exUMI: Extensible Robot Teaching System with Action-aware Task-agnostic Tactile Representation

Yue Xu[1], **Litao Wei**[1], **Pengyu An**[1], **Qingyu Zhang**[1], **Yong-Lu Li**[1,2*]
[1]Shanghai Jiao Tong University, [2]Shanghai Innovation Institute
{silicxuyue,oscar0731,anpengyu,leozhangchina,yonglu_li}@sjtu.edu.cn

Abstract: Tactile-aware robot learning faces critical challenges in data collection and representation due to data scarcity and sparsity, and the absence of force feedback in existing systems. To address these limitations, we introduce a tactile robot learning system with both hardware and algorithm innovations. We present **exUMI**, an extensible data collection device that enhances the vanilla UMI with robust proprioception (via AR MoCap and rotary encoder), modular visuo-tactile sensing, and automated calibration, achieving 100% data usability. Building on an efficient collection of over **1 M** tactile frames, we propose Tactile Prediction Pretraining (**TPP**), a representation learning framework through action-aware temporal tactile prediction, capturing contact dynamics and mitigating tactile sparsity. Real-world experiments show that TPP outperforms traditional tactile imitation learning. Our work bridges the gap between human tactile intuition and robot learning through co-designed hardware and algorithms, offering open-source resources to advance contact-rich manipulation research.
Project page: https://silicx.github.io/exUMI.

Keywords: Tactile Sensing, Robot Data Collection System, Imitation Learning

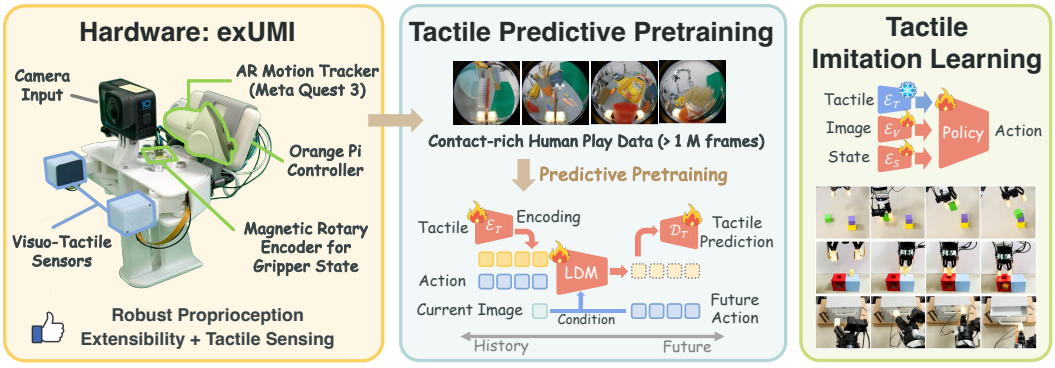

Figure 1: We present a co-design of hardware and algorithm for tactile-aware robot learning. We present **exUMI**, an extensible upgrade for the UMI [1] system (left). With the hardware, we learn tactile representation by temporal tactile prediction (middle), encoding the tactile dynamics conditioned on robot action. We evaluate our representation on multiple real-world robotic tasks (right).

1 Introduction

Collecting robot manipulation data is crucial for developing autonomous robots in real-world environments. While teleoperation techniques [2, 3, 4] yield accurate robot learning data, they are labor-intensive, inefficient, and expensive. Learning from human demonstration [5, 6, 7, 8, 9, 10, 11] is

*Corresponding author.

9th Conference on Robot Learning (CoRL 2025), Seoul, Korea.

cheap and massive, but inaccurate for robot embodiment. Between these two extremes, portable hand-held devices offer a promising middle ground. Recently, UMI [1] was proposed to alleviate the deployment gap, which is a portable physical twin of the robot gripper for in-the-wild demonstration collection. It shows remarkable capabilities in collecting kinesthetic teaching data, enabling efficient transfer of human demonstrations to robotic systems.

However, when it comes to tactile-aware robot learning, traditional data collection systems face challenges. First, human demonstrators rely on tactile feedback to adjust manipulation strategies, making it hard or impossible for the teleoperation systems to collect demonstrations for force-sensitive tasks. Second, tactile signals in robot learning are severely sparse, with valid contacts occupying less than 10% of manipulation trajectories [12]. This undermines conventional tactile learning approaches: (1) Direct imitation learning [13, 14] suffers from data scarcity. (2) Self-supervised pretraining [15, 16, 17, 18] learns with the potentially incorrect inductive bias (*e.g* , translation invariance). (3) Visual-tactile alignment paradigm [19, 20, 21] overlooks that vision and tactile modalities have a one-to-many relation considering the contact force. Therefore, the resulting representations often fail to generalize beyond narrow task-specific scenarios.

To address these challenges, we present a solution of tactile representation learning with the co-design of both novel hardware and an algorithm. We propose **exUMI** (Sec. 3), an **ex**tensible upgrade to the **UMI** robot teaching system [1]. exUMI is a portable hand-held data collection device with visuo-tactile sensors and a motion capture system for teaching robot trajectory. Our system has three key innovations: (1) A robust proprioception subsystem of AR-based motion capture (Meta Quest 3) and magnetic rotary encoder (AS5600), achieving nearly 100% data usability by replacing the vulnerable SLAM and ArUco systems. (2) A central controller enabling modality extensibility, coupled with temporal sensor alignment protocols during human demonstrations with $<$50 ms error. (3) Visuo-tactile integration of an upgraded design of 9DTact [22] for stable quality control and better durability. With exUMI, users could efficiently collect robot learning and tactile sensing data with the least effort. For a simple pick-and-place task, a user could collect **100** demonstrations in **20** minutes to achieve **100%** data usability and over **70%** task success rate by behavior cloning.

Building on the hardware basis, we propose an *action-aware* but *task-agnostic* tactile representation learning framework (Sec. 4). To exploit exUMI's unique data properties and address the previously mentioned potential issues, we propose **T**actile **P**redictive **P**retraining (**TPP**), learning tactile representation by the proxy task of **temporal tactile prediction**. We pretrain a tactile diffusion model to predict future tactile frames, conditioned on the action sequence and the current camera image. The representation is learned considering the physical action on the sensor, mirroring human haptic perception, where contact dynamics could be inferred from future movement. To enable the action-aware pretraining, we use exUMI to efficiently collect large-scale *human play* data by randomly interacting with objects, producing a contact-rich tactile-action aligned dataset of $>$**1 M frames** with over **10 times** efficiency than teleoperation. The pretrained representation model could be embedded in an imitation learning policy, and our tactile embeddings empirically achieve significant success gain in force-sensitive tasks (*e.g* , "pull drawer, peg in hole") versus vision-only baselines.

This work establishes a new paradigm for tactile-aware robot learning to overcome fundamental bottlenecks in tactile learning. Our primary contributions include: (1) **exUMI**, a tactile robot data system that enhances UMI with 100% reliable proprioception and tactile sensing; (2) **Tactile Prediction Pretraining (TPP)**, an action-aware and task-agnostic tactile representation learning method that exploits contact dynamics through forward tactile prediction; (3) A large-scale tactile-action aligned robot dataset with over 1 M frames; (4) Empirical validation showing over 20% performance gains over tactile learning baselines.

2 Related Work

Robot Data Collection Systems are essential for training generalizable manipulation policies. Traditional approaches include teleoperation [23, 4, 24, 25], which offers high-quality data but is costly; and human data based methods [5, 6, 7, 26, 27, 8, 28, 29, 9, 10, 30, 11], which lack precision.

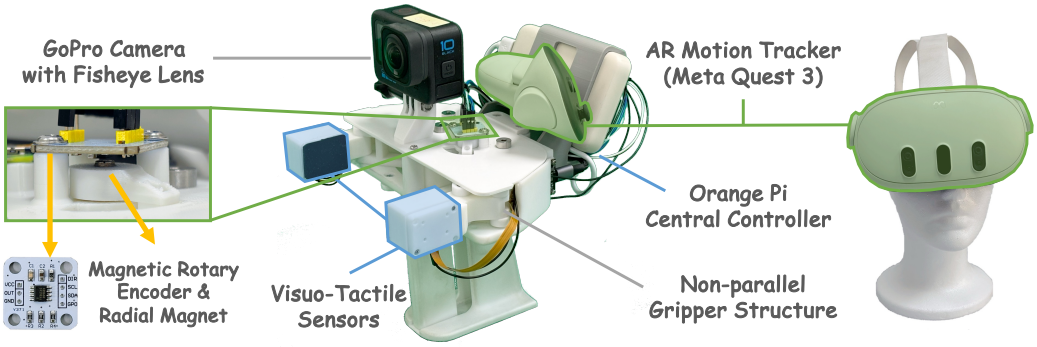

Figure 2: **exUMI** hardware system. We extend the UMI framework by disentangling proprioception into an AR motion capture system and a rotary encoder for precise gripper width. A central controller with automatic latency calibration enables additional sensor integration with maximal mobility. We attach two visuo-tactile sensors to the gripper fingertips.

Portable systems like UMI [1] and AirExo [31] employ a handheld or exoskeleton as a physical twin of a robot, enabling efficient in-the-wild demonstrations. Recent variants address existing limitations: Fast-UMI [32] enhances the SLAM system of UMI with RealSense T265, while ForceMimic [33] integrates a force-torque sensor for contact-rich tasks. These advancements improve scalability, adaptability, and multimodal data acquisition for robotic learning.

Tactile Representation Learning aims to extract meaningful tactile features for robots to perceive geometry properties and interaction dynamics. Current trending involves: (1) *Directly imitation training*: minimally process the tactile data [13] or even directly use it without any processing [14] for training via reinforcement learning methods. Further learning process involves modality fusion models such as transformer [19, 20, 21]. (2) *Intermediate Representation:* transform raw data into a manual representation space that captures task-relevant semantics, such as converting into point clouds [34, 35], or reconstruct with NeRF [36]. (3) *Self-Supervised Learning* (SSL): leverages the inherent structure and temporal-spatial correlations within raw tactile data through proxy tasks. Rodriguez et al. [15] extends contrastive learning to paired tactile data. Guzey et al. [16], Yu et al. [17] extend traditional SSL methods like BYOL to tactile tasks and train on play data or task-specific data. Wu et al. [18] utilizes the masked learning method. Feng et al. [37] combine the pixel level SSL and contrastive learning to learn a cross-sensor tactile representation. Zhao et al. [38] exploit multimodal and multitask joint representation learning for a semantically meaningful tactile model.

Real-World Tactile Datasets have been developed to advance tactile perception and manipulation. Early progress in tactile sensing was driven by data collected using GelSight and, more recently, DIGIT sensors. Notable examples include Calandra *et. al.* [39, 40], SSVTP [41], datasets collected through a self-supervised automated process, respectively utilizing GelSight and DIGIT; VisGel [42], a synchronized vision-touch dataset comprising diverse everyday objects; X-Capture [43], adding depth information and acoustic information. Touch and Go [44], a single model approach that utilizes a portable device, enabling human data collection and introducing a large diversity in scenes. In comparison, our dataset addresses the challenges of efficient collection and proprioception alignment, resulting in a significantly larger data volume than any existing dataset.

3 Hardware System

We present **exUMI**, an enhanced hardware design upon UMI [1], guided by three key principles:

- **Precise robot proprioception**: Precise tracking of end-effector 6D pose and gripper width. The vanilla UMI system relies on visual SLAM and ArUco tracking, which is vulnerable (Fig. 3).
- **Extensibility**: Seamless integration of additional sensors through centralized control.
- **Portability**: Ensuring in-the-wild data collection without fixed infrastructure (*e.g* , base station).

Figure 3: Hard scenarios for SLAM and marker detection (clean background, occlusion).

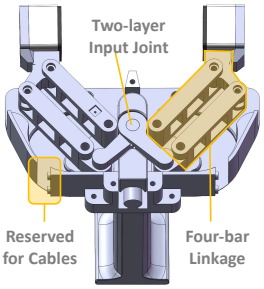

Figure 4: Non-parallel gripper mechanism.

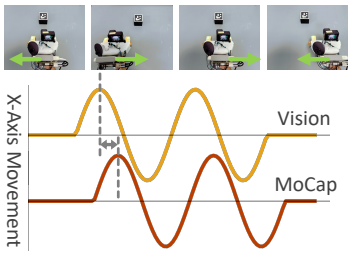

Figure 5: The latency calibration process of AR MoCap and RGB camera.

Therefore, we develop the hardware system with the following key components. Due to the page limit, please refer to Appendix A for full details.

3.1 Hardware Design

Magnetic Rotary Encoder. The gripper state estimation in the original UMI design relies on ArUco markers, which suffer from occlusion and severe fisheye distortion, contributing to approximately 10% of data preprocessing failures and notable errors. To achieve accurate and robust gripper width tracking, we propose a low-cost AS5600 magnetic rotary encoder solution. As shown in Fig. 2, we modify the top plate to attach a radial magnet and the rotary sensor above a joint.

AR Motion Capture System. To overcome the limitations of vision-based tracking (SLAM) in occluded or complex scenarios, use a Meta Quest 3 headset for end-effector 6D pose capture following Chen et al. [45], which is accurate and robust to occlusion. We integrate the left VR controller through a custom-designed mount attached to the UMI body, which also provides space for the power supply and an Orange Pi controller, serving as a universal sensor hub, synchronously capturing data from the AR headset, rotary encoder, and any additional sensors, such as tactile sensors. Our MoCap system achieves less than 10 mm error on average compared to FastUMI [32].

Fingertip Visuo-tactile Sensors. We embed visuo-tactile sensors on the fingertips for modality extension. We propose a low-cost visuo-tactile sensor based on 9DTact [22]. We redesigned the sensor model with a contact protection structure, which could hold the silicon gel and secure it from large tangent forces. The cable and power connection are also optimized for durability. We also use a customized mold to ensure the same and stable thickness of the silicon gel for consistent tactile sensing. The upgraded sensor design achieves significantly enhanced durability and stability.

Visual Input. We employ the same GoPro setting to Wu et al. [32] for a wider and clearer view.

Non-Parallel Gripper Mechanical Design. We design an additional mechanical system (Fig. 4) for the users of non-parallel grippers such as Flexiv Grav and Robotiq 2F.

Cost and Accessibility. Our system is low-cost and DIY-friendly with a default configuration starting at **$ 698**, making it suitable for research/education. All CAD files will be released.

3.2 Data Processing

Our careful design enables robust in-the-wild data collection with minimal calibration overhead.

System Calibration. Our system requires two **one-time** calibrations: (1) *AR Controller Calibration*: the user aligns the exUMI with the base coordinates in the AR space and records the controller's transform, which is then used to correct the pose tracking. (2) *Gripper State Calibration*: incrementally positioned the gripper at 1 cm intervals and recorded AS5600 reading, which is then interpolated and used for mapping from AS5600 readings to absolute gripper width.

Latency Calibration. We designed a calibration protocol to synchronize AR motion capture (and the tactile signals) with the visual inputs (Fig. 5). At the beginning of data collection, the user horizontally sweeps above an ArUco marker. We extract the x-axis movement of the AR MoCap and the marker trajectories, then use an MSE minimization algorithm to find the latency offset between the two curves. This approach ensures tight synchronization between the two systems.

Overall, exUMI could efficiently collect robot demonstrations with accurate 6D pose trajectory and gripper width, aligned RGB image input, and additional tactile signals. Our data collection pipeline is significantly simplified and more robust, leading to a nearly 100% data processing success rate compared to less than 60% of vanilla UMI. Our system could significantly enhance data efficiency and also enable multi-modality data collection at an affordable price.

4 Methodology

4.1 Taxonomy of Tactile Representation Learning

Tactile representation learning converts high-dimensional raw touch signals into compact features, which is fundamental to alleviating challenges like sensor heterogeneity, data dimension, data scarcity, and the need for real-world generalization. Regularly, its target is to learn a tactile encoder $\mathcal{E}_T$ for further multimodal policy learning: $\pi(\mathbf{a}_t|\mathcal{E}_S(\mathbf{s}_t), \mathcal{E}_T(\mathbf{T}_t), \mathcal{E}_V(\mathbf{V}_t))$, where $\mathbf{a}$ is the action, $\mathbf{T}$, $\mathbf{V}$, $\mathbf{s}$ are tactile, visual and robot state inputs, and $\mathcal{E}_S$, $\mathcal{E}_T$, $\mathcal{E}_V$ are the encoders. Currently, we classify the learning methods of $\mathcal{E}_T$ to three paradigms, which will be detailed in Appendix B.

(a) Direct Multimodal Imitation Learning [13, 14, 19, 20, 21]: end-to-end learning $\mathcal{E}_T$ during the training of π, which suffers from tactile data scarcity since both the data size and the proportion of valid tactile contacts are limited. **(b) Spatial Self-Supervised Learning**: methods like contrastive learning [15, 16, 17] and masked learning [18] learn tactile embeddings $\mathcal{E}_T(\mathbf{T}_t)$ through proxy objectives. But these method usually imposes incorrect inductive biases borrowed from vision *e.g* , geometrical self-consistency and translation invariance, which may not exist in tactile sensing. **(c) Visual-Tactile Alignment** [42, 43]: learns joint embeddings by maximizing similarity $s\left(\mathcal{E}_T(\mathbf{T}_t), \mathcal{E}_V(\mathbf{V}_t)\right)$. It assumes a coarse *one-to-one visuo-tactile mapping*, regardless of the actual *one-to-many relation* when different contact forces are applied.

Therefore, to overcome data scarcity of tactile sensing and for task transferability, representation pretraining (such as (b), (c)) is critical. However, current pretraining approaches face limitations that stem from a shared oversight: treating tactile signals as static observations rather than *action-aware dynamic processes*. Human tactile understanding intrinsically combines contact mechanics with motion intent (*e.g* , "*if I push harder, drag the object left, the slip risk decreases, and the tactile signal will be more significant*"). Our framework bridges this gap by reformulating tactile learning as an action-conditioned temporal prediction problem, explicitly modeling the forward tactile dynamics that underpin real-world contact interactions.

4.2 Action-aware Tactile Data Collection

For the tactile pretraining process, we efficiently collect tactile-action aligned human play data leveraging the portability of the exUMI system. The collectors randomly manipulate diverse objects across **10** real-world environments, interacting with **300+** objects spanning from rigid tools to deformable fabrics and granular materials. Finally, we collect a total of **1 M** frames of aligned images-tactile-action data. Our data has rich contacts with over 60% active tactile frames, compared to less than 10% of regular data collection [12]. The contact richness further enhances our collection efficiency, and the **480 K** tactile frames are collected from just 5 hours of human interaction, which would take 10× the time for a teleoperation system. Although having different purposes and granularity, the dataset is significantly larger than the previous tactile datasets (*e.g* , TVL [12] has 43.7 K frames), which is sufficient for our tactile learning.

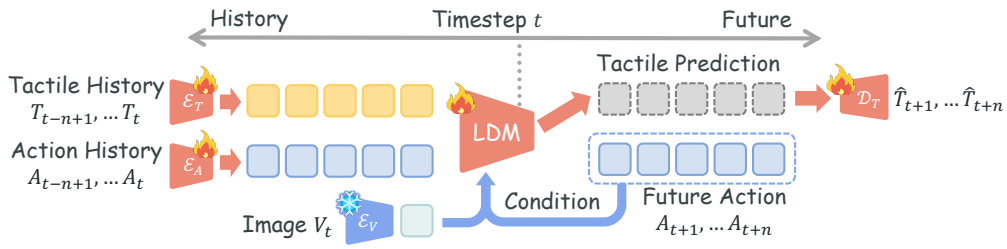

Figure 6: The proposed tactile representation learning pipeline. The representation model $\mathcal{E}_T$ is learned during the temporal tactile prediction task. The history tactile and action features are fused and mapped as the prediction of future tactile features with a latent diffusion model (LDM). The future action and the current image are encoded as the condition of the denoising process.

4.3 Tactile Predictive Pretraining

Given the exUMI hardware, we present an action-aware, and task-agnostic tactile representation learning framework that addresses the challenges of contact dynamics modeling in manipulation. We propose **T**actile **P**redictive **P**retraining (**TPP**), formulating tactile representation learning as a conditional future prediction: $p_\theta\left(\mathbf{T}_{t+1:t+n} | \mathcal{E}_T(\mathbf{T}_{t-n+1:t}), \mathcal{E}_V(\mathbf{V}_t), \mathcal{E}_A(\mathbf{A}_{t-n+1:t+n})\right)$, The tactile encoder $\mathcal{E}_T$ would learn an informative representation involving tactile dynamics in the predictive pretraining. The predictive model is learned from our 1 M human play data. We adopt the diffusion model and masked autoencoder structure following [46] (Fig. 6), with the following components:

Multimodal Encoding. We employ the VAE model as the encoder and decoder of the tactile modality $(\mathcal{E}_T, \mathcal{D}_T)$, and each tactile image is patchified and converted into a sequence of embeddings. Different from UVA [46], the VAE model is learnable for tactile representation learning.

Tactile Prediction. Following Li et al. [46], we apply random masking to the history tactile patch embeddings and action features and fuse the two modalities with a transformer. The n history latents are forwarded to a latent diffusion model to predict the latents of future tactile signals, where the embeddings of future action $A_{t+1:t+n}$ and current RGB image V_t serve as the condition. The tactile image is reconstructed by the VAE decoder. The predictive model is constrained by hybrid losses: (1) $\mathcal{L}_{diff}$: the regular diffusion loss between the predicted and actual noise perturbations. (2) $\mathcal{L}_{recon}$: the reconstruction MSE loss between reconstructed and original tactile images.

Tactile History	Action Input	RGB Image	MSE Error
✗	✓	✓	0.0298
✓	✗	✗	0.0132
✓	✗	✓	0.0125
✓	✓	✗	0.0117
✓	✓	✓	**0.0099**

Table 1: Tactile predictive pretraining with different input settings.

Policy Learning. After learning the predictive representation, we *freeze* the tactile encoder $\mathcal{E}_T$ for all downstream policies. We learn the multimodal policy with common imitation learning. The rich tactile dynamics knowledge has been encoded in the tactile model; hence, our approach could generate a more robust representation and alleviate the issues of data scarcity and low diversity, avoiding the overfitting in low-shot learning scenarios for tactile-aware tasks. Note that although the pretraining process requires dense computational resources, the pretrained tactile model could be seamlessly adopted in all the downstream policy learning tasks without further finetuning.

5 Experiment

5.1 Tactile Predictive Pretraining

We first pretrain the TPP model on our collected large-scale dataset with a total of over 1M frames.

Implementation Details. exUMI collects two tactile images on the two sides. We convert the images to a calibrated grayscale image following Lin et al. [22], and extract the convex and concave pixel maps and stack them as a 3-channel image for a richer representation of tactile contacts. We train the TPP model on 4 NVIDIA H100 GPUs for 120 hours. Refer to Appendix C for details.

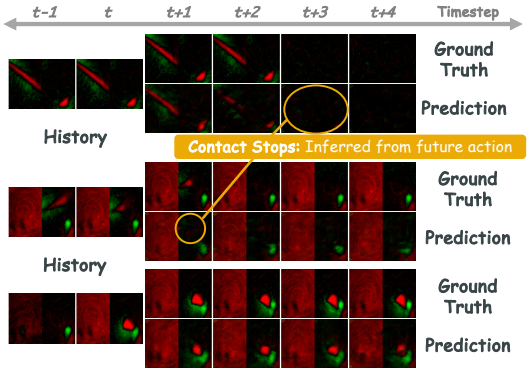

Figure 7: Examples of temporal tactile prediction of TPP on the validation set.

Task	Number of Demos	Data Collection Duration †	Success Rate
Pick Cube	204	42 min	**85%**
Pick Carrot	135	31 min	**80%**
Pick Broccoli	170	47 min	**60%**
Stack Cube	201	26 min	**60%**
Insert Pen	139	60 min	**65%**
Put Ball	181	66 min	**70%**
Open Bottle	270	79 min	**20%**
Pull Drawer	202	70 min	**40%**
Peg in Hole	163	56 min	**50%**

Table 2: Data collection efficiency and policy performance on non-tactile tasks. †: total collection time of **all** demonstrations, including the environment resetting.

Results. We compare the prediction MSE error on the val set in Tab. 1, showing that multimodal conditioning of visual and action sequences could reduce the tactile prediction error. Among all the settings, our action-aware prediction achieved the best performance, indicating that the policy implicitly learns the forward tactile dynamics with full consideration of the action sequence.

In Fig. 7. We visualize the tactile signals on the validation set (unseen data) following Lin et al. [22], where red and green represent the concave and convex areas. Our model manifests to learn clues from the action sequence. For example, in the first case, the model infers from the future action that the tactile contact will cease after two frames, which is *impossible* for models without the reasoning of action-informed tactile dynamics. Thus, our action-aware predictive pretraining enables robust tactile representation for future tactile-aware policy learning.

5.2 Experiment Settings for Imitation Learning

Environment Settings. We evaluate our learning system on a Flexiv Rizon 4 robot arm with a Flexiv Grav adaptive gripper, and use a GoPro camera as the only visual input. We adopt diffusion policy [47] with a ViT image backbone model, and direct feature concatenation for multimodal inputs. We evaluate our policy for 20 trials.

Task Settings. Our evaluation covers regular manipulation tasks: (1) *Pick cube / carrot / broccoli*: pick up and place it into a container; (2) *Insert pen*: move a pen to another cup; (3) *Stack cubes*: stack a small cube on top of another.

For tactile-aware policy learning, we evaluate on more complex tasks: (1) *Put Ball*: pick up a soft ball and place it in a cup. (2) *Open Bottle*: rotate the bottle cap until it is fully unscrewed. (3) *Pull Drawer*: pull out one drawer, which is either empty ("Empty") or contains a random amount of stones ("Random"), requiring tactile clues to determine the pulling direction. (4) *Peg in hole*: insert a block into a slot, requiring a precision and force-aware adjustment. We split the task into "Grasp" and "Insert" stages. Please refer to Appendix C for more details.

Demonstration Collection. We collect 100 to 200 demonstrations with exUMI for robot teaching, and the total data collection time is shown in Tab. 2. Note that the time duration includes the task. With its efficiency and tactile feedback, an expert could complete the data collection within half an hour. And our system has a nearly 100% effective data ratio. In comparison, users should spend **50%** or more data collection time with the regular UMI system for the same amount of valid data.

5.3 Real World Evaluation Result

Non-tactile Imitation Learning. We first train a vision-only diffusion policy to evaluate the data collection quality of exUMI (Tab. 2). The policy achieves decent performance across tasks, demonstrating both sufficient spatial demonstration quality and its capability for visual imitation learning. Our policy achieves an over 80% success rate on simple pick-and-place tasks, indicating sufficient

Input Representation	Put Ball	Open Bottle	Pull Drawer		Peg in Hole	
			Empty	Random	Grasp	Insert
Vision Only	70%	20%	100%	40%	100%	50%
Vision & Tactile	70%	50%	100%	50%	100%	60%
Vision & Tactile w/ TPP (Ours)	**85%**	**60%**	100%	**95%**	100%	**80%**

Table 3: Real world evaluation of tactile-aware policies on complex tasks.

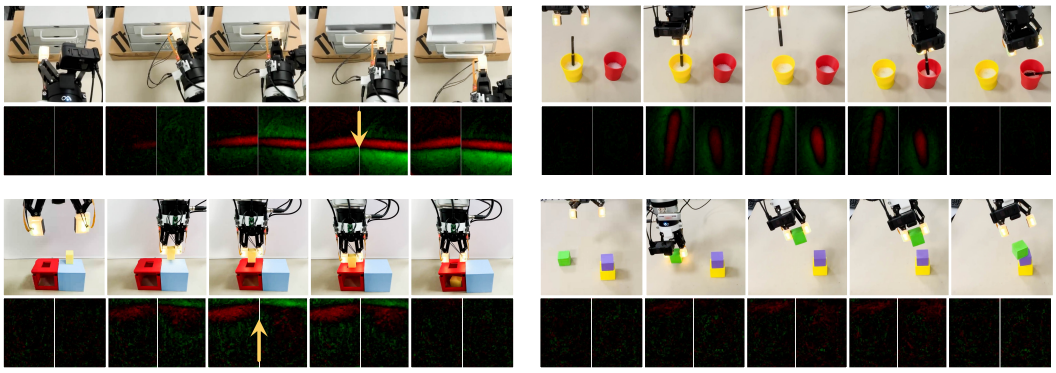

Figure 8: Real world rollout of tactile-aware policy. Yellow arrows indicate the tangent force, which points from red area (concave due to pressure) to green area (convex bump of silicone gel).

spatial demonstration quality. The success rate of "pick broccoli" task slightly drops to 60% due to the slippery surface and irregular geometry. Tasks involving precision manipulation ("stack" and "insert") are more challenging, but we still reach over 60% due to the accurate robot proprioception of exUMI. Complex tasks requiring reasoning with tactile feedback show performance degradation. "Pull drawer" achieves merely a 40% success rate, and the majority of failures occur when the policy applies excessive force in incorrect directions. For "peg in hole", the policy usually fails to align the peg merely on visual input. These tasks require further tactile robot learning.

Tactile-aware Imitation Learning. We augment our policy model with tactile sensing and evaluate across contact-sensitive tasks (Tab. 3 and visualized in Fig. 8). While both vision-only and tactile-aware policies achieve a high success rate on simple settings, tactile integration yields significant improvements, particularly in harder stages (*e.g* drawer pulling). For the "put ball" and "rotate bottle" tasks, TPP brings >15% performance gain over vanilla tactile policy. And though tactile policy is only comparable to vision policy on "put ball" task, we empirically observe that tactile-aware policy slightly adjusts the grasp pose until accurately holding the center of the ball. For the "pull drawer" and "peg in hole" tasks, all policies with or without tactile sensing achieve a high success rate of 100% at non-tactile stages (pulling the empty drawer or simply grasping the object). But for the force-sensitive stages, tactile-aware policies continuously bring improvement, and our TPP boosts the performance to over 100% and 80%, respectively. For these tasks, tactile clues play an important role in the trajectory planning. As shown in Fig. 8, for the "pull drawer" task, the tactile signals inform the grasp pose of the handle (red area), which is critical for the selection of pulling direction. This indicates that the insertion success depends critically on post-grasp tactile feedback.

These results validate that effective tactile integration requires both careful tactile feature engineering (via pretraining). The performance gaps between our method and baselines highlight the importance of learning the temporal dynamics of tactile for real-world manipulation.

6 Conclusion

In this work, we propose a co-design of the exUMI hardware with its reliable proprioception and scalable tactile sensing, and the TPP framework, which learns tactile features by predictive proxy task. Our system achieves significant performance gain on complex tactile-aware tasks, highlighting the importance of grounding human-style contact dynamics reasoning in physical interaction.

7 Limitations

While our system provides a user-friendly interface for tactile-aware robot data collection, several critical challenges remain to be addressed in future research.

Hardware Limitations. (1) Although our AR headset-based motion capture system demonstrates robustness, user feedback highlights two ergonomic concerns of thermal discomfort and neck strain. While we considered alternative tracking solutions (*e.g* , dedicated motion capture trackers like HTC Vive Tracker), these trackers usually rely on external base stations for more accurate tracking, which conflicts with our design goal of maintaining portability for in-situ AR-assisted data acquisition. Future work could explore ergonomic upgrades like neck supports. (2) The durability and consistency have always been critical concerns of tactile sensors. We adopt the 9DTact sensor for a low-cost tactile solution. While we have notably improved tactile sensor consistency to 9DTact, further enhancements remain possible.

Algorithm Limitations. Our predictive framework for learning interaction representations faces inherent constraints. The interaction and movement information is limited due to low action dimension and limited camera angle, resulting in imperfect tactile prediction performance. We plan to address these by integrating force-torque measurements and multi-view vision inputs in the future.

References

[1] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024.

[2] Y. Zhu, A. Joshi, P. Stone, and Y. Zhu. Viola: Imitation learning for vision-based manipulation with object proposal priors. In *Conference on Robot Learning*, pages 1199–1210. PMLR, 2023.

[3] A. Mandlekar, Y. Zhu, A. Garg, J. Booher, M. Spero, A. Tung, J. Gao, J. Emmons, A. Gupta, E. Orbay, et al. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In *Conference on Robot Learning*, pages 879–893. PMLR, 2018.

[4] S. Luo, Q. Peng, J. Lv, K. Hong, K. R. Driggs-Campbell, C. Lu, and Y.-L. Li. Human-agent joint learning for efficient robot manipulation skill acquisition. *arXiv preprint arXiv:2407.00299*, 2024.

[5] L. Shao, T. Migimatsu, Q. Zhang, K. Yang, and J. Bohg. Concept2robot: Learning manipulation concepts from instructions and human demonstrations. *The International Journal of Robotics Research*, 40(12-14):1419–1434, 2021.

[6] A. S. Chen, S. Nair, and C. Finn. Learning generalizable robotic reward functions from" in-the-wild" human videos. *arXiv preprint arXiv:2103.16817*, 2021.

[7] S. Bahl, A. Gupta, and D. Pathak. Human-to-robot imitation in the wild. *arXiv preprint arXiv:2207.09450*, 2022.

[8] A. Simeonov, Y. Du, A. Tagliasacchi, J. B. Tenenbaum, A. Rodriguez, P. Agrawal, and V. Sitzmann. Neural descriptor fields: Se (3)-equivariant object representations for manipulation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6394–6400. IEEE, 2022.

[9] K. Schmeckpeper, A. Xie, O. Rybkin, S. Tian, K. Daniilidis, S. Levine, and C. Finn. Learning predictive models from observation and interaction. In *European Conference on Computer Vision*, pages 708–725. Springer, 2020.

[10] Y. Qin, Y.-H. Wu, S. Liu, H. Jiang, R. Yang, Y. Fu, and X. Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision*, pages 570–587. Springer, 2022.

[11] T. Xiao, I. Radosavovic, T. Darrell, and J. Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022.

[12] L. Fu, G. Datta, H. Huang, W. C.-H. Panitch, J. Drake, J. Ortiz, M. Mukadam, M. Lambeta, R. Calandra, and K. Goldberg. A touch, vision, and language dataset for multimodal alignment. *arXiv preprint arXiv:2402.13232*, 2024.

[13] E. Su, C. Jia, Y. Qin, W. Zhou, A. Macaluso, B. Huang, and X. Wang. Sim2real manipulation on unknown objects with tactile-based reinforcement learning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9234–9241, 2024. doi: 10.1109/ICRA57147.2024.10611113.

[14] K.-W. Lee, Y. Qin, X. Wang, and S.-C. Lim. Dextouch: Learning to seek and manipulate objects with tactile dexterity. *IEEE Robotics and Automation Letters*, 9(12):10772–10779, Dec. 2024. ISSN 2377-3774. doi:10.1109/lra.2024.3478571. URL http://dx.doi.org/10.1109/LRA.2024.3478571.

[15] S. Rodriguez, Y. Dou, W. van den Bogert, M. Oller, K. So, A. Owens, and N. Fazeli. Contrastive touch-to-touch pretraining, 2024. URL https://arxiv.org/abs/2410.11834.

[16] I. Guzey, B. Evans, S. Chintala, and L. Pinto. Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play, 2023. URL https://arxiv.org/abs/2303.12076.

[17] K. Yu, Y. Han, Q. Wang, V. Saxena, D. Xu, and Y. Zhao. Mimictouch: Leveraging multi-modal human tactile demonstrations for contact-rich manipulation, 2025. URL https://arxiv.org/abs/2310.16917.

[18] T. Wu, J. Li, J. Zhang, M. Wu, and H. Dong. Canonical representation and force-based pretraining of 3d tactile for dexterous visuo-tactile policy learning, 2025. URL https://arxiv.org/abs/2409.17549.

[19] B. Romero, H.-S. Fang, P. Agrawal, and E. Adelson. Eyesight hand: Design of a fully-actuated dexterous robot hand with integrated vision-based tactile sensors and compliant actuation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1853–1860, 2024. doi:10.1109/IROS58592.2024.10802778.

[20] H. Lin, R. Corcodel, and D. Zhao. Generalize by touching: Tactile ensemble skill transfer for robotic furniture assembly. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9227–9233, 2024. doi:10.1109/ICRA57147.2024.10610567.

[21] V. Pattabiraman, Y. Cao, S. Haldar, L. Pinto, and R. Bhirangi. Learning precise, contact-rich manipulation through uncalibrated tactile skins, 2024. URL https://arxiv.org/abs/2410.17246.

[22] C. Lin, H. Zhang, J. Xu, L. Wu, and H. Xu. 9dtact: A compact vision-based tactile sensor for accurate 3d shape reconstruction and generalizable 6d force estimation. *IEEE Robotics and Automation Letters*, 2023.

[23] W. Fan, X. Guo, E. Feng, J. Lin, Y. Wang, J. Liang, M. Garrad, J. Rossiter, Z. Zhang, N. Lepora, et al. Digital twin-driven mixed reality framework for immersive teleoperation with haptic rendering. *IEEE Robotics and Automation Letters*, 2023.

[24] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.

[25] M. Seo, S. Han, K. Sim, S. H. Bang, C. Gonzalez, L. Sentis, and Y. Zhu. Deep imitation learning for humanoid loco-manipulation through human teleoperation. In *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*, pages 1–8. IEEE, 2023.

[26] Y. J. Ma, S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, and A. Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.

[27] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023.

[28] C. Pan, B. Okorn, H. Zhang, B. Eisner, and D. Held. Tax-pose: Task-specific cross-pose estimation for robot manipulation. In *Conference on Robot Learning*, pages 1783–1792. PMLR, 2023.

[29] W. Shen, G. Yang, A. Yu, J. Wong, L. P. Kaelbling, and P. Isola. Distilled feature fields enable few-shot language-guided manipulation. *arXiv preprint arXiv:2308.07931*, 2023.

[30] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.

[31] H. Fang, H.-S. Fang, Y. Wang, J. Ren, J. Chen, R. Zhang, W. Wang, and C. Lu. Airexo: Low-cost exoskeletons for learning whole-arm manipulation in the wild. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 15031–15038. IEEE, 2024.

[32] Z. Wu, T. Wang, C. Guan, Z. Jia, S. Liang, H. Song, D. Qu, D. Wang, Z. Wang, N. Cao, et al. Fast-umi: A scalable and hardware-independent universal manipulation interface. *arXiv preprint arXiv:2409.19499*, 2024.

[33] W. Liu, J. Wang, Y. Wang, W. Wang, and C. Lu. Forcemimic: Force-centric imitation learning with force-motion capture system for contact-rich manipulation. *arXiv preprint arXiv:2410.07554*, 2024.

[34] Y. Yuan, H. Che, Y. Qin, B. Huang, Z.-H. Yin, K.-W. Lee, Y. Wu, S.-C. Lim, and X. Wang. Robot synesthesia: In-hand manipulation with visuotactile sensing. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6558–6565, 2024. doi: 10.1109/ICRA57147.2024.10610532.

[35] B. Huang, Y. Wang, X. Yang, Y. Luo, and Y. Li. 3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing, 2025. URL https://arxiv.org/abs/2410.24091.

[36] S. Suresh, H. Qi, T. Wu, T. Fan, L. Pineda, M. Lambeta, J. Malik, M. Kalakrishnan, R. Calandra, M. Kaess, J. Ortiz, and M. Mukadam. Neuralfeels with neural fields: Visuotactile perception for in-hand manipulation. *Science Robotics*, 9(96):eadl0628, 2024. doi:10.1126/scirobotics.adl0628. URL https://www.science.org/doi/abs/10.1126/scirobotics.adl0628.

[37] R. Feng, J. Hu, W. Xia, T. Gao, A. Shen, Y. Sun, B. Fang, and D. Hu. Anytouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors. *arXiv preprint arXiv:2502.12191*, 2025.

[38] J. Zhao, Y. Ma, L. Wang, and E. H. Adelson. Transferable tactile transformers for representation learning across diverse sensors and tasks. *arXiv preprint arXiv:2406.13640*, 2024.

[39] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3(4):3300–3307, 2018.

[40] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, and S. Levine. The feeling of success: Does touch sensing help predict grasp outcomes? *arXiv preprint arXiv:1710.05512*, 2017.

[41] J. Kerr, H. Huang, A. Wilcox, R. Hoque, J. Ichnowski, R. Calandra, and K. Goldberg. Self-supervised visuo-tactile pretraining to locate and follow garment features. *arXiv preprint arXiv:2209.13042*, 2022.

[42] Y. Li, J.-Y. Zhu, R. Tedrake, and A. 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.

[43] S. Clarke, S. Wistreich, Y. Ze, and J. Wu. X-capture: An open-source portable device for multi-sensory learning. *arXiv preprint arXiv:2504.02318*, 2025.

[44] F. Yang, C. Ma, J. Zhang, J. Zhu, W. Yuan, and A. Owens. Touch and go: Learning from human-collected vision and touch, 2022. URL https://arxiv.org/abs/2211.12498.

[45] S. Chen, C. Wang, K. Nguyen, L. Fei-Fei, and C. K. Liu. Arcap: Collecting high-quality human demonstrations for robot learning with augmented reality feedback. *arXiv preprint arXiv:2410.08464*, 2024.

[46] S. Li, Y. Gao, D. Sadigh, and S. Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025.

[47] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.

[48] C. Higuera, A. Sharma, C. K. Bodduluri, T. Fan, P. Lancaster, M. Kalakrishnan, M. Kaess, B. Boots, M. Lambeta, T. Wu, et al. Sparsh: Self-supervised touch representations for vision-based tactile sensing. *arXiv preprint arXiv:2410.24090*, 2024.

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

[50] A. L. Burka. *Instrumentation, data, and algorithms for visually understanding haptic surface properties*. PhD thesis, University of Pennsylvania, 2018.

[51] R. Gao, Y. Dou, H. Li, T. Agarwal, J. Bohg, Y. Li, L. Fei-Fei, and J. Wu. The object-folder 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.

[52] S. Rodriguez, Y. Dou, M. Oller, A. Owens, and N. Fazeli. Touch2touch: Cross-modal tactile generation for object manipulation. *arXiv preprint arXiv:2409.08269*, 2024.

[53] K. Yu, Y. Han, Q. Wang, V. Saxena, D. Xu, and Y. Zhao. Mimictouch: Leveraging multi-modal human tactile demonstrations for contact-rich manipulation. *arXiv preprint arXiv:2310.16917*, 2023.

Appendix

A Details of exUMI Hardware Design

We gave a brief introduction to the hardware and algorithms for exUMI due to the page limit. Below are the details of our system.

A.1 Hardware

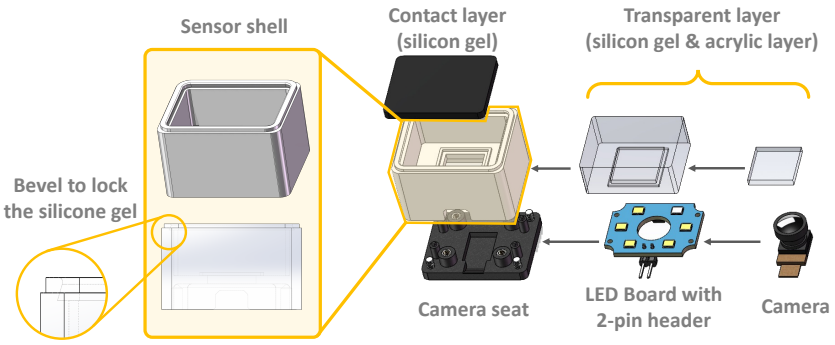

Figure 9: An exploded view of the enhanced tactile sensor design. We enhance the 9DTact for stability and quality control. We add an bevel to the sensor shell to secure the black silicon gel and prevent it from detaching.

Magnetic Rotary Encoder. We propose a low-cost AS5600 magnetic rotary encoder solution to achieve accurate and robust gripper state capture. As shown in Fig. 11, we modify the top cover of UMI to attach a radial magnet to one of the joints of the mechanical assembly, with the Hall sensor positioned above it at an appropriate distance ($\sim$ 2 mm). The AS5600 provides high-resolution 12-bit position readings (4,096 positions per revolution) and communicates with the single-board computer through the I^2C protocol. This solution offers a higher sampling rate and resolution, immunity to visual occlusion, and negligible computational overhead.

AR Motion Capture System. To overcome the limitations of vision-based tracking (SLAM) in occluded or complex scenarios, we adopt an AR-based approach for end-effector pose estimation. Following ARCap [45], our system uses a Meta Quest 3 headset for 6D pose motion capture, which is accurate and robust to occlusion. We integrate the left VR controller through a custom-designed mount attached to the UMI body, and use the headset to track the 6D pose of the controller. The mount also provides additional space for the power supply and an Orange Pi controller, serving as a universal sensor hub, synchronously capturing data from the AR headset, rotary encoder, and any additional sensors, such as tactile sensors. The tracking range of the system is 3 meters, sufficient for tasks at a typical robot workspace. While our system can also easily adapt to other trackers (*e.g.* HTC Vive trackers), our goal of using VR is the real-time evaluation of tracking and future extension of a user-friendly interface to guide the crowdsourcing.

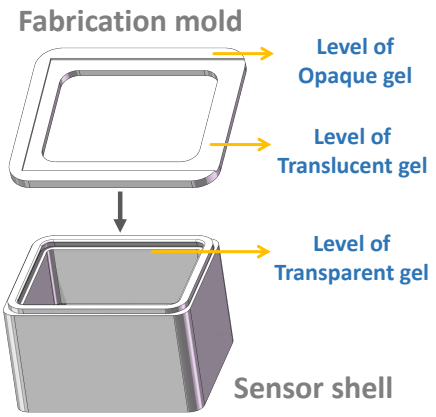

Figure 10: The mold for stable fabrication of the tactile sensor.

The orientation of the VR controller is arbitrary since the transformation between the controller and UMI coordinate frames will be determined through our calibration pipeline. This flexible mounting approach simplifies the assembly and avoids precise physical alignment.

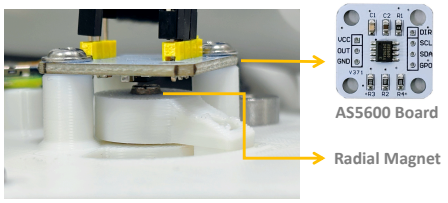

Figure 11: Detailed view of AS5600 sensor on exUMI.

Component	Base Cost ($)
GoPro 11 + Accessories	298
Meta Quest VR Headset	299
Orange Pi 3B	35
AS5600 Magnetic Encoder	1
3D Printed Parts	15
Visuo-Tactile Sensors	30
Misc. (Power Bank, Cables/Screws/Nuts)	20
Total Cost of exUMI	**698**

Table 4: Bill of materials (BOM) of exUMI.

Visual Input. Same as the UMI [1], we employ a GoPro camera with a fisheye lens as our primary visual input. We move the camera forward, following FastUMI [32] for a wider and clearer view. The camera positioning eliminates body occlusion in the field of view to enhance transferability.

Fingertip Visuo-tactile Sensors. We improve 9D-Tact [22] as a low-cost and DIY-friendly tactile senser for exUMI. We use 9DTact for low-cost. It is not as accurate as GelSight for surface geometry, but it provides sufficiently rich information about contact forces (normal+tangent) with fast deformation recovery, which is justified by our experiments.
As shown in Fig. 9 and Fig. 10, our enhancements involve: (1) We redesigned the sensor shell to securely anchor the top silicone layer, enhancing its resistance to tangent forces and ensuring long-term stability. (2) The LED board was modified to incorporate a more robust 2-pin header connector for stable power delivery. Compared to USB cables, Dupont connectors offer superior cable management flexibility. Plus, the LEDs were rearranged to minimize power consumption—a critical factor for our embedded system's efficiency. (3) A custom mold was developed to precisely control the silicone layer's thickness. The mold is affixed to the sensor shell, allowing controlled pouring of transparent, translucent, or opaque liquid silicone until it reaches the desired level (Fig. 10). Excess silicone is then removed by carefully scraping along the mold's surface with a spatula.
The upgraded sensor design achieves significantly enhanced durability and stability. Please refer to the appendix for more fabrication details.

Cost and Accessibility. We show the overall bill of materials in Tab. 4. Our system is low-cost with a minimal configuration starting at **$ 698**, which can be further reduced by substituting the GoPro with alternative fisheye cameras. The battery duration of Meta Quest headset is around 4 hours, and that of the Orange Pi system is over 10 hours. Our design is DIY-friendly and use readily available components, making it suitable for research and education. All CAD files will be released.

A.2 Data Collection and Processing

AR Capture Interface. Building upon the remarkable engineering of ARCap [45], we simplify the socket-based data transfer interface for the 6D pose capture process. The collection procedure is as follows:

1. Initialize the server program on the Raspberry Pi.
2. Launch the client application on the Meta Quest headset.
3. Set up the base coordinate frame in AR space. Then the headset can be **optionally** placed on a stand for convenience.
4. Begin data streaming of real-time 6D controller poses to the Raspberry Pi.

Figure 12: VR Headset Stand

Calibration of AR Latency. To synchronize AR motion capture with the visual inputs, we designed a calibration protocol involving horizontal sweeps in front of a stationary ArUco marker. We extract the x-axis movement of the AR MoCap system and the

Algorithm 1 Latency alignment algorithm

Input: Trajectories $f(t)$ and $g(t)$, timesteps $\{t_i\}_{i=1}^{T}$, bounds of latency δ_{min} δ_{max}
Input: Constants: $\epsilon = 0.0001$, search window N, search splits M
Output: Latency δ^* of $g(t)$ such that $f(t) \approx g(t + \delta^*)$

repeat
  Interpolate the interval $[\delta_{min}, \delta_{max}]$ into M segments: $\delta_0, \delta_1, \cdots, \delta_M$
  $k = \min_k \sum_{i=1}^{T} \|f(t_i) - g(t_i + \delta_k)\|_2^2$
  $\delta^* = \delta_k$
  $\delta_{min}, \delta_{max} \leftarrow \delta_{k-N}, \delta_{k+N}$ (update the search range to the neighborhood of δ_k)
until $\delta_{max} - \delta_{min} < \epsilon$

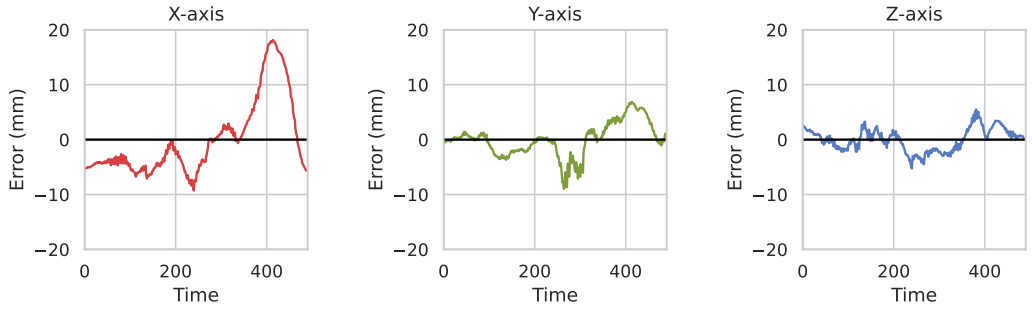

Figure 13: Comparison of AR MoCap trajectory and ground truth trajectory.

camera-detected marker trajectories, then use a bisection-style optimization algorithm to align the two trajectories and compute the latency of the AR system. Specifically, given the two 1-dimension trajectories on timesteps $\{t_i\}_{i=1}^{T}$, we convert them to two function $f(t)$ $g(t)$ *w.r.t* time t by interpolation, which is then calibrated by minimizing the MSE error between the two trajectory. The details are given in Alg. 1.

Data Collection and Processing Pipeline. Compared to the original UMI system, our data collection and processing pipeline is significantly simplified and more robust:

1. Set up the desired environment.
2. Initialize AR tracking system on the Raspberry Pi.
3. Record latency calibration sequence (one video).
4. Record demonstration videos.
5. Calculate and apply temporal latency correction.
6. Align AR capture data to the video frames through interpolation.
7. Pack synchronized data.

A.3 System Evaluation

Proprioception Precision. We first evaluated the precision of our proprioception system, particularly for the AR-based MoCap system. We obtained both ground truth and AR controller trajectories by mounting the AR controller on the robot end-effector and teleoperating the robot. We evaluate the system by moving the robot within a 50 cm range. The resulting 6D pose differences between the estimated and ground truth values are presented in Fig. 13 The system demonstrates remarkable accuracy, achieving mean position errors of 5.4 / 2.3 / 1.7 mm at each axis. The rotation errors are below 1 degree (notably small due to the robot's limited rotation range). The x-axis error reaches a maximum of 20 mm since it is the depth axis in the Flexiv coordinate system and inherently presents greater measurement challenges. These high-precision measurements enable efficient robot policy learning by providing high-quality training data.

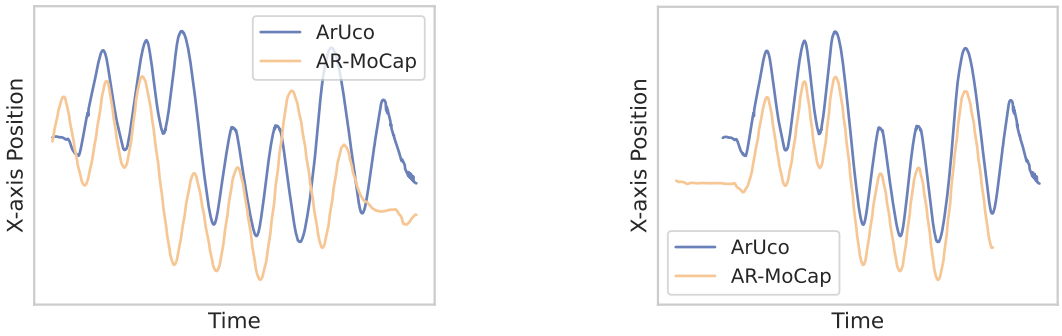

Figure 14: An example of AR MoCap and Vision trajectory before (left) and after (right) alignment.

Latency Calibration. We give a qualitative visualization of our latency calibration algorithm in Fig. 14. The x-axis trajectory of the AR MoCap system and the visual input (represented as the trajectory of the ArUco marker) are perfectly aligned after our calibration system, and we can read the time offset for further modality alignment. With proper sweeping frequency, our system could consistently achieve less than 5 ms latency error.

Key Takeaways

We provide an accurate, extensible, cost-effective, and DIY-friendly enhancement to the UMI system. Our solution is particularly valuable if you need:

- **Enhanced Tracking Precision**: Achieve $\sim$100% data effective ratio in diverse environments by our AR-based MoCap system and the magnetic rotary encoder for gripper state.
- **Non-Parallel Gripper Support**: An alternative design is provided for more popular industrial grippers like Flexiv Grav or Robotiq 2F series with an adaptable mechanical design.
- **Multimodal Sensing**: Including tactile, audio, or other custom sensors through our modular hardware interface.

All components are commercially available, and the fabrication details will be opened soon.

B Detailed Taxonomy of Tactile Representation Learning

We give more details of our discussion on the current taxonomy of tactile representation learning.

The target of tactile representation learning is to learn a tactile encoder $\mathcal{E}_T$ for the tactile data, to facilitate further multimodal policy learning:

$$\pi(\mathbf{a}_{t+1}|\mathcal{E}_S(\mathbf{s}_t), \mathcal{E}_T(\mathbf{T}_t), \mathcal{E}_V(\mathbf{V}_t)). \tag{1}$$

Current tactile representation methods fall into three dominant paradigms, each with specific advantages and fundamental constraints:

(a) Direct Multimodal Imitation Learning [13, 14, 19, 20, 21]. This approach trains end-to-end multimodal policies in Eq. 1 using paired tactile-visual-action data, and learn the tactile representation ϕ_T. While effective for narrow tasks, it suffers from tactile data scarcity, since the tactile contacts only occur in very few frames in most regular robot tasks (*e.g* $\sim$ 3000 frames in 100 demos of pick and place task). The method also inherently couples task objectives with tactile features, limiting cross-task transferability.

(b) Spatial Self-Supervised Learning [15, 16, 17, 18]. Self-supervised learning (SSL) is adopted for a generic and transferable tactile representation and learn tactile embeddings $\mathcal{E}_T(\mathbf{T}_t)$ through proxy objectives on task-agnostic unlabeled data. Mostly, SSL is adopted for tactile pretraining **spatially**, *i.e* treating tactile frames as images and applying image-level SSL algorithms. However, these spatial SSL methods potentially exploit incorrect inductive bias for tactile learning. Methods like contrastive learning [15, 16, 17] usually assumes translation invariance, which may not exist

Dataset	Data Scale	Tactile Sensor	Proprioception / Action	Collection Source
Calandra et al. [39]	6.5 K	GelSight	✓	Robot
Calandra et al. [40]	9.3 K	GelSight	✓	Robot
VisGel [42]	12.0 K	GelSight	✓	Robot
Burka [50]	1.1 K	Multiple	✓	Human
Touch and Go [44]	13.9 K	GelSight	✗	Human
ObjectFolder Real [51]	3.0 K	GelSight	✓	Robot
SSVTP [41]	4.5 K	DIGIT	✓	Robot
TVL [12]	43.7 K	DIGIT	✓	Robot
Touch2Touch [52]	32.3 K	Multiple	✓	Robot
X-Capture [43]	3.0 K	DIGIT	✗	Human
Ours	480.9 K (raw frames)	9DTact+	✓	Human

Table 5: Comparison of real-world tactile datasets.

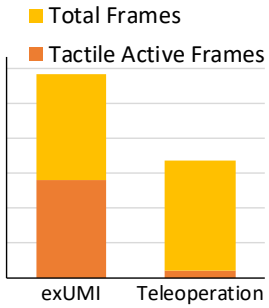

Figure 15: Comparison of per-hour tactile data collection efficiency.

in tactile sensing: any translation or shifting of tactile frames result in different contact point information. Similarly, masked learning [18, 48, 37] methods assumes that some image patches could be recovered from other patches, but tactile images do not hold geometrical self-consistency: *e.g* , a three-finger press and a four-finger press on a tactile sensor produce different signals, but partial masking could produce an identical image showing only two contact points. This ambiguity makes it impossible to recover the correct original input from the masked data. Therefore, it is challenging to design a self-supervising proxy objective for a robot task.

(c) Visual-Tactile Alignment [42, 43, 37]. Cross-modal alignment learns joint embeddings by maximizing in-pair similarity $s\left(\mathcal{E}_T(\mathbf{T}_t), \mathcal{E}_V(\mathbf{V}_t)\right)$ of visual and tactile modalities. Though it is effective for visual-language learning, it fundamentally assumes a coarse *one-to-one visuo-tactile mapping*, regardless of the actual *one-to-many relation*: with different contact forces, identical visual scenes yield divergent tactile signals. The multimodal alignment also overlooks that visual and tactile sensing are complementary rather than well-aligned. For robot learning, tactile sensors are a complementary information to the visual input, but the alignment method discards this privileged information.

Beside these approaches, some methods use pure generative models to learn a compact latent that preserves most tactile clues, such as auto-encoder [38] or VQ-GAN [49], which is more reasonable. To take a step further, we consider the proxy task of temporal prediction, by reformulating tactile representation as an action-conditioned temporal prediction problem, explicitly modeling the forward tactile dynamics that underpin real-world contact interactions.

C More Implementation Details

Tactile Data Curation. Since active tactile signals are sparse in real-world data collection, we adopt a data rejection strategy during data sampling to avoid trivial samples. For each data chunk, we check the proportion of active pixels for each tactile frame. If the active proportion of all frames is below a certain threshold, the data chunk will be discarded and resampled.

Implementation Details. For each timestep, our exUMI collects two tactile images on the two sides of the gripper. We convert the images to a calibrated grayscale image following 9DTact [22], and extract the convex and concave pixel map by comparing the grayscale image to the reference image (tactile signal at no contact). The grayscale image, convex map, and concave map are stacked as a 3-channel image for a richer representation of tactile contacts.

We pretrain the tactile representation on our large-scale human play dataset, which is randomly split into a training and validation set by 15:1. The action sequence is represented as the relative pose and the gripper state. The images on two sides are concatenated and then downsampled to 224×224 resolution. We use a pretrained VAE model (KL-F16) as the encoder and decoder for tactile learning. The tactile prediction is conducted in 8 temporal frames, where 4 random frames

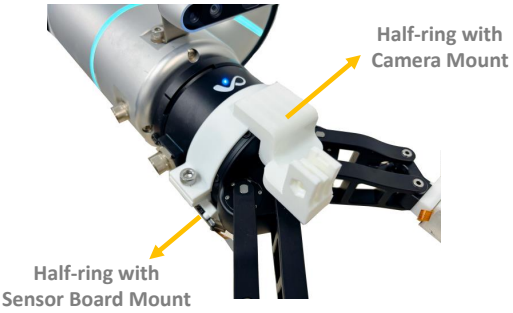

Figure 16: Pipe clamp style GoPro mount for deployment.

Modality Tactile Learning	V /	V+T Direct	V+T BYOL	V+T TPP (Ours)
Put Ball	70%	70%	80%	**85%**
Peg in Hole	50%	60%	50%	**80%**

Table 6: Real world evaluation of tactile represention learning algorithms. V: vision; T: tactile.

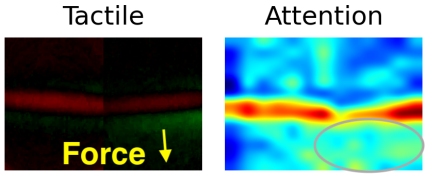

Figure 17: Attention Map.

in the first half are regarded as input and 4 frames in the second half are the prediction target. We adopt a larger frequency for action following Li et al. [46]. The tactile prediction model reaches quick convergence due to the simpler distribution of tactile sensing.

Environment Details. The camera and sensors are connected to a master computer with a RTX 4070 GPU, which controls the robot at a frequency of 10 Hz. As illustrated in Fig. 16, we designed a pipe clamp-style camera mount and gripper tactile sensor mount to exactly replicate the end-effector sensor placement of the exUMI system. The mount consists of two half-rings, one integrated with a standard GoPro mount. This design can be attached to the Flexiv Grav gripper base, and also adaptable on various other end effectors.

Task Details. (1) *Pick cube / carrot / broccoli*: picking up an object and place it into a container. The object and the target container are randomized within a 30cm×30cm area. (2) *Insert pen*: moving a pen from one cup to another cup. The cups are randomized within a 30cm×30cm area, and the pen is randomly placed in the cup. The colors are fixed. (3) *Stack cubes*: stacking a 5cm×5cm×5cm cube on top of another, requiring precise manipulation ability. The cubes are randomized within a 30cm×30cm area. The colors are fixed.

For tactile-aware policy learning, we evaluate on more complex tasks: (1) *Put Ball*: pick up a yellow soft ball (radius= 3cm) and place it in a red cup. The cubes are randomized within a 30cm×30cm area. (2) *Open Bottle*: rotate the bottle cap until it is fully unscrewed. The bottle has diameter= 6.5cm and height= 13cm, and is randomized within a 20cm×20cm area. (3) *Pull Drawer*: pull out a drawer, which is either empty ("Empty") or contains a random amount of stones ("Random", randomized from 50g to 1000g), requiring tactile clues to determine the pulling direction. (4) *Peg in hole*: insert a 4 cm × 3 cm yellow block into a 4.3 cm × 3.3 cm slot, requiring a precision and force-aware adjustment. The yellow block is randomly put on a 10cm×10cm blue cube. We split the task into "Grasp" and "Insert" stages.

D Ablation Experiments

D.1 Attention Visualization

We visualize the attention map of tactile encoder pretrained by TPP. We give an example in Fig. 17 on "pull drawer" task, where the arrow shows the tangent force direction. The pretrained tactile

model focuses on the area that indicates force magnitude (red area), and also the direction (in the circle).

D.2 Tactile Learning Comparison

To compare our representaion learning algorithm, we implement the direct learning method, and a spatial self-supervised learning method BYOL following MimicTouch [53], which is pretrained on our collected tactile dataset. The results are shown in Tab. 6. TPP shows the best success rate on two tasks.

E Broader Impact

The open-source design and affordability ($698, and we are working on making it less than $500) of the exUMI system democratize tactile robotics research by lowering technical and financial barriers for resource-constrained labs and educational institutions. This accessibility can accelerate innovation and broaden participation in the field. Beyond the research community, this work has direct societal applications. In assistive robotics, it enables robots to perform delicate tasks for the elderly or individuals with disabilities. In industrial safety, sophisticated tactile sensing allows robots to operate more safely alongside human workers and enabling the reliable handling of fragile components. By making advanced tactile learning more accessible, our work helps pave the way for robots that can interact with the physical world more safely and intelligently.