ᗡᑕ BeingBeyond

UniTacHand: Unified Spatio-Tactile Representation for Human-to-Dexterous-Hand Skill Transfer

Chi Zhang[1,2,*]   **Penglin Cai**[1,2,*]   **Haoqi Yuan**[1,2]
Chaoyi Xu[2]   **Zongqing Lu**[1,2,†]

[1]Peking University   [2]BeingBeyond

https://beingbeyond.github.io/UniTacHand/

Abstract

Tactile sensing is crucial for robotic hands to achieve human-level dexterous manipulation, especially in scenarios with visual occlusion. However, its application is often hindered by the difficulty of collecting large-scale real-world robotic tactile data. In this study, we propose to collect low-cost human manipulation data using haptic gloves for tactile-based robotic policy learning. The misalignment between human and robotic tactile data makes it challenging to transfer policies learned from human data to robots. To bridge this gap, we propose **UniTacHand**, a unified representation to align robotic tactile information captured by dexterous hands with human hand touch obtained from gloves. First, we project tactile signals from both human hands and robotic hands onto a morphologically consistent 2D surface space of the MANO hand model. This unification standardizes the heterogeneous data structures and inherently embeds the tactile signals with spatial context. Then, we introduce a contrastive learning method to align them into a unified latent space, trained on only 10 minutes of paired data from our data collection system. Our approach enables zero-shot tactile-based policy transfer from humans to a real robot, generalizing to objects unseen in the pre-training data. We also demonstrate that co-training on mixed data, including both human and robotic demonstrations via **UniTacHand**, yields better performance and data efficienct compared with using only robotic data. **UniTacHand** paves a path toward general, scalable, and data-efficient learning for tactile-based dexterous hands.

Date: December 17, 2025

1 Introduction

Dexterous manipulation for contact-rich tasks remains a grand challenge in robotics [19, 25, 36, 39], largely because the systems lack the rich, multi-modal tactile perception that humans use. While vision-based policies [6, 49, 54] are dominant, they struggle with fundamental limitations like occlusion and ambiguity in complex interactions [23, 50]. Tactile sensing provides a distinct, complementary information stream that addresses these shortcomings [7]. However, despite the advantage of diverse, high-resolution tactile sensors, effectively processing and unifying heterogeneous tactile data remains an open problem [2].

---

*Equal contribution.
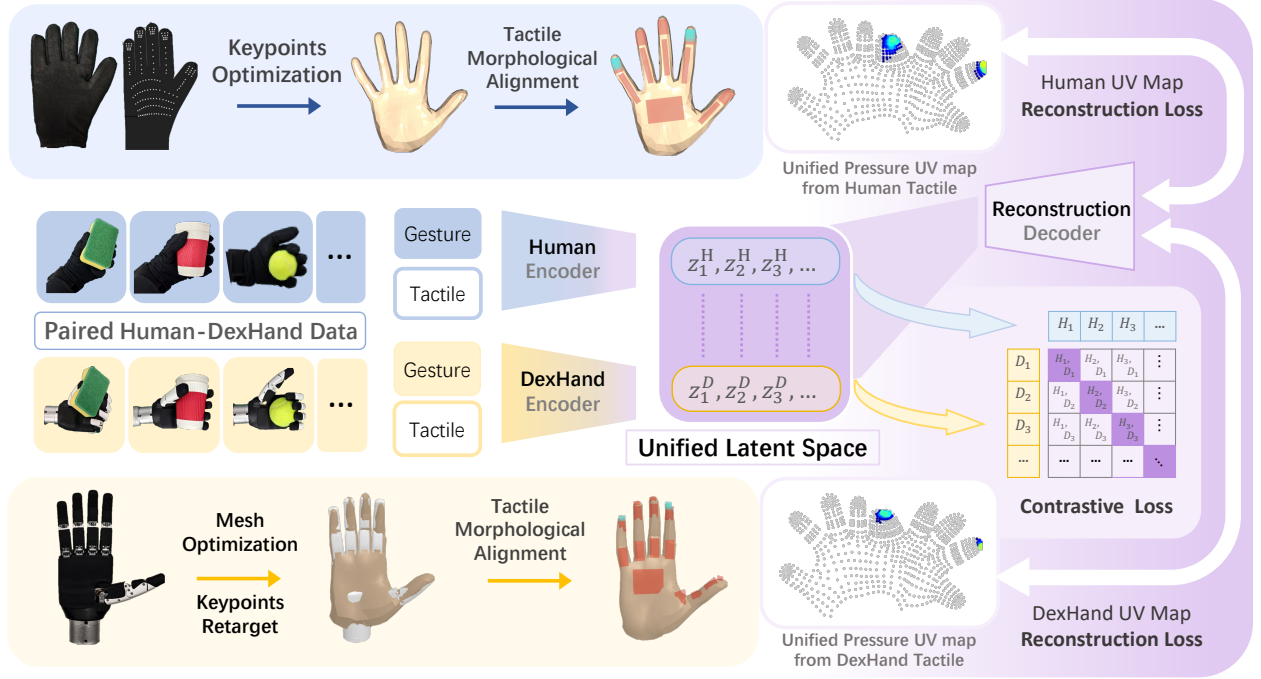†Correspondence to Zongqing Lu <lu@beingbeyond.com>.

Figure 1: **An overview of UniTacHand.** (Left) Stage 1: We project tactile data from both human haptic gloves and robotic hands onto a unified MANO UV map. (Right) Stage 2: We introduce contrastive learning with reconstruction and adversarial losses to align the latent representations. We align the tactility and hand gesture from both sources to the same latent space using a contrastive framework trained with paired data. The unified pressure UV maps serve as accurate prior knowledge to supervise the domain-specific encoders, thereby enriching such a latent space with tactile-grounded information.

To endow robots with autonomous manipulation capabilities, data-driven methods have become increasingly prominent. Reinforcement learning (RL) for robotics, while showing promise for some tasks, often suffers from high sample complexity [17] and the well-known sim-to-real gap – especially pronounced in tasks with complex contact dynamics [3, 4, 16]. Imitation learning (IL) and learning from demonstration (LfD) methods have emerged as a promising alternative, directly mimicking expert data to acquire manipulation skills. However, applying IL to tactile-based dexterous manipulation is fundamentally limited by the general scarcity and low quality of real robot teleoperation data [51, 53]. LfD methods [5, 15, 29], especially those that utilize human data collected by wearable devices, are hindered by the embodiment gap between human and robotic hands [41].

Current efforts to bridge the embodiment gap have largely focused on kinematic retargeting [15, 47], leaving the tactile morphological gap between different robotic hands largely unaddressed. A common workaround is to equip human hands with the same robotic tactile sensors used on the robot for data collection [1, 42]. However, such approaches rely on hardware uniformity, which can be difficult to achieve for many robotic hands and tends to be less cost-effective or scalable for data collection. In this research, we aim to use human data collected by wearable haptic gloves to learn tactile-based robotic manipulation policies. Our focus thus shifts to the following question: how can we design a method to transfer tactile manipulation knowledge from haptic gloves to dexterous robotic hands?

To bridge such cross-embodiment gaps, leveraging a shared canonical representation is essential. In the realm of kinematics and vision, the MANO hand model [34] has been established as a cornerstone, providing a parametric space that unifies diverse hand shapes and poses [13]. MANO model has been widely adopted as a general representation for learning policies on dexterous hands [31, 37, 40, 43, 47]. We argue that this geometric consistency can be extended beyond kinematics to unify tactile perception.

Inspired by this, we introduce **UniTacHand**, a novel unified representation that bridges the tactile gap. Our core idea is to project disparate tactile signals, from human gloves and robotic hands, onto a canonical, morphologically consistent 2D surface space defined by the MANO UV map. Such unification standardizes the heterogeneous data structures and inherently embeds the tactile signals with spatial context. We utilize a contrastive learning framework to align the two domains into a unified latent space. Based on this unified structure, we demonstrate that 10-minute paired data are sufficient to enable, to our knowledge, the first instance of zero-shot tactile-based policy transfer from human data to a physical robot. Policies trained solely on human data can be directly deployed to perform complex downstream tasks. Furthermore, we demonstrate that co-training on mixed data, including both human and (a small number of) robotic demonstrations via UniTacHand, yields a superior capability. This enables policies to achieve better data efficiency with only one-shot demonstrations on the real robot. UniTacHand provides a general, scalable, and data-efficient pathway for transferring human haptic intelligence to policies on tactile dexterous hands.

Our contributions are as follows:

- We propose **UniTacHand**, a novel unified spatio-tactile representation that projects heterogeneous tactile data from different hand morphologies to a canonical 2D figure (MANO UV map), and to a unified latent space (by representation learning from paired human-robot data).

- We demonstrate zero-shot human-to-robot policy transfer for tactile-based manipulation. Policies trained solely on human data via UniTacHand can be directly deployed on a robot to perform complex tactility-based downstream tasks.

- We show that mixing one-shot real robot tactile demonstration with human data provides a superior knowledge for policies, leading to improved performance and data efficiency in terms of human-to-robot policy transfer.

2 Related Work

Canonical hand models as representation space. Our work is built upon the foundation of parametric 3D hand models. Loper [28] first introduced SMPL, a learned parametric model for human bodies. Building on this, Romero [34] proposed MANO, a powerful and expressive model of hand shape and articulation derived from thousands of high-resolution 3D scans. MANO has become a standard in the computer vision community, which is widely employed for visual hand pose and shape estimation from images or videos [11, 32, 48]. Besides, as a kind of motion representation, MANO is also widely used in motion generative models and even vision-language-action (VLA) models [10, 24, 29, 40, 56]. However, its potential as a canonical coordinate frame for tactile data has been largely unexplored. Prior work may project contact onto a mesh [30]. Our work is the first to leverage the intrinsic, morphologically consistent MANO UV map as a unified spatio-tactile representation space specifically for cross-domain human-to-robot policy transfer.

Unified representations for heterogeneous tactile sensors. The diversity of tactile sensors (e.g., visual-tactile sensors [20, 50]) poses a significant domain adaptation challenge. Early efforts to learn a shared latent space [9, 22], typically via auto-encoders, discarded the precise spatial geometry essential for manipulation policies. Recent works seek unified representations, but diverge in their target alignment space. UniTouch [44] aligns to the semantic space of CLIP [33] for zero-shot semantic tasks. Some other works including T3 [52] use physics-based priors (e.g., forces [38]) or learn data-hungry implicit embeddings, which require millions of samples. On the other hand, paired video-tactility data have been leveraged to align force information with visual contact, extracting knowledge and skills from human demonstrations [35, 43]. **UniTacHand** presents a more data-efficient and geometrically grounded alternative. We utilize such spatial context as useful signals, and align it to an explicit morphological space. By leveraging the MANO UV map as a powerful inductive bias, our model learns a aligned embedding from a fraction of the data (a 10-minute collected paired set).

Policy transfer from human data. Learning from demonstration (LfD) has been well-established for visual or kinematic trajectories [3, 21, 29, 55], yet transferring *contact-rich* policies via tactile data remains

an open problem. The challenge stems from the embodiment gap – the profound morphological and sensory mismatch between a human hand and a robotic end effector. For instance, multiple human demonstration datasets are widely used to learn behaviors from human manipulation videos or trajectories [8, 12, 18, 26, 27], as such datasets provide strong priors of performing manipulation tasks [29, 45, 56].

In the field of tactile sensing, prior research has often tailored its approach according to the morphological complexity of the target system. For instance, Feel the Force (FTF) [1] achieved compelling zero-shot transfer by unifying representations of contact forces; however, this was demonstrated only on a parallel-jaw gripper, a simple structure that circumvents the challenges of high-DoF, multi-finger coordination. When it comes to more complex, anthropomorphic hands, MimicTouch [46] relies on a sophisticated three-stage pipeline that includes online residual reinforcement learning. This dependence on an online RL patch to correct the policy on the physical robot reveals that the embodiment gap has not been fundamentally addressed at the representation level. In contrast, our work focuses on resolving the embodiment gap by aligning different domains into a canonical space. This approach enables genuine zero-shot transfer for dexterous hands and enriches the learned knowledge by integrating human demonstrations with one-shot data from real robots.

3 Method

Our objective is to learn a unified tactile-pose representation that bridges the significant morphological and sensor-level gap between human hands and robotic dexterous hands. Our method, **UniTacHand**, achieves this through a two-stage process, as illustrated in figure 1. First, we introduce a canonical representation by projecting heterogeneous tactile data from both human and robotic domains to the 2D UV map of the MANO hand model. Such approaches unify the data structure with the same dimensionality and embeds spatial context, making it possible for policies to learn from a mixture of both human and robotics data. Second, we leverage a small, paired dataset to train encoders using a contrastive learning framework, aligning the representations from both domains into a shared, domain-invariant latent space.

3.1 Unified Representation via MANO UV Maps

The MANO model [34] provides a low-dimensional, differentiable parameterization for the human hand, defined by shape parameters β and pose parameters θ. Critically, it offers a consistent mesh topology ($N_v = 778$ vertices) and a corresponding 2D UV map, which we utilize as our canonical surface space for decoders.

Human haptic data projection. We acquire human demonstration data using a motion capture glove, which provides 3D keypoints for hand pose, and a pressure-sensitive tactile glove, which yields sparse tactile readings $T_H \in \mathbb{R}^{N_H}$ (where N_H is the number of human sensors). To project these sparse signals onto the continuous MANO surface, we perform a one-time morphological annotation. We manually identify and associate the four corner vertices of each rectangular sensing region on the tactile glove with the corresponding vertex indices on the MANO meshes, which defines a patch on the mesh. The pressure reading from a sensor is then distributed across its associated patch, with values for vertices inside the region computed via bilinear interpolation. This process yields a dense tactile representation on the mesh, which is then rasterized into a tactile UV map $U_H^{ori} \in \mathbb{R}^{W \times H}$.

Robotic hand tactile data projection. For the dexterous robotic hand, we leverage its known URDF model and the predefined geometry of its N_R tactile sensors, which provide readings $T_R \in \mathbb{R}^{N_R}$. To bridge the morphological gap to the MANO model, we first perform a one-time optimization to find the optimal MANO shape parameters β^*. This optimization minimizes a joint loss function $\mathcal{L}_{align} = \mathcal{L}_{CD} + w(t)\mathcal{L}_{key}$ (using a static reference pose for both hands). $\mathcal{L}_{CD}$ is the Chamfer Distance loss, computed between 4096 points sampled from the robotic hand's URDF mesh and the MANO mesh. $\mathcal{L}_{key}$ is the keypoint position difference loss based on 21 corresponding keypoints. The keypoint loss ensures correct initial alignment and its weight $w(t)$ is gradually decayed as the optimization proceeds.

Once the optimal shape β^* is fixed, we address the dynamic pose. For each frame, we solve a real-time keypoint-based retargeting problem. We directly optimize for the MANO pose parameters θ that minimize the positional difference between the 21 keypoints derived from the robot's current joint state P_R and the
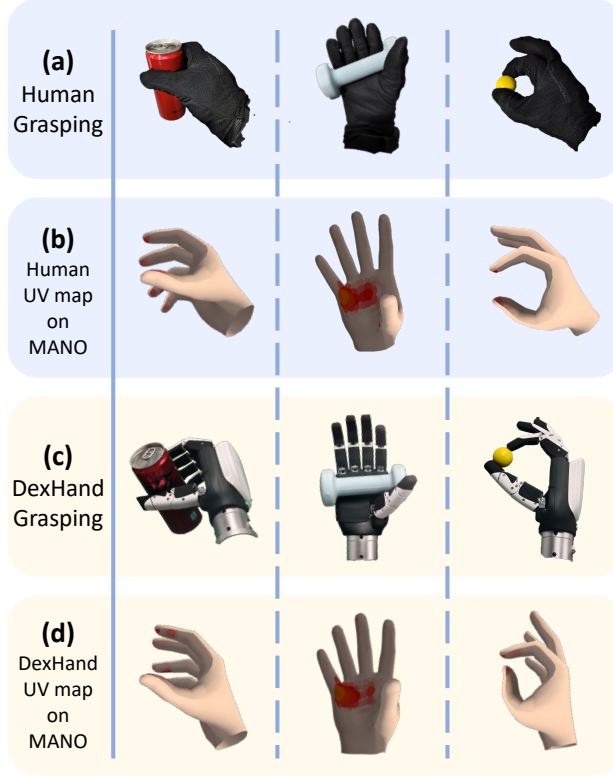
Figure 2: **UV mapping results.** When a human hand (or robotic dexterous hand) grasps an object, the activated tactile lattices on a MANO hand are highlighted in red, with the poses of the hand (actions of fingers and rotations of the wrist) rendered at the same time. In this way, we achieve the unification of hand actions and spatial tactility, as well as the alignment between human hands and dexterous hands in terms of tactile information.

corresponding keypoints on the MANO model (defined by β^* and θ). We then project the robot's tactile sensor readings T_R onto this mesh. We model each sensor as a 2D area in the URDF and project this area onto the optimized MANO mesh, transferring the sensor's reading to the corresponding set of mesh vertices using a weighted interpolation scheme. This process generates a corresponding robotic tactile UV map $U_R^{ori} \in \mathbb{R}^{W \times H}$.

Post-processing and masking. To ensure spatial continuity and account for minor alignment errors, we first apply a Gaussian smoothing kernel to the populated texels for both U_H^{ori} and U_R^{ori}. Let these smoothed maps be U_H^{smooth} and U_R^{smooth}.

To focus the learning process only on active sensing regions, we then create binary masks M_H and M_R. These masks are 1 for texels that correspond to a sensor region and 0 otherwise. The final unified representations, which are used in all subsequent steps, are defined as:

$$U_H = U_H^{smooth} \odot M_H \quad \text{and} \quad U_R = U_R^{smooth} \odot M_R$$

where $\odot$ denotes element-wise multiplication.

3.2 Latent Space Alignment via Representation Learning

Given the unified tactile UV maps and corresponding pose data, $d_H = (U_H, P_H)$ and $d_R = (U_R, P_R)$, where P_H represents human hand keypoints and P_R is the robot hand's 6-DoF pose and joint state, our goal is to learn two encoders, E_H and E_R, that map these inputs into a shared latent space, which is shown in figure 3.
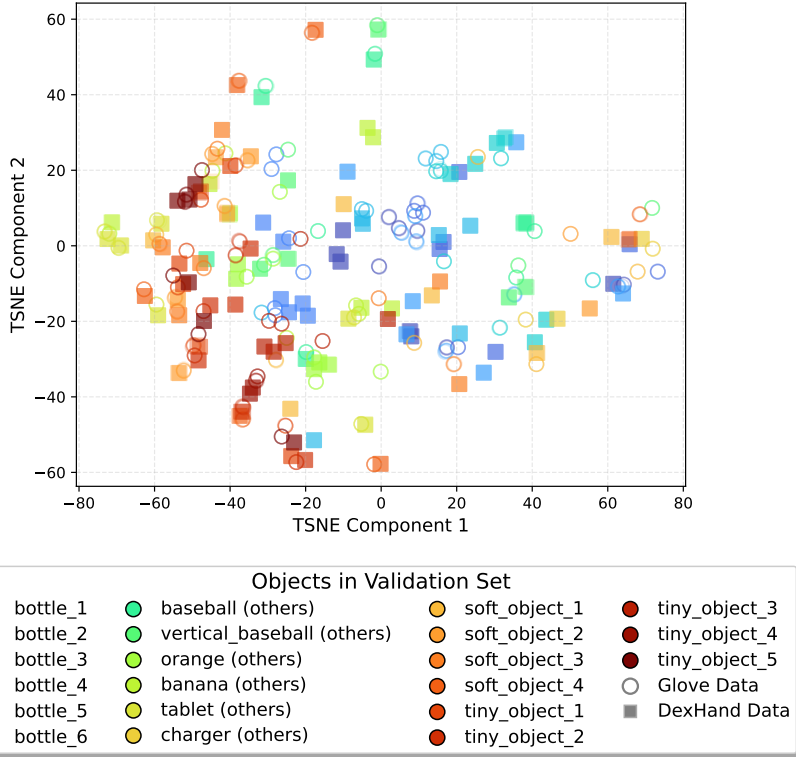
5

Figure 3: **t-SNE visualization of latent space on unseen validation data.** The plot shows latent representations of objects from the validation set (not used in training). Each color represents an object category (e.g., shades of blue for bottle types) and shape indicates the data source (circles: Glove, squares: DexHand). We observe that semantically similar objects, like different bottle or soft_object samples, cluster closely. This demonstrates that the learned latent space is meaningful and captures generalizable object characteristics.

Architecture. Our model consists of two domain-specific encoders, E_H and E_R. Each encoder has a two-stream architecture: a CNN backbone processes the high-dimensional tactile UV map (U_H or U_R), while a separate MLP processes the low-dimensional pose information (P_H or P_R). The features from both streams are then fused to produce a unified latent representation. The latent features are passed to three separate heads: (1) a shared projection head P that maps features into an embedding space for the contrastive loss; (2) two domain-specific decoders, D_H and D_R, trained to reconstruct the original tactile UV maps $\hat{U}_H = D_H(E_H(d_H))$ and $\hat{U}_R = D_R(E_R(d_R))$; and (3) a domain classifier C_D for adversarial alignment.

Paired dataset and augmentation. We collect a paired dataset of human and robotic hand interactions. This dataset involves manipulating a set of 50 common objects, totaling 688 trajectories of data (16k frames at 40 Hz).

As high-fidelity paired tactile data is challenging to acquire at scale, we focus on data-efficient learning. We introduce a novel augmentation strategy based on the observation that linear interpolations of paired trajectories remain valid pairs. Given two paired data samples from the dataset, (d_H^i, d_R^i) and (d_H^j, d_R^j), we generate a new synthetic pair (d_H^{new}, d_R^{new}), where $d_H^{new} = (U_H^{new}, P_H^{new})$ and $d_R^{new} = (U_R^{new}, P_R^{new})$. These new components are generated by interpolating the tactile maps and pose data independently:

$$(U_H^{new}, P_H^{new}) = \left(\lambda_1 U_H^i + (1 - \lambda_1) U_H^j, \ \lambda_2 P_H^i + (1 - \lambda_2) P_H^j \right) \tag{1}$$

$$(U_R^{new}, P_R^{new}) = \left(\lambda_1 U_R^i + (1 - \lambda_1) U_R^j, \ \lambda_2 P_R^i + (1 - \lambda_2) P_R^j \right) \tag{2}$$

where $\lambda_1, \lambda_2 \in [0, 1]$ are randomly sampled coefficients to fully utilize all paired information. This interpolation is applied to the full data and, more powerfully, to subsets of fingers (both their motion and tactile data simultaneously). This physics-informed augmentation significantly expands the effective training data. We also apply standard augmentations such as Gaussian noise and sensor dropout. As shown in figure 4, the ablation results demonstrate that our method together with the physics-informed augmentation strategy is effective, outperforming a baseline method using contrastive learning on a VAE network or our method without data augmentation.
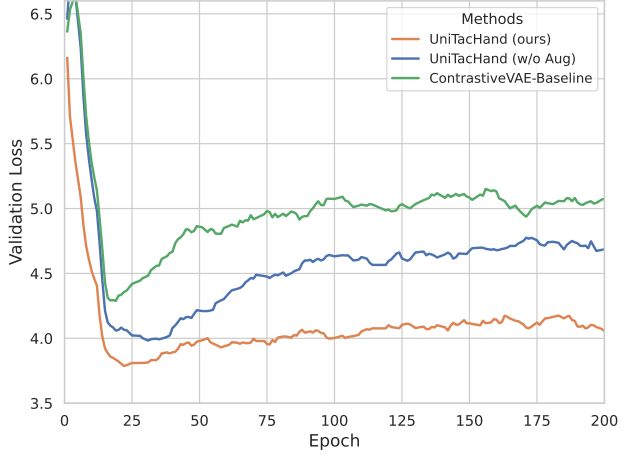

Figure 4: **Ablation study on representation learning.** The validation loss comparison demonstrates that our full UniTacHand framework (orange) outperforms both the ContrastiveVAE Baseline (green) and a variant without our data augmentation strategy (w/o Aug, blue), highlighting the role of augmentation in achieving superior performance.

Training objectives. Our model is trained end-to-end with a composite loss function.

1) Contrastive Alignment Loss: We use a symmetric InfoNCE loss to align the latent embeddings from paired samples. For a batch of B pairs, let $z_H^i = P(E_H(d_H^i))$ and $z_R^i = P(E_R(d_R^i))$ be the projected embeddings. The loss is:

$$\mathcal{L}_{\text{CON}} = -\frac{1}{B} \sum_{i=1}^{B} \left[\log \frac{\exp(s(z_H^i, z_R^i)/\tau)}{\sum_{j=1}^{B} \exp(s(z_H^i, z_R^j)/\tau)} + \log \frac{\exp(s(z_R^i, z_H^i)/\tau)}{\sum_{j=1}^{B} \exp(s(z_R^i, z_H^j)/\tau)} \right] \quad (3)$$

where $s(\cdot, \cdot)$ is the cosine similarity and τ is a temperature hyperparameter. This loss pulls positive pairs (z_H^i, z_R^i) together while pushing negatives (z_H^i, z_R^j) for $i \neq j$ apart.

2) Reconstruction Loss: To ensure the latent space retains high-fidelity information, we task the decoders with reconstructing the original (masked) UV maps. We use a Mean Squared Error (MSE) loss based on the Frobenius norm, computed only over the active sensor regions (as U_H and U_R' are pre-masked):

$$\mathcal{L}_{\text{REC}} = \mathbb{E}_{d_H} \left[||\hat{U}_H - U_H||_F^2 \right] + \mathbb{E}_{d_R} \left[||\hat{U}_R - U_R||_F^2 \right] \quad (4)$$

This loss ensures that the encoders do not discard crucial tactile information during the alignment.

3) Domain-Adversarial Loss: To forge a truly unified latent space for manipulation, we need to ensure that the features are invariant to their domain (human or robot). So our goal is not merely to align tactile maps, but to create a holistic, manipulation-centric latent space where both action (P) and sensation (U).

Therefore, we apply a domain classifier C_D to the fused latent representation $E(d)$. The pose inputs (P_H and P_R) are heterogeneous, and it is a deliberate design choice. By applying adversarial pressure via a Gradient Reversal Layer (GRL) to the fused representation, we force the encoders to align tactile data and normalize the heterogeneous pose structures into a canonical, domain-invariant representation of hand pose.

This ensures the final latent space z is aligned and agnostic to its origin, which is critical for downstream tasks that depend on the interplay of action and touch. The optimization uses a standard Binary Cross-Entropy (BCE) loss:

$$\mathcal{L}_{\text{ADV}} = \mathbb{E}_{d_H}[\text{BCE}(C_D(E_H(d_H)), 0)] + \mathbb{E}_{d_R}[\text{BCE}(C_D(E_R(d_R)), 1)] \quad (5)$$

7

(a) : Tactile Glove and Motion
Capture Glove

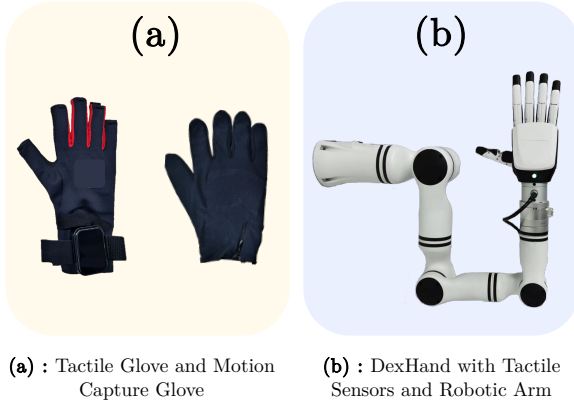(b) : DexHand with Tactile
Sensors and Robotic Arm

Figure 5: **Hardware settings.** (a) The human contact hardware includes a 137-dim tactile glove with pressure-sensitive fiber sensors on its palm and a motion capture glove. (b) The robotic platform features a 6-DoF RealMan arm and a 6-DoF Inspire tactile hand with 1062-dim arrayed pressure sensors on the surfaces of the fingers and palm.

Figure 6: **Examples of the paired dataset.** A subset of our paired object dataset, enabling unified representation alignment by paired manipulation. We choose a variety of objects including fruits, drinks, toys, and other daily items.

Overall objective. The final loss function is a weighted sum of these three components:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{CON}} + \lambda_{\text{REC}}\mathcal{L}_{\text{REC}} + \lambda_{\text{ADV}}\mathcal{L}_{\text{ADV}} \tag{6}$$

where λ_{REC} and λ_{ADV} are hyperparameters balancing the contribution of each term. By optimizing this objective, our encoders E_H and E_R learn to produce representations that are information-rich, and domain-invariant, enabling effective zero-shot and one-shot policy transfer.

4 Experiments

To validate the effectiveness of **UniTacHand**, we conduct a series of experiments designed to answer the following questions:

(1) Is our unified spatial-tactile UV Map effective to bridge the domain gap for simple zero-shot tactile tasks?

(2) Through learning from paired human-robot tactile data, can our unified representation be capable of zero-shot human-to-robot policy transfer, with policies trained solely on human data?

(3) By mixing one-shot robot tactile demonstrations with human data, can policies learn from mixed data via **UniTacHand** show performance gain and data efficiency, demonstrating the capability of performing complex tasks in terms of human-to-robot transfer?

4.1 Experimental Setup

Hardware platform. Our experimental setup, shown in figure 5, consists of a human-side data collection suite and a robot-side deployment platform.

- **Human-side:** We use a custom pressure-sensitive tactile glove providing 137-dimensional tactile readings across the hand. Hand pose is captured by a motion capture glove, which provides 21 keypoint coordinates relative to the wrist.

- **Robot-side:** We use an Inspire tactile hand, which is a 5-finger, 6-DoF anthropomorphic hand equipped with 1062-dimensional tactile sensing capabilities. The hand is mounted on a 6-DoF RealMan robot arm for manipulation tasks. For the robot arm, we use end-effector (EEF) control as the action space, and conduct inverse kinematics (IK) to calculate the target joint angles. For the Inspire hand, the 6 degree-of-freedoms(DoF) naturally corresponds to its 6-dim action space.

Table 1: **Simple zero-shot task evaluation.** Experimental results on `SoftHardPickPlace` and `ObjectLocating` tasks. We test each for 20 times and calculate the success rate (%). Our policies trained using UV Map projection significantly outperform the PatchMatch baselines.

Task	PatchMatch	UV-Direct (Ours)
SoftHardPickPlace	25.0	**85.0**
ObjectLocating	55.0	**100.0**

Paired dataset for unified alignment in representation learning. During representation learning, we train the alignment encoders using a small, 10-minute paired dataset collected from both human and robotic dexterous hands simultaneously figure 6. During data collection, we use retargeting algorithms including DexPilot [13] to map MANO keypoints captured from human hand to the Inspire dexterous hand by teleoperation, ensuring that both human and the robot hand are performing the same action in a similar behavioral manner. We collect such paired data on 50 kinds of diverse paired objects , performing the same manipulation process including grasping simultaneously with various hand poses. We collect 688 trajectories containing 16k frames in total, at a frequency of 40 Hz.

Compared methods and baselines. We compare our approach against several baselines:

- **Simple patch-wise mapping (PatchMatch):** A baseline that directly matches tactile sensor patches from the human glove to corresponding patches on the robotic hand based on pre-defined spatial correspondence. The lattices on both hands are divided into several partitions, with each paired partition performing a one-to-one mapping.

- **Direct transfer via UV Map (UV-Direct):** A policy is trained on human tactile data projected to the MANO UV map. Such a policy is then directly deployed on the robot hand, using the robotic tactile data projected to the UV map as input.

- **UniTacHand (ours):** Our full method, which uses contrastive learning to align the latent spaces of the human and robot tactile data, and train an encoder-decoder framework in a manner of representation learning. The policy is trained on human data using the human encoder, and transferred to the real robot by using the robot encoder during inference.

4.2 Unified UV Map for Zero-Shot Policy Transfer

We first evaluate the fundamental effectiveness of our unified UV map representation on simple, open-loop tactile tasks in a zero-shot setting. In this part, we conduct evaluations on two different tactility-based tasks:

1. **SoftHardPickPlace:** the robot first picks up the object on the table, and places it into the corresponding basket depending on whether it is a soft or hard object. This task requires the ability to distinguish tactile senses according to the softness of the object during contact-rich manipulation.

2. **ObjectLocating:** the robot hand touches an object on the tabletop to locate the object according to the tactility, and adjust its own wrist pose following the locating result before finally grasping. In this task, the initial location of the object is randomly sampled within a rectangular range of 10 cm × 20 cm.

Our policies are trained using only human data and transferred directly to the robot, with results listed in table 1. The experimental results show that our policies trained using UV Map projection significantly outperform the PatchMatch baselines, indicating the effectiveness of our canonical UV Map projection.

4.3 Unified Representation Learning from Paired Human-Robot Data for Policy Transfer

To evaluate our unified representation learning framework in terms of human-to-robot policy transfer, we further investigate two more complex zero-shot transfer tasks (as shown in figure 7):
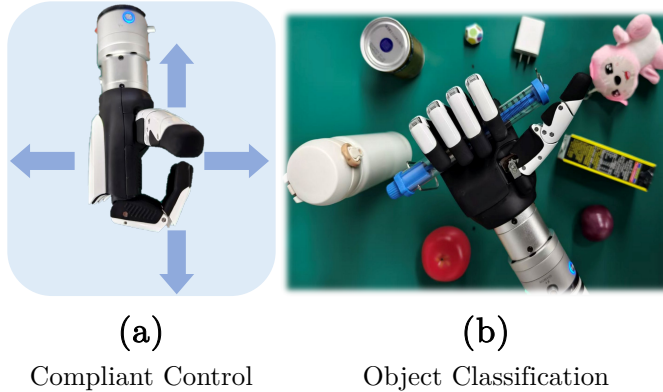
(a)
Compliant Control

(b)
Object Classification

Figure 7: **Illustration of the tasks `CompliantControl` and `ObjectClassification`.** (a) For `CompliantControl`, we fix the action of the dexterous hand at a constant pose, and drag parts of the hand to apply external forces on the tactile sensors. The aim of the task is to make the end effector move towards the same direction as the external forces. (b) For `ObjectClassification`, we choose 10 objects that are totally unseen in the paired data used to train the representation encoders. The aim of the task is to classify the correct category of the objects according to tactile information when grasped in hand.

Table 2: **Detailed results of the tasks `CompliantControl` and `ObjectClassification`.** We calculate the accuracy (%) for each task on each method.

Tasks		PatchMatch	UV-Direct	UniTacHand (Ours)
CompliantControl		10.0	36.0	**40.0**
ObjectClassification	Human Validation Set	43.2	**71.6**	59.5
	Real Robot Test Set	15.7	18.9	**38.6**

1. **CompliantControl:** while the robot hand holds a fixed action, the end-effector of the robot arm should follow the human when external forces are applied to the dexterous hand. Such a task requires the capability of spatial perception related to the direction of the external force, which ultimately determines the moving direction of the end-effector. This task is considered successful only when the moving direction inferred from the policy is consistent with the external force applied by human.

2. **ObjectClassification:** we choose 10 objects different in shape, textures, geometric features, as well as physical characteristics, to form an object set for the classification task. Holding the object in hand and taking the corresponding tactile information, the policy should infer the category of the object. These 10 objects contain both seen objects in the paired data used for representation learning, and unseen ones.

We compare the performances of **UniTacHand**, **UV-Direct**, as well as the **PatchMatch** baselines in table 2. For the task `CompliantControl`, we test in five different directions for 50 times and calculate the accuracy (consistency with the direction of the external force). For the task `ObjectClassification`, we evaluate the classification accuracy on both a held-out human validation set and on the physical robot test set (for zero-shot transfer).

According to the experimental results, due to the cluttered data of human tactile demonstrations, simply applying one-to-one PatchMatch does not work on the task `CompliantControl`, while our UniTacHand manages to overcome the difficulty of distribution shift from human data to real robot deployment. Besides, according to the task `ObjectClassification`, the results clearly demonstrate that while direct UV map projection is a useful representation (which yields a relatively high accuracy in human validation set), it is insufficient to bridge the domain gap, especially due to the different manipulation poses, gestures, and

Table 3: **One-shot manipulation task success rate (%).** We compare our methods with both visual-only baselines and methods that are solely trained on real robot data.

	Methods	Success Rate
Robot Data Only	**R-Visual-Only**	43.3
	R-Visual-Tactile	50.0
Human Data w/ One-Shot Robot Data	**PatchMatch**	56.7
	UV-Direct	63.3
	UniTacHand (Ours)	**73.3**

the morphological mismatch in human-robot pairs (a relatively low accuracy in the real robot test set). In this case, our contrastive alignment framework (UniTacHand) successfully learns a shared embedding space, enabling robust zero-shot transfer of a complex classification policy.

Overall, evaluations on both tasks demonstrate that with a more precise and physically consistent representation learned from paired human-robot data, not only do our policies gain better performances, but also they are more capable of generalization in the face of unseen objects and out-of-distribution cases.

4.4 One-Shot Learning for Human-to-Robot Policy Transfer

Finally, we evaluate the ability of our UniTacHand when a one-shot robot demonstration is given. We consider a task with multi-modal inputs: both vision and tactility. In this task, we choose three pairs of objects that share similar appearances, but result in different distributions of tactility due to their physical attributes. For instance, a soft orange and a hard one, an empty bottle and the exact same one full of water – each pair of objects can be similar in vision, but can be distinct in tactility when being held in hand.

We evaluate our methods and the **PatchMatch** baseline on this task, together with visual-only policies (**R-Visual-Only**) as well as policies that only learn from real robot data (**R-Visual-Tactile**). As listed in table 3, our methods strongly outperform various baselines. The results that **R-Visual-Only** performs worse than methods that utilize tactile information as inputs indicate that in such cases that visual information often makes policy confusing, adding tactility can help the policy distinguish and make feasible decisions. On the other hand, the results that **R-Visual-Tactile** performs no better than those trained on mixed data (human data with one-shot real robot data) demonstrate the effectiveness of human data in the domain of human-to-robot policy transfer. Despite the fact that human data cannot be directly applied during real robot deployment, such rich and diverse information still provides essential knowledge and helps the process of policy learning.

5 Conclusion

In this paper, we propose **UniTacHand** to unify tactile representations between human hands and dexterous robotic hands. By mapping heterogeneous tactile information from different embodiments onto MANO UV maps and using contrastive learning, our method achieves either zero-shot or one-shot human-to-robot policy transfer for tactile-based manipulation tasks. Such unified representations, serving as precise and physically consistent mappings between human hands and dexterous hands, make it possible to utilize large-scale tactile data collected from humans, which is much easier to obtain compared to robot data. Future work could scale human tactile data collection and incorporate tactile modality into multi-modal foundation models, such as vision–language–action models, to build generalist robotic systems.

References

[1] Ademi Adeniji, Zhuoran Chen, Vincent Liu, Venkatesh Pattabiraman, Raunaq Bhirangi, Siddhant Haldar, Pieter Abbeel, and Lerrel Pinto. Feel the force: Contact-driven learning from humans. arXiv preprint arXiv:2506.01944, 2025.

[2] Alessandro Albini, Mohsen Kaboli, Giorgio Cannata, and Perla Maiolino. Representing data in robotic tactile perception–a review. arXiv preprint arXiv:2510.10804, 2025.

[3] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. Robotics and autonomous systems, 57(5):469–483, 2009.

[4] Yevgen Chebotar, Ankur Handa, Viktor Makoviychuk, Miles Macklin, Jan Issac, Nathan Ratliff, and Dieter Fox. Closing the sim-to-real loop: Adapting simulation randomization with real world experience. In 2019 International Conference on Robotics and Automation (ICRA), pp. 8973–8979. IEEE, 2019.

[5] Zoey Qiuyu Chen, Karl Van Wyk, Yu-Wei Chao, Wei Yang, Arsalan Mousavian, Abhishek Gupta, and Dieter Fox. Dextransfer: Real world multi-fingered dexterous grasping with minimal human demonstrations. arXiv preprint arXiv:2209.14284, 2022.

[6] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. The International Journal of Robotics Research, 2023.

[7] Ravinder S. Dahiya, Giorgio Metta, Maurizio Valle, and Giulio Sandini. Tactile sensing—from humans to humanoids. IEEE Transactions on Robotics, 26(1):1–20, 2010. doi: 10.1109/TRO.2009.2033627.

[8] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 12943–12954, 2023.

[9] 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.

[10] Qifan Fu, Xu Chen, Muhammad Asad, Shanxin Yuan, Changjae Oh, and Gregory Slabaugh. Handrawer: Leveraging spatial information to render realistic hands using a conditional diffusion model in single stage. arXiv preprint arXiv:2503.02127, 2025.

[11] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10833–10842, 2019.

[12] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 18995–19012, 2022.

[13] Ankur Handa, Karl Van Wyk, Wei Yang, Jacky Liang, Yu-Wei Chao, Qian Wan, Stan Birchfield, Nathan Ratliff, and Dieter Fox. Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system. In 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 9164–9170. IEEE, 2020.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. CoRR, 2015.

[15] Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. arXiv preprint arXiv:2505.11709, 2025.

[16] Ziye Huang, Haoqi Yuan, Yuhui Fu, and Zongqing Lu. Efficient residual learning with mixture-of-experts for universal dexterous grasping. arXiv preprint arXiv:2410.02475, 2024.

[17] Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. The International Journal of Robotics Research, 32(11):1238–1274, 2013.

[18] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 10138–10148, 2021.

[19] Arjun Lakshmipathy and Nancy S Pollard. Contactmpc: Towards online adaptive control for contact-rich dexterous manipulation. In 2nd Workshop on Dexterous Manipulation: Design, Perception and Control (RSS), 2024.

[20] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian H. Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, Dinesh Jayaraman, and Roberto Calandra. DIGIT: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. CoRR, 2020.

[21] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: Contrastive unsupervised representations for reinforcement learning. In Proceedings of the 37th International Conference on Machine Learning (ICML), 2020.

[22] Michelle A Lee, Yuke Zhu, Peter Zachares, Matthew Tan, Krishnan Srinivasan, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Learning multimodal representations for contact-rich tasks. IEEE Transactions on Robotics, 36(3):582–596, 2020.

[23] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quigley. Learning hand-eye coordination for robotic grasping with deep learning. The International Journal of Robotics Research, 37(4-5):421–436, 2018.

[24] Qixiu Li, Yu Deng, Yaobo Liang, Lin Luo, Lei Zhou, Chengtang Yao, Lingqi Zeng, Zhiyuan Feng, Huizhi Liang, Sicheng Xu, Yizhong Zhang, Xi Chen, Hao Chen, Lily Sun, Dong Chen, Jiaolong Yang, and Baining Guo. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos, 2025. URL https://arxiv.org/abs/2510.21571.

[25] Zhixuan Liang, Yao Mu, Yixiao Wang, Tianxing Chen, Wenqi Shao, Wei Zhan, Masayoshi Tomizuka, Ping Luo, and Mingyu Ding. Dexhanddiff: Interaction-aware diffusion planning for adaptive dexterous manipulation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 1745–1755, 2025.

[26] Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21740–21751, 2024.

[27] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21013–21022, 2022.

[28] M. et al. Loper. Smpl: A skinned multi-person linear model, 2015.

[29] Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. Being-h0: Vision-language-action pretraining from large-scale human videos. arXiv preprint arXiv:2507.15597, 2025.

[30] Yiyao Ma, Kai Chen, Kexin Zheng, and Qi Dou. Contact map transfer with conditional diffusion model for generalizable dexterous grasp generation. arXiv preprint arXiv:2511.01276, 2025.

[31] Zhao Mandi, Yifan Hou, Dieter Fox, Yashraj Narang, Ajay Mandlekar, and Shuran Song. Dexmachina: Functional retargeting for bimanual dexterous manipulation. arXiv preprint arXiv:2505.24853, 2025.

[32] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1496–1505, 2022.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. CoRR, 2021.

[34] J. et al. Romero. Embodied hands: Modeling and capturing hands and bodies together, 2017.

[35] Yuxin Ray Song, Jinzhou Li, Rao Fu, Devin Murphy, Kaichen Zhou, Rishi Shiv, Yaqi Li, Haoyu Xiong, Crystal Elaine Owens, Yilun Du, Yiyue Luo, Xianyi Cheng, Antonio Torralba, Wojciech Matusik, and Paul Pu Liang. Opentouch: Bringing full-hand touch to real-world interaction, 2025. URL https://arxiv.org/abs/2512.16842.

[36] HJ Suh, Tao Pang, Tong Zhao, and Russ Tedrake. Dexterous contact-rich manipulation via the contact trust region. arXiv preprint arXiv:2505.02291, 2025.

[37] Congyi Wang, Feida Zhu, and Shilei Wen. Memahand: Exploiting mesh-mano interaction for single image two-hand reconstruction. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 564–573, 2023.

[38] Tianhao Wu, Jinzhou Li, Jiyao Zhang, Mingdong Wu, and Hao Dong. 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.

[39] Zheng Wu, Wenzhao Lian, Vaibhav Unhelkar, Masayoshi Tomizuka, and Stefan Schaal. Learning dense rewards for contact-rich manipulation tasks. In 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 6214–6221. IEEE, 2021.

[40] Pengfei Xie, Wenqiang Xu, Tutian Tang, Zhenjun Yu, and Cewu Lu. Ms-mano: Enabling hand pose tracking with biomechanical constraints. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2382–2392, 2024.

[41] Mengda Xu, Zhenjia Xu, Cheng Chi, Manuela Veloso, and Shuran Song. Xskill: Cross embodiment skill discovery. In Conference on robot learning, pp. 3536–3555. PMLR, 2023.

[42] Mengda Xu, Han Zhang, Yifan Hou, Zhenjia Xu, Linxi Fan, Manuela Veloso, and Shuran Song. Dexumi: Using human hand as the universal manipulation interface for dexterous manipulation. arXiv preprint arXiv:2505.21864, 2025.

[43] Xinyue Xu, Jieqiang Sun, Siyuan Chen, Lanjie Ma, Ke Sun, Bin Zhao, Jianbo Yuan, Sheng Yi, Haohua Zhu, Yiwen Lu, et al. Dexcanvas: Bridging human demonstrations and robot learning for dexterous manipulation. arXiv preprint arXiv:2510.15786, 2025.

[44] 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, pp. 26340–26353, 2024.

[45] Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, et al. Egovla: Learning vision-language-action models from egocentric human videos. arXiv preprint arXiv:2507.12440, 2025.

[46] Kelin Yu, Yunhai Han, Qixian Wang, Vaibhav Saxena, Danfei Xu, and Ye Zhao. Mimictouch: Leveraging multi-modal human tactile demonstrations for contact-rich manipulation. arXiv preprint arXiv:2310.16917, 2023.

[47] Haoqi Yuan, Bohan Zhou, Yuhui Fu, and Zongqing Lu. Cross-embodiment dexterous grasping with reinforcement learning. arXiv preprint arXiv:2410.02479, 2024.

[48] Haoqi Yuan, Yu Bai, Yuhui Fu, Bohan Zhou, Yicheng Feng, Xinrun Xu, Yi Zhan, Börje F Karlsson, and Zongqing Lu. Being-0: A humanoid robotic agent with vision-language models and modular skills. arXiv preprint arXiv:2503.12533, 2025.

[49] Haoqi Yuan, Ziye Huang, Ye Wang, Chuan Mao, Chaoyi Xu, and Zongqing Lu. Demograsp: Universal dexterous grasping from a single demonstration. arXiv preprint arXiv:2509.22149, 2025.

[50] Wenzhen Yuan, Siyuan Dong, and Edward H. Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. Sensors, 17(12), 2017. ISSN 1424-8220. doi: 10.3390/s17122762. URL https://www.mdpi.com/1424-8220/17/12/2762.

[51] Han Zhang, Songbo Hu, Zhecheng Yuan, and Huazhe Xu. Doglove: Dexterous manipulation with a low-cost open-source haptic force feedback glove. arXiv preprint arXiv:2502.07730, 2025.

[52] Jialiang Zhao, Yuxiang Ma, Lirui Wang, and Edward H. Adelson. Transferable tactile transformers for representation learning across diverse sensors and tasks, 2024. URL https://arxiv.org/abs/2406.13640.

[53] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. arXiv preprint arXiv:2304.13705, 2023.

[54] Yifan Zhong, Xuchuan Huang, Ruochong Li, Ceyao Zhang, Zhang Chen, Tianrui Guan, Fanlian Zeng, Ka Num Lui, Yuyao Ye, Yitao Liang, et al. Dexgraspvla: A vision-language-action framework towards general dexterous grasping. arXiv preprint arXiv:2502.20900, 2025.

[55] Bohan Zhou, Haoqi Yuan, Yuhui Fu, and Zongqing Lu. Learning diverse bimanual dexterous manipulation skills from human demonstrations. arXiv preprint arXiv:2410.02477, 2024.

[56] Bohan Zhou, Yi Zhan, Zhongbin Zhang, and Zongqing Lu. Megohand: Multimodal egocentric hand-object interaction motion generation. arXiv preprint arXiv:2505.16602, 2025.

A Implementation Details

This appendix provides comprehensive details regarding data processing, network architectures, and training procedures utilized in our **UniTacHand** framework, supplementing the methodology described in section 3.

A.1 Data Unification and Preprocessing

MANO Model Configuration. We utilize the standard MANO model [34] featuring $N_v = 778$ vertices and $N_f = 1538$ faces. All tactile projections are rasterized into a normalized 2D UV map with a resolution of $W \times H = 1024 \times 1024$.

Human Data Projection. The morphological alignment between the tactile glove and the MANO mesh is performed via a one-time manual annotation. The glove contains $N_H = 137$ sensor arrays. We associate the four corners of each sensor's rectangular region with corresponding vertex indices on the MANO mesh. During processing, pressure values are distributed to vertices within these patches using bilinear interpolation to ensure smooth gradients.

Robotic Hand Alignment. To bridge the morphological gap, we perform a two-stage alignment process. First, an offline shape optimization is conducted to determine the optimal MANO shape parameters β^* by minimizing the alignment loss over 10000 iterations:

- **Chamfer Distance ($\mathcal{L}_{\mathbf{CD}}$):** Computed between 4096 points uniformly sampled from the robot's URDF mesh surface and 4096 points from the MANO mesh.

- **Keypoint Loss ($\mathcal{L}_{\mathbf{key}}$):** Defined as the L_2 distance between 21 semantically corresponding keypoints on both hands.

- **Loss Scheduling:** The weight $w(t)$ for $\mathcal{L}_{\text{key}}$ follows a linear decay schedule $w(t) = \max(0, 1.0 - t/2500)$, prioritizing global structural alignment in early iterations and surface detail refinement in later stages.

Second, for online processing, we employ a real-time pose retargeting module. We optimize the MANO pose parameters θ frame-by-frame using the Adam optimizer to minimize the keypoint discrepancies between the robot's current joint state and the MANO model.

Post-processing. To mitigate rasterization artifacts, a Gaussian smoothing kernel (size 5×5, $\sigma = 0.5$) is applied to the initial UV maps U_H^{ori} and U_R^{ori}. The resulting maps are masked by binary validity masks M_H and M_R and normalized to $[0, 1]$ before network input.

A.2 Network Architecture

The UniTacHand architecture adopts a partitioned, region-aware design to handle the morphological discrepancies between the human glove and the robotic hand. It comprises domain-specific encoders (E_H, E_R) for tactile and pose modalities, followed by a shared latent fusion module.

Tactile Encoders (E_H^{tac}, E_R^{tac}). Unlike standard monolithic encoders, we propose a partition-based encoding strategy that processes specific morphological regions independently before global aggregation.

- **Human Tactile Encoder (E_H^{tac}):** The 137-dimensional glove input is divided into 7 semantic regions: *thumb, index, middle, ring, pinky, palm,* and *bend.* Each region $\mathcal{R}_i$ is processed by a specific lightweight MLP branch.
    - *Branch Architecture:* The finger branches utilize a 2-layer MLP (hidden dims $[8, 8]$ and $[8, 4]$), while the palm branch uses a deeper structure $[16, 8, 8]$ to capture complex contact patterns.
    - *Aggregation:* The features from all branches are concatenated and passed through a global mixing MLP with hidden units $[128]$ and an output dimension of 64 (D_{tac}).

- **Robot Tactile Encoder (E_R^{tac}):** The 1062-dimensional tactile array is segmented into 17 distinct spatial patches (e.g., *finger_phalanges, palm*). Each patch is treated as a local image and processed by a dedicated CNN branch selected based on the patch resolution:

    - *Small Patches (3×3):* Processed by a single convolutional layer ($C_{out} = 1, k = 2, s = 1$).

    - *Standard Patches ($12 \times 8, 10 \times 8$):* Processed by a 2-layer CNN block. Layer 1: ($C_{out} = 1, k = 3, s = 2$); Layer 2: ($C_{out} = 2, k = 2, s = 1$).

    - *Large Patches (Palm 8×14):* Processed by a 2-layer CNN block with a larger receptive field. Layer 1: ($C_{out} = 1, k = 4, s = 2$); Layer 2: ($C_{out} = 2, k = 2, s = 1$).

    - *Aggregation:* Feature maps from all branches undergo adaptive average pooling to a fixed 4×4 spatial size, are flattened, concatenated, and fused via a global mixing MLP (hidden [128], output 64).

Pose Encoders (E_H^{pose}, E_R^{pose}). Pose information is encoded via multi-layer perceptrons to align the kinematic spaces.

- **Human Side:** The 60-dimensional input is processed by a 4-layer MLP with hidden dimensions $[64, 32, 64]$ and a final projection to 32 (D_{pose}).

- **Robot Side:** The 6-dimensional joint angles are processed by a 3-layer MLP with hidden dimensions $[32, 64]$ and a final projection to 32 (D_{pose}).

Fusion and Projection. The extracted tactile and pose features are concatenated ($D_{fused} = 32 + 64 = 96$) and mapped to the shared latent space.

- **Post-Concat Fusion:** A transition MLP with hidden size 128 reduces the dimensionality to 64.

- **Shared Projection Head:** To obtain the final unified representation $z \in \mathbb{R}^{32}$, the features pass through a shared MLP block with hidden dimensions $[128, 128]$.

Decoders and Discriminator. To validate the representation quality, we employ symmetric decoders (D_H, D_R) to reconstruct the UV representations from the shared latent space z. Unlike the partition-based encoders, the decoders utilize a global Multi-Layer Perceptron (MLP) architecture. Specifically, the network expands the latent vector through hidden layers with dimensions $[128, 256]$ and projects it to a final output dimension of 391, corresponding to the UV feature space for both the human glove and the Inspire hand. The domain discriminator C_D retains a 3-layer MLP structure $[128, 64, 32, 16, 1]$ with LeakyReLU activations, receiving input via a Gradient Reversal Layer (GRL) to enforce domain invariance.

A.3 Training Protocol

Dataset Split. The paired dataset (688 trajectories, $\sim$16k frames) is split into training and validation sets with a ratio of 70%:30%. Splits are stratified by object instance to ensure zero-shot evaluation on unseen objects.

Optimization. We train end-to-end using the Adam optimizer ($lr = 5e - 5$, weight decay$=1e - 5$) with a batch size of $B = 1024$ for 200 epochs.

Loss Configuration. The total loss weights are set to $\lambda_{\text{REC}} = 1.0$ and $\lambda_{\text{ADV}} = 0.5$. The InfoNCE temperature is $\tau = 0.1$.

Augmentation Details. Our proposed interpolation augmentation samples mixing coefficients $\lambda_1, \lambda_2 \sim U(0, 1)$. Standard augmentations include additive Gaussian noise ($\sigma = 0.03$) and random sensor dropout (probability $p = 0.005$).
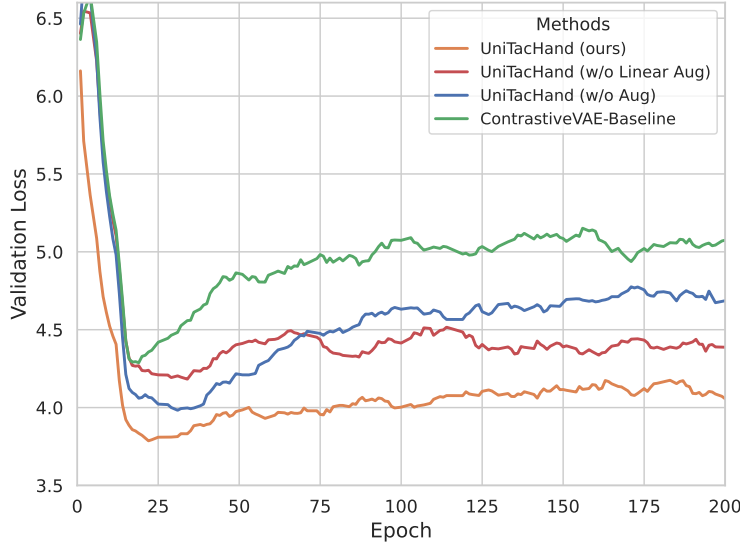
Figure 8: **Ablation on augmentation strategy.** Comparison of validation loss curves. We compare our full method (Orange) against a variant without our specific linear interpolation (Red) and a baseline without any augmentation (Blue), demonstrating the additive value of our proposed strategy.

B Ablation Study

To validate the design choices of UniTacHand, we conduct extensive ablation studies focusing on data augmentation strategies and loss components.

B.1 Impact of Augmentation Strategies

One key contribution of our method is the physics-informed interpolation augmentation. To rigorously quantify its impact and disentangle it from standard regularization techniques, we trained two additional variants of our model:

- **w/o Linear Aug (Red):** Removes only our proposed paired interpolation strategy.
- **w/o Aug (Blue):** Removes all data augmentations, including Gaussian blur and dropout.

As shown in figure 8, the model trained with our full augmentation strategy (Orange curve) exhibits the fastest convergence and the lowest final validation loss.

Comparing the variants, removing all augmentations (Blue curve) results in the highest loss. However, simply adding these standard augmentations is not enough; the gap between the red curve (w/o Linear Aug) and the orange curve (Ours) demonstrates that our physics-informed interpolation effectively densifies the training manifold, significantly boosting generalization to unseen contact configurations beyond what standard noise injection can achieve.

B.2 Impact of Decoder Configurations

We further investigate the effectiveness of our latent representation learning by comparing different decoder configurations. Specifically, we evaluate our default **Dual Decoder** against single-branch variants (**DexHand-only** and **Human-only**) and a baseline that directly reconstructs **Raw Tactile** signals.

As illustrated in figure 9, the Raw Tactile Decoder (Red curve) yields the highest validation loss throughout the training. We attribute this to the high dimensionality and noisy details inherent in raw sensor data, which make it challenging for the model to extract meaningful semantic representations compared to the structured geometry of UV maps.
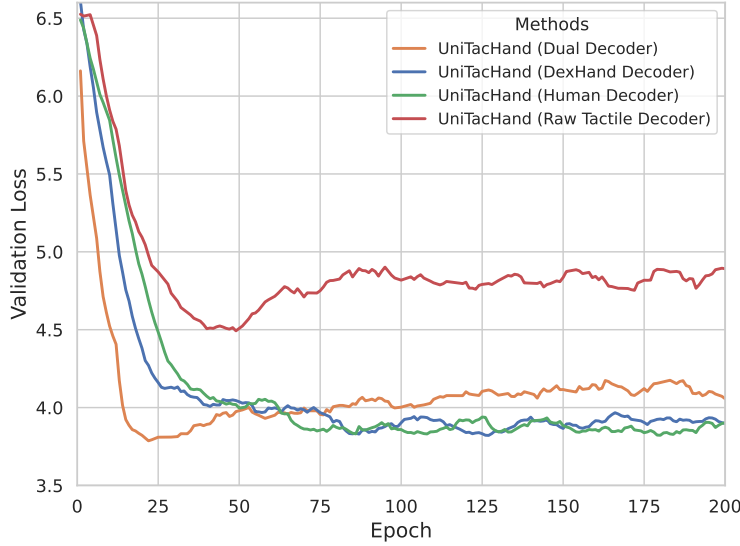
Figure 9: **Comparison of decoder architectures.** The Dual Decoder (Orange) achieves the fastest initial convergence, while the Raw Tactile baseline (Red) performs the worst due to signal noise. The subsequent rise in the Orange curve indicates overfitting due to limited paired data capacity.

In contrast, all UV-based decoders achieve significantly lower losses, confirming that both Human and DexHand UV maps provide effective supervision signals. Notably, the **Dual Decoder (Orange)** exhibits the fastest convergence rate in the early stages, suggesting that leveraging constraints from both modalities simultaneously accelerates feature learning.

However, we also observe that the Dual Decoder begins to overfit after approximately 25 epochs (loss increases), whereas the single-decoder variants (Blue and Green) maintain stable convergence. We hypothesize that while the dual-objective provides stronger gradients initially, our current volume of paired data may be insufficient to support the simultaneous high-fidelity reconstruction of two complex modalities without overfitting. Consequently, while the Dual setup is powerful, single decoders offer a trade-off with slightly slower convergence but greater stability on this dataset.

B.3 Alignment Visualization

To qualitatively assess the alignment quality, we visualize the cosine similarity matrix of latent representations for a paired trajectory from the validation set. As shown in figure 10, the x and y axes represent the time steps of the Human and Robot tactile sequences, respectively. An ideal alignment should manifest as a sharp, high-similarity diagonal.

Comparing the visualizations reveals distinct behaviors:

- **Ours (figure 10a):** Displays a crisp diagonal line with a clean background. This indicates that our method successfully learns a temporally consistent and distinct representation, where specific contact states are uniquely matched across modalities.

- **w/o UV Decoder (figure 10b):** Exhibits minor block-like confusion around the diagonal. Without the dense geometric supervision from the UV map, the model struggles to differentiate between spatially adjacent tactile patterns, leading to local aliasing.

- **w/o Augmentation (figure 10c):** Suffers from large-scale confusion areas (high similarity in off-diagonal regions). This confirms that without our physics-informed augmentation, the model overfits to specific signal amplitudes rather than learning robust contact features, causing it to confuse temporally distant frames.

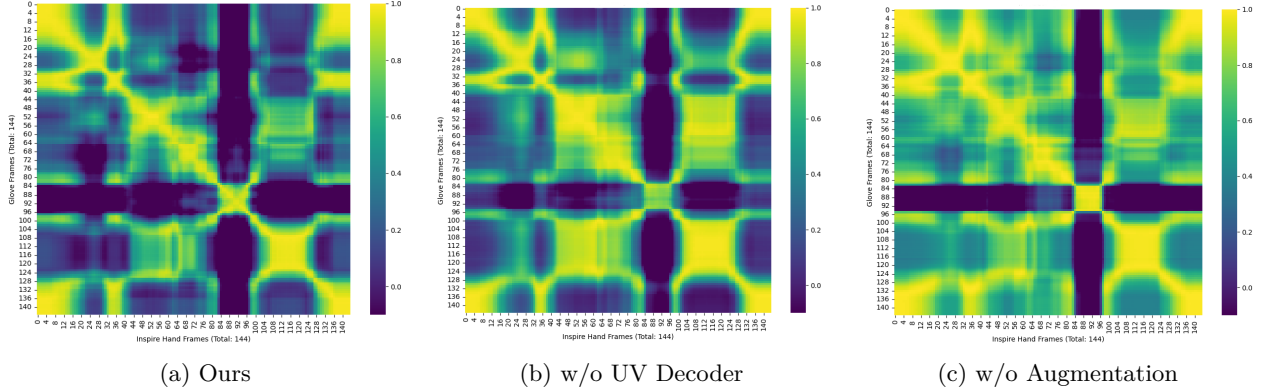| (a) Ours | (b) w/o UV Decoder | (c) w/o Augmentation |

Figure 10: **Latent similarity heatmaps.** Comparison of feature alignment quality on a paired validation trajectory. (a) **Ours** shows a distinct diagonal with clear boundaries. (b) Removing the UV decoder leads to local ambiguities (small blocky confusion). (c) Removing augmentation results in widespread confusion off the diagonal.
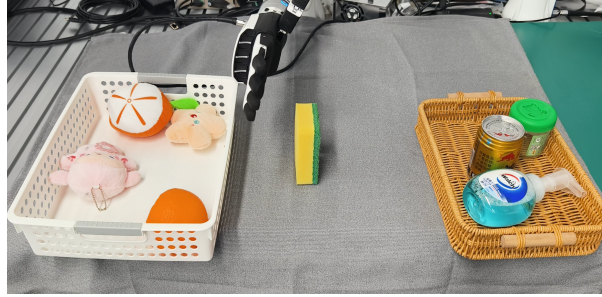

Figure 11: **An example of the task `SoftHardPickPlace`.** The robot picks up the object, and places it into one of the two baskets according to whether the object is soft or hard.

C   Tasks and Policy Training

In this section, we describe the details of the task definitions, including the basic settings of observation and action spaces, the location of the objects, as well as the success conditions. We also present the details of training policies for these tasks.

C.1   Task Description

In our experiments section 4, we choose five different tasks varying from perceptions on the numerical magnitude of tactile values to perceptions on the morphological distributions of tactility, among which four are manipulation tasks:

1. **SoftHardPickPlace.** In this task, we divide our objects into two categories according to hardness. We choose two soft objects and two hard ones as seen objects used to collect human demonstrations, with another two soft objects and two hard ones as unseen objects. We prepare two baskets, one on the left and the other on the right. The robot (arm and hand) picks the object up with only tactility and proprioception, and places the object into a basket. This task is considered success only if the object is placed into the right basket (soft in the left, hard in the right). A demonstration example is shown in figure 11.

2. **ObjectLocating.** Such a task requires the robot to find the correct location of the object after a pre-grasping contact. For each test time, the end-effector moves to a fixed location and touches the object on the table. The object is initially placed within a rectangular range of 10 cm × 20 cm, so that
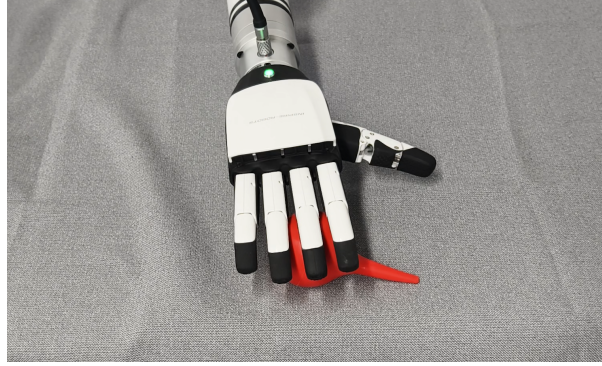
20

Figure 12: **An example of the task `ObjectLocating`.** The robot touches the object first, then adjusting the movement of the end-effector, so as to grasp the object successfully.

Figure 13: **An example of the task `CompliantControl`.** A person grabs the dexterous hand by applying external forces to a certain direction. The policy should infer the correct direction and move the end-effector according to tactility.

for each test, objects at different locations will result in different regions of tactile activation on the dexterous hand. For instance, in some cases the tactile regions on the index finger are activated, while in other cases the object is touched by the hand on the pinky. Once touching, the policy makes decisions representing the movement of the end-effector according to tactile information in the inference phase. The aim is to move the end-effector towards the correct direction with a proper distance so that the object can be grasped successfully with a direct grasp after moving the end-effector. A demonstration example is shown in figure 12.

3. **CompliantControl.** In this task, the pose of the dexterous hand is fixed similar to a "fist". A person "drags" the hand towards a specific direction. Due to the clenching gesture of the "fist", dragging towards different directions will result in different regions of tactility on the hand to be activated. The policy infers the moving direction according to tactile information, which should be consistent with the intention of the person who applies external forces on the dexterous hand. A demonstration example is shown in figure 13.

4. **One-Shot Manipulation.** Such a task is used to test the capability of the policy when one-shot real robot demonstration is mixed with human data. Besides, we use both visual and tactile information, to verify the competence of tactility in terms of those cases where vision only is not enough. In this task, we choose three pairs of objects (six in total, as shown in figure 15). Within each pair, the two objects are similar in vision but have different physical properties. For instance, two identical bottles, one empty and another full of water, look similar but can result in different tactility due to their significant

Figure 14: **An example of the one-shot manipulation task.** The robot picks up the bottle from the table, and infers the category of the object based on both visual input and tactile information, and finally places the object to the right place accordingly.

Table 4: **Basic settings of each manipulation task.** We list the inputs and outputs of the policies in our manipulation tasks: `SoftHardPickPlace`, `ObjectLocating`, `CompliantControl`, and the last one-shot manipulation task.

Task	Input State	Output Action
SoftHardPickPlace	Tactility + Proprioception (3-DoF End-Effector Location)	3-DoF End-Effector Location (X, Y, Z)
ObjectLocating	Tactility	2-DoF End-Effector Location (X, Y)
CompliantControl	Tactility	2-DoF End-Effector Delta Movement $(\Delta X, \Delta Y)$
One-Shot Manipulation	RGB Image + Tactility + Proprioception (arm joints)	6-DoF arm joints

difference in weight. During test time, the robot picks up the object (*e.g.*, a bottle) at a fixed location, and decides the following actions according to both the visual observation (from a third-person camera) and the tactile sense on the hand. The robot is required to place the object to the correct pre-defined location (each bottle full of water corresponding to a unique place; all empty bottles corresponding to an empty basket). A demonstration example is shown in figure 14.

For our manipulation tasks, `SoftHardPickPlace`, `ObjectLocating`, `CompliantControl`, as well as the last one-shot manipulation task, the inputs (states) and outputs (actions) of our policies are listed in table 4. For each task, we choose a specific suitable state representation and the corresponding action control.

Besides, we conduct experiments on object classification based on tactility over objects of different shapes, geometries, and physical features, denoted as our `ObjectClassification` task. In this task, we choose ten categories for classification (nine objects with another category identified as "empty"), as listed in figure 16.

C.2 Training Policies for Tasks

Here we present the details of training policies for each task in the expertiments (section 4).

We list the implementation details of the four policies (networks and architectures) in table 5. For our one-shot manipulation task, we utilize a pre-trained ResNet-18 [14] as the visual backbone, and concatenate the visual embedding with the tactile embedding, which is then fed into the policy network.

D Latent Space Analysis

To demonstrate the robustness and interpretability of the unified representation learned by UniTacHand, we visualize the latent space properties using paired human-robot data from the unseen validation set. We organize our analysis into two key aspects: cross-modal reconstructability and the linearity of the latent manifold.

Figure 15: **Six objects used in our one-shot manipulation task.** These objects are divided into three pairs of bottles, including both full and empty ones.

Figure 16: **Ten categories (nine objects and another category identified as "empty") used in our `ObjectClassification` task.** These objects are unseen in the paired data used to train unified representations, ranging in various shapes, geometries, and physical features.

Figures 17 and 18 present these properties using a 3×3 grid layout for two different pairs of object interactions. The layout is defined as follows:

- **Left Column (Human Input):** Shows two distinct human tactile patterns (H_1, H_2) and their pixel-wise average $((H_1 + H_2)/2)$.

- **Middle Column (Robot Ground Truth):** Shows the corresponding paired robotic tactile maps (R_1, R_2) and their pixel-wise average $((R_1 + R_2)/2)$.

- **Right Column (Model Reconstruction):** The first two rows show the robot UV maps reconstructed purely from the human inputs (i.e., $D_R(E_H(H_i))$). The third row shows the decoding result of the *averaged latent feature* (i.e., $D_R(\frac{z_1+z_2}{2})$).

D.1 Cross-Modal Reconstructability

A critical requirement for human-to-robot transfer is the ability to map human tactile sensations to their robotic counterparts accurately. As seen in the first two rows of figure 17 and figure 18, our model successfully performs this cross-modal translation. The reconstructed robot maps (Right Column, Rows 1-2) closely resemble the ground-truth robot maps (Middle Column, Rows 1-2) in terms of contact location and intensity distribution. This indicates that the Human Encoder (E_H) effectively extracts a canonical representation that the Robot Decoder (D_R) can interpret, successfully bridging the morphological gap.

Table 5: **Implementation details of the four policies.** We list the details including network architectures and hyperparameters of the policies in our manipulation tasks. D_{tac} is the dimension of tactility. For the following modalities: (1) raw data of tactile gloves, (2) raw data of Inspire Hand tactility, (3) our UV Map universal representation, (4) the embedding space of our unified representation learned from paired data, (5) the PatchMatch baseline that matches Inspire to gloves patch by patch, the dimensions are $D_{\text{tac}} = 137$, $D_{\text{tac}} = 1062$, $D_{\text{tac}} = 391$, $D_{\text{tac}} = 32$, $D_{\text{tac}} = 137$, respectively.

Implementation	SoftHardPickPlace	ObjectLocating	CompliantControl	One-Shot Manipulation
Network Architecture	MLP w/ Gumbel Routing	Multi-Head MLP	Tactile Encoder (MLP) + Policy (MLP)	Image Encoder (ResNet-18), Tactile Encoder (MLP), Policy (MLP)
Input Dim	$D_{\text{tac}} + 3$	D_{tac}	D_{tac}	$D_{\text{tac}} + 6$ (+ image)
Hidden Dim	512	[512, 256]	256	64
Output Dim	3	2	3	6
Action Chunk Size	16	64	1	16
Batch Size	20	32	32	120
Learning Rate	0.001	0.001	0.001	1×10^{-5}

D.2 Linearity and Additivity in the Latent Space

Beyond simple reconstruction, a well-structured latent space should exhibit linearity, where algebraic operations in the latent space correspond to meaningful morphological transformations in the output space. We investigate this by averaging the latent vectors of two distinct grasps: $z_{avg} = \frac{1}{2}(E_H(H_1) + E_H(H_2))$. We then decode this averaged vector to produce $\hat{R}_{avg} = D_R(z_{avg})$.

As illustrated in the third row of the figures, the decoded result from the latent average (Bottom Right) is morphologically consistent with the pixel-wise average of the ground truth robot data (Bottom Middle). This "Additivity" property suggests that our contrastive learning objective has shaped the latent manifold to be continuous and linear. Such a property is vital for generalization, as it implies the policy can smoothly interpolate between learned tactile primitives rather than overfitting to discrete training samples.
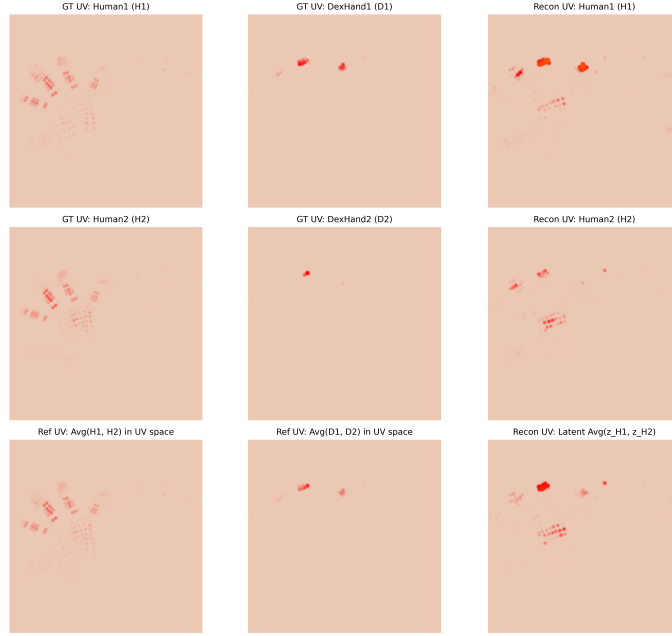
Figure 17: **Visualization of latent space properties (sample A).** This 3×3 grid demonstrates both reconstruction and linearity. **(Right Col, Top & Middle)**: The model accurately reconstructs robot tactile maps from human input, matching the ground truth in the Middle Column. **(Right Col, Bottom)**: The decoded result of the averaged latent vector matches the physical average of the robot maps (Middle Col, Bottom), confirming the linearity and additivity of the latent space.
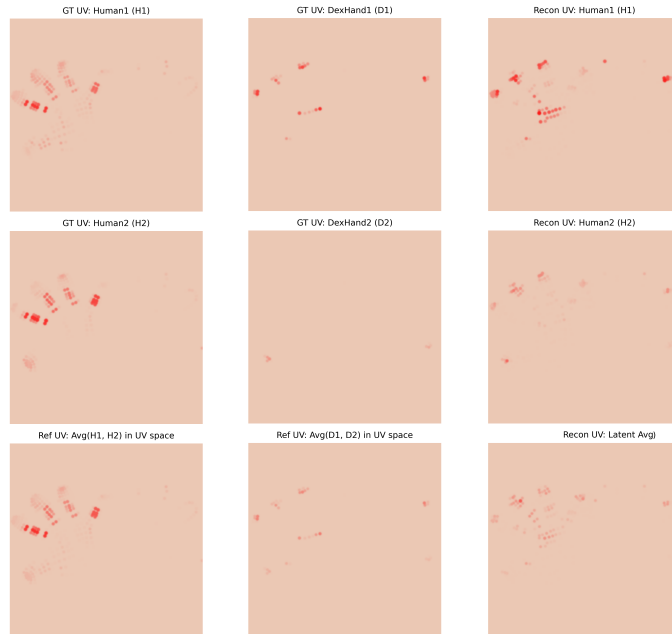

Figure 18: **Visualization of latent space properties (sample B).** Another example from the validation set. The consistent similarity between the decoded latent average (Bottom Right) and the ground truth average (Bottom Middle) reinforces that the learned representation captures the underlying continuous structure of tactile manipulation skills.