo-tactile imprints throughout the interaction. The contact patterns evolve smoothly from initial contact to final grasp, reflecting more consistent estimation of contact states. This enables more timely and appropriate force adjustments, leading to improved grasp stability and force efficiency.

4.2.2 Learned Visuo-Tactile Policies

We evaluate whether physically grounded hierarchical tactile representations improve policy learning and generalization in end-to-end visuo-tactile manipulation.

Setup. We adopt the VTLA framework [29] as the policy backbone and train on the OmniVTA dataset [8], which provides paired visual observations, tactile sequences, and action

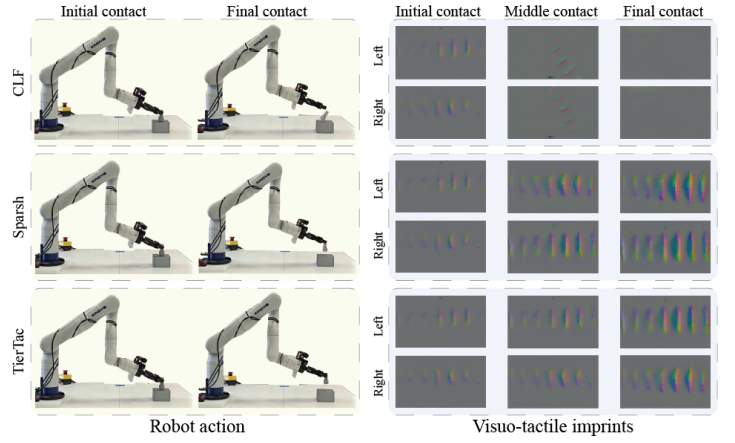


Figure 7: Closed-loop grasping behavior comparison. Robot actions (left) and visuo-tactile imprints (right) across grasping stages for CLF, Multi-Head baseline (Sparsh), and TierTac.

trajectories for contact-rich manipulation tasks. The policy takes visual inputs, tactile observations, and task instructions to predict continuous end-effector actions. To ensure a controlled comparison, all components of the policy, including the backbone, action representation, and training pipeline, are kept identical across variants; only the tactile encoder is replaced. We compare three encoders: a generic ViT encoder, a tactile-specific Sparsh encoder, and the proposed TierTac encoder. Performance is evaluated using task success rate (SR) and success rate under unseen conditions (SR-unseen).

Results. As shown in Fig. 8, TierTac consistently outperforms both ViT and Sparsh across all six manipulation tasks. The gains are particularly evident in contact-sensitive tasks such as "Assembly" and "Grasp", where accurate tactile interpretation is critical. TierTac also achieves higher SR-unseen, indicating stronger generalization beyond the training distribution. These improvements suggest that structured tactile representations provide more reliable contact-state information for policy learning.

Behavior analysis. We further analyze policy behaviors to understand these improvements. Compared to ViT and Sparsh, TierTac produces more stable and consistent action sequences, with fewer abrupt adjustments during execution. This difference is most pronounced in contact-rich stages, where accurate interpretation of tactile feedback is essential. As shown in Fig. 9, policies using ViT and Sparsh often fail to maintain stable contact during alignment and interaction stages. In contrast, TierTac maintains more consistent contact-aware behavior across execution stages, leading to successful task completion. This indicates that structured tactile representations enable more reliable contact-state estimation, which in turn improves policy stability and decision-making.

5 Conclusion

We presented TierTac, a hierarchical contact-aware reasoning framework for transferable visuo-tactile perception across heterogeneous sensors. By integrating sensor-agnostic representation learning with hierarchical neuro-symbolic reason-

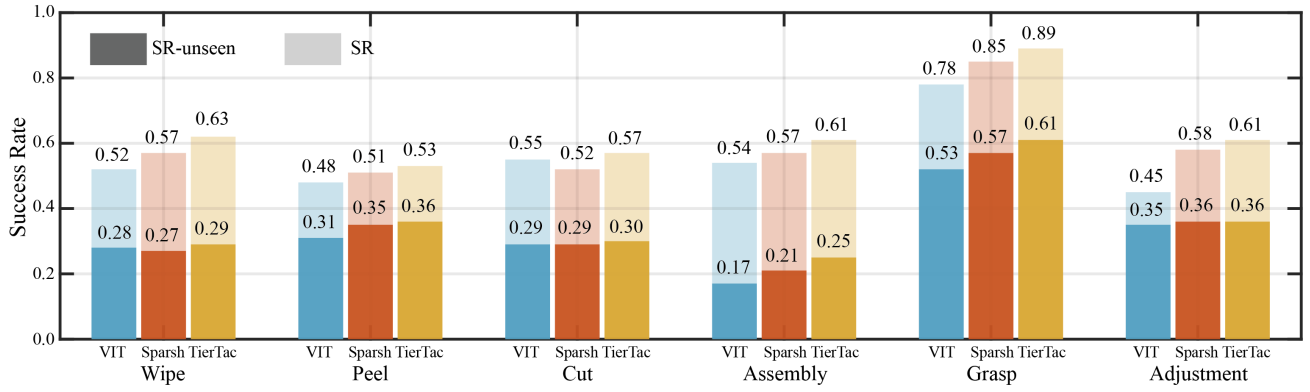


Figure 8: Policy performance on six contact-rich manipulation tasks from the OmniVTA dataset. We compare ViT, Sparsh, and TierTac using task success rate (SR, light bars) and success rate under unseen conditions (SR-unseen, dark bars)

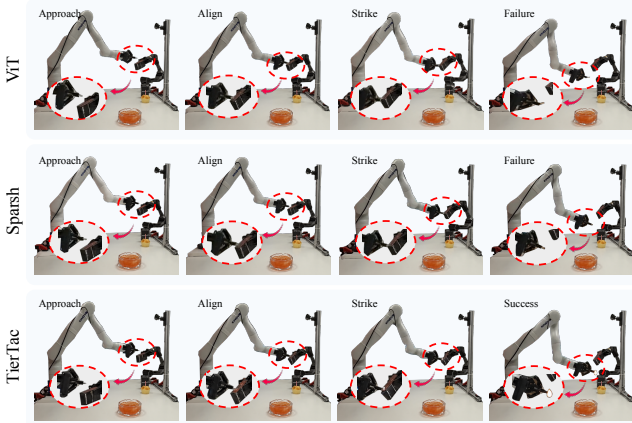


Figure 9: Qualitative comparison of learned visuo-tactile policy behaviors in a contact-rich manipulation task. The figure shows representative execution stages, including approach, alignment, and contact interaction, for policies using ViT, Sparsh, and TierTac tactile encoders under the same VTTLA framework.

ing, TierTac enables physically consistent multi-attribute inference while reducing hardware-dependent bias. Extensive experiments show that TierTac learns contact-centric representations, improves attribute prediction, and generalizes robustly to unseen sensors. These perception-level gains translate into more stable and force-efficient interaction in both classical control and learned policy settings on a real-world manipulator. Overall, our results highlight the importance of modeling tactile perception as a structured physical reasoning process. Future work will extend TierTac to broader manipulation tasks and online adaptation in unstructured environments.

Acknowledgements

This work was supported in part by the Research Grants Council (RGC) of Hong Kong under grant 15231023, and in part by the PolyU FENG Intra-Faculty Interdisciplinary Project under grant WZ8B.

References

- [1] Shoujie Li, Zihan Wang, Changsheng Wu, Xiang Li, Shan Luo, Bin Fang, Fuchun Sun, Xiao-Ping Zhang, and Wenbo Ding. 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.
- [2] Yuyang Li, Yinghan Chen, Zihang Zhao, Puhao Li, Tengyu Liu, Siyuan Huang, and Yixin Zhu. Simultaneous tactile-visual perception for learning multimodal robot manipulation. *IEEE Robotics and Automation Letters*, 2026.
- [3] Ruoxuan Feng, Jiangyu Hu, Wenke Xia, Tianci Gao, Ao Shen, Yuhao Sun, Bin Fang, and Di Hu. Anytouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors. *arXiv preprint arXiv:2502.12191*, 2025.
- [4] Samanta Rodriguez, Yiming Dou, William van den Bogert, Miquel Oller, Kevin So, Andrew Owens, and Nima Fazeli. Contrastive touch-to-touch pretraining. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5857–5863. IEEE, 2025.
- [5] Jialiang Zhao, Yuxiang Ma, Lirui Wang, and Edward H Adelson. Transferable tactile transformers for representation learning across diverse sensors and tasks. *arXiv preprint arXiv:2406.13640*, 2024.
- [6] Carolina Higuera, Akash Sharma, Chaithanya Krishna Bodduluri, Taosha Fan, Patrick Lancaster, Mrinal Kalakrishnan, Michael Kaess, Byron Boots, Mike Lambeta, Tingfan Wu, and Mustafa Mukadam. Sparsh: Self-supervised touch representations for vision-based tactile sensing. 2024.
- [7] Ning Cheng, Jinan Xu, Jialing Chen, Bin Fang, and Wenjuan Han. Stola: Self-adaptive touch-language framework for tactile commonsense reasoning in open-ended scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 18198–18206, 2026.
- [8] Zhengxue Cheng, Yiqian Zhang, Wenkang Zhang, Haoyu Li, Keyu Wang, Li Song, and Hengdi Zhang. Omnivta:

- Vision-tactile-language-action model with semantic-aligned tactile sensing. *arXiv preprint arXiv:2508.08706*, 2025.
- [9] Wenzhen Yuan, Yuchen Mo, Shaoxiong Wang, and Edward H Adelson. Active clothing material perception using tactile sensing and deep learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4842–4849. IEEE, 2018.
- [10] Samson Yu, Kelvin Lin, Anxing Xiao, Jiafei Duan, and Harold Soh. Octopi: Object property reasoning with large tactile-language models. *arXiv preprint arXiv:2405.02794*, 2024.
- [11] Ruoxuan Feng, Yuxuan Zhou, Siyu Mei, Dongzhan Zhou, Pengwei Wang, Shaowei Cui, Bin Fang, Guocai Yao, and Di Hu. Anytouch 2: General optical tactile representation learning for dynamic tactile perception. *arXiv preprint arXiv:2602.09617*, 2026.
- [12] Tong Li, Yuhang Yan, Chengshun Yu, Jing An, Yifan Wang, and Gang Chen. A comprehensive review of robot intelligent grasping based on tactile perception. *Robotics and Computer-Integrated Manufacturing*, 90:102792, 2024.
- [13] Qiwei Wu, Haidong Wang, Jiayu Zhou, Xiaogang Xiong, and Yunjiang Lou. Tars: Tactile affordance in robot synesthesia for dexterous manipulation. *IEEE Robotics and Automation Letters*, 2024.
- [14] Kang-Won Lee, Yuzhe Qin, Xiaolong Wang, and Soo-Chul Lim. Dextouch: Learning to seek and manipulate objects with tactile dexterity. *IEEE Robotics and Automation Letters*, 2024.
- [15] Ying Yuan, Haichuan Che, Yuzhe Qin, Binghao Huang, Zhao-Heng Yin, Kang-Won Lee, Yi Wu, Soo-Chul Lim, and Xiaolong Wang. Robot synesthesia: In-hand manipulation with visuotactile sensing. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6558–6565. IEEE, 2024.
- [16] Haoran Yuan, Weigang Yi, Zhenyu Zhang, Wendi Chen, Yuchen Mo, Jiashi Yin, Xinzhuo Li, Xiangyu Zeng, Chuan Wen, Cewu Lu, et al. Vtam: Video-tactile-action models for complex physical interaction beyond vlas. *arXiv preprint arXiv:2603.23481*, 2026.
- [17] Zhengtong Xu, Raghava Uppuluri, Xinwei Zhang, Cael Fitch, Philip Glen Crandall, Wan Shou, Dongyi Wang, and Yu She. Unit: Data efficient tactile representation with generalization to unseen objects. *IEEE Robotics and Automation Letters*, 2025.
- [18] Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, et al. Binding touch to everything: Learning unified multimodal tactile representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26340–26353, 2024.
- [19] Liang Heng, Haoran Geng, Kaifeng Zhang, Pieter Abbeel, and Jitendra Malik. Vitacformer: Learning cross-modal representation for visuo-tactile dexterous manipulation. *arXiv preprint arXiv:2506.15953*, 2025.
- [20] Samanta Rodriguez, Yiming Dou, Miquel Oller, Andrew Owens, and Nima Fazeli. Cross-sensor touch generation. *arXiv preprint arXiv:2510.09817*, 2025.
- [21] Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting touch and vision via cross-modal prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10609–10618, 2019.
- [22] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. *arXiv preprint arXiv:2211.12498*, 2022.
- [23] Yuhang Zheng, Songen Gu, Weize Li, Yupeng Zheng, Yujie Zang, Shuai Tian, Xiang Li, Ce Hao, Chen Gao, Si Liu, et al. Omnivta: Visuo-tactile world modeling for contact-rich robotic manipulation. *arXiv preprint arXiv:2603.19201*, 2026.
- [24] Amir-Hossein Shahidzadeh, Gabriele M Caddeo, Koushik Alapati, Lorenzo Natale, Cornelia Fermüller, and Yiannis Aloimonos. Feelanyforce: Estimating contact force feedback from tactile sensation for vision-based tactile sensors. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 251–257. IEEE, 2025.
- [25] Ruohan Gao, Yiming Dou, Hao Li, Tanmay Agarwal, Jeannette Bohg, Yunzhu Li, Li Fei-Fei, and Jiajun Wu. The objectfolder benchmark: Multisensory learning with neural and real objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17276–17286, 2023.
- [26] Sudharshan Suresh, Zilin Si, Stuart Anderson, Michael Kaess, and Mustafa Mukadam. Midastouch: Monte-carlo inference over distributions across sliding touch. In *Conference on Robot Learning*, pages 319–331. PMLR, 2023.
- [27] Justin Kerr, Huang Huang, Albert Wilcox, Ryan Hoque, Jeffrey Ichnowski, Roberto Calandra, and Ken Goldberg. Self-supervised visuo-tactile pretraining to locate and follow garment features. *arXiv preprint arXiv:2209.13042*, 2022.
- [28] Letian Fu, Gaurav Datta, Huang Huang, William Chung-Ho Panitch, Jaimyn Drake, Joseph Ortiz, Mustafa Mukadam, Mike Lambeta, Roberto Calandra, and Ken Goldberg. A touch, vision, and language dataset for multimodal alignment. In *Forty-first International Conference on Machine Learning*, 2024.
- [29] Chaofan Zhang, Peng Hao, Xiaoge Cao, Xiaoshuai Hao, Shaowei Cui, and Shuo Wang. Vtla: Vision-tactile-language-action model with preference learning for insertion manipulation. *arXiv preprint arXiv:2505.09577*, 2025.