LIFELONGTOUCH: LIFECYCLE-INVARIANT TACTILE REPRESENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-based tactile sensors (VBTSs) provide rich contact information, yet their observations are susceptible to variations across sensors and state changes during long-term use. Existing tactile representation learning methods primarily focus on cross-sensor generalization, typically treating each sensor as a static domain and thus struggling with evolving sensor states during long-term deployment. To address this limitation, we propose **LifelongTouch**, a lifecycle-invariant tactile representation learning framework for generalization across sensors and lifecycle states. LifelongTouch jointly models contact observations and no-contact references under the current sensor state, and combines *reference-conditioned physical canonicalization*, *physical geometry supervision*, and *sensor-state physical alignment* to learn consistent physical representations across sensors and lifecycle states. Furthermore, we develop a lifecycle-aligned tactile simulation engine that provides cross-sensor, cross-state tactile observations with geometric supervision for representation learning. Experiments on object classification, contact pose estimation, force distribution estimation, and surface reconstruction demonstrate that LifelongTouch generalizes robustly to unseen sensors and lifecycle states, providing an effective approach toward reliable lifelong tactile perception. The data and code will be publicly released upon acceptance.

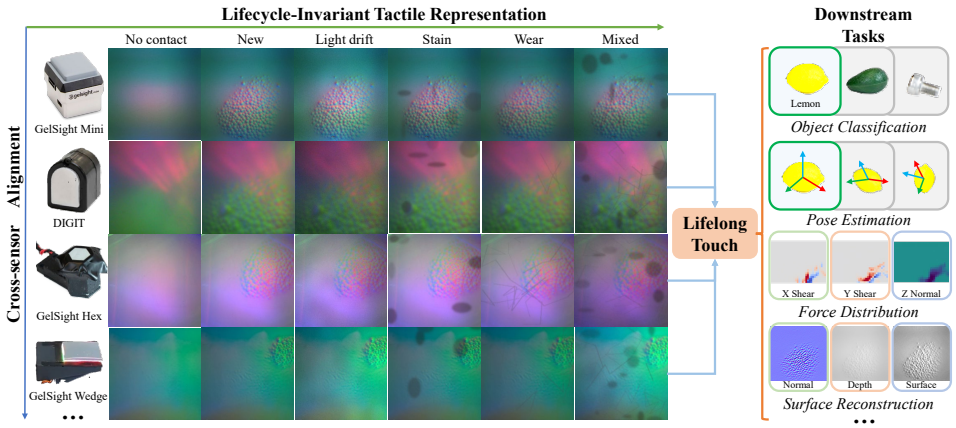

Figure 1: **Lifelong tactile perception.** LifelongTouch is a lifecycle-invariant tactile representation learning framework that generalizes across sensors and lifecycle states, enabling robust transfer to diverse downstream tasks.

1 INTRODUCTION

Tactile perception enables robots to acquire contact and physical interaction information that is difficult to perceive directly through vision. Vision-based tactile sensors (VBTS), particularly GelSight and its variants (Yuan et al., 2017a; Wang et al., 2021; Lambeta et al., 2020; Zhang et al., 2022; Li et al., 2026), convert subtle deformations of compliant surfaces into high-resolution visual signals through optical imaging, thereby providing rich geometric and contact information. Therefore, VBTS have been widely used for object reconstruction (Suresh et al., 2024; Comi et al., 2025), material perception (Yuan et al., 2018; Fu et al., 2024), slip detection (Dong et al., 2017; 2019; Feng

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2026), force estimation (Chen et al., 2024; Higuera et al., 2025), and fine-grained robotic manipulation (Yu et al., 2025; Xue et al., 2025; Niu et al., 2026).

However, VBTS observations are highly dependent on sensor-specific characteristics and lifecycle states. Differences in cameras, illumination, materials, and manufacturing processes across sensors, together with long-term changes such as illumination drift, gel aging, coating wear, and surface contamination, can significantly alter the distribution of tactile observations (Abad & Ranasinghe, 2020; Cottrell et al., 2026). As a result, pre-trained models may struggle to generalize not only across different sensors but also across the evolving states of the same sensor, while frequent recalibration, data recollection, and model updates are costly in practical deployment.

To address these challenges, one line of work learns general-purpose tactile representations through large-scale self-supervised learning, such as Sparsh (Higuera et al., 2025) and UniT (Xu et al., 2025), primarily aiming to improve transfer across diverse perception and manipulation tasks. Building on this, methods such as T3 (Zhao et al., 2025) and AnyTouch (Feng et al., 2025) incorporate data from multiple tactile sensors to improve generalization across both sensors and tasks. Another line of work directly focuses on cross-sensor tactile representation learning, such as SITR (Gupta et al., 2025) and CTTP (Rodriguez et al., 2025), which enable cross-sensor transfer through sensor-invariant representation learning. However, most existing methods still treat each sensor as a static domain with a fixed observation distribution, overlooking the continual state changes that occur within the same sensor during long-term use.

To fill this gap, we propose **LifelongTouch**, a lifecycle-invariant tactile representation learning framework designed to learn consistent physical representations across heterogeneous sensors and their evolving lifecycle states. Specifically, LifelongTouch jointly models a contact image with a no-contact reference acquired under the current sensor state. Since sensor-state variations affect tactile observations through complex optical and material responses, their effects cannot be reliably removed by direct subtraction. LifelongTouch therefore introduces *reference-conditioned physical canonicalization*, which adaptively modulates contact features using the reference to reduce state-dependent variations. However, reference conditioning alone may still allow the model to exploit appearance-related shortcuts. To encourage physically meaningful representations, *physical geometry supervision* constrains the learned features with dense contact geometry. Finally, *sensor-state physical alignment* explicitly aligns representations of the same physical contact across different sensors and lifecycle states, further promoting sensor- and state-invariant tactile representations.

To support the above learning paradigm, we develops a lifecycle-aligned tactile simulation engine that systematically varies sensor designs and lifecycle states while preserving the underlying physical contact, generating precisely aligned cross-sensor and cross-state tactile observations with dense geometric supervision. This enables the model to learn a unified tactile representation that preserves contact-related physical information while remaining robust to lifecycle variations. Experiments on object classification, contact pose estimation, force distribution estimation, and surface reconstruction demonstrate that LifelongTouch generalizes robustly to unseen sensors and lifecycle states, validating its effectiveness for reliable long-term tactile perception.

The main contributions are summarized as follows:

- Lifelong tactile perception is formulated to extend tactile generalization from static cross-sensor transfer to evolving sensor states during long-term deployment.
- LifelongTouch is proposed to learn lifecycle-invariant tactile representations through reference-conditioned canonicalization, physical geometry supervision, and sensor-state physical alignment.
- A lifecycle-aligned tactile simulation engine is developed to provide aligned cross-sensor and cross-state observations with geometric supervision. Extensive experiments across diverse downstream tasks demonstrate robust generalization to unseen sensors and lifecycle states.

2 Related Work

General-purpose Tactile Representation. Since VBTSs encode physical contact as high-dimensional image-like signals, visual representation learning has been widely adopted for tactile perception (Li et al., 2019; Calandra et al., 2018). Early studies mainly explored contrastive learning and visuo-tactile alignment to learn transferable tactile features (Yuan et al., 2017b; Yang et al., 2022;

Kerr et al., 2023; Guzey et al., 2023). With the emergence of large-scale tactile datasets, recent works have increasingly focused on general-purpose representations that transfer across sensors, tasks, and interaction settings. TacMAE (Cao et al., 2023) learns dense tactile features through masked autoencoding, UniT (Xu et al., 2025) learns a compact latent space with VQGAN, and Sparsh (Higuera et al., 2025) performs large-scale self-supervised pre-training across multiple VBTS datasets. T3 (Zhao et al., 2025) combines sensor-specific encoders with a shared Transformer to transfer knowledge across sensors and tasks, while UniTouch (Yang et al., 2024) introduces sensor-specific tokens for tactile-semantic alignment. AnyTouch (Feng et al., 2025) further learns unified representations from multiple tactile sensors, and AnyTouch2 (Feng et al., 2026) extends this paradigm from static perception to dynamic interaction.

Cross-Sensor Tactile Representation. Despite increasing efforts to learn from multi-sensor tactile data, substantial domain gaps persist due to variations in imaging systems, optical configurations, materials, and sensor geometries. To bridge these gaps, recent studies have explicitly explored sensor-invariant representations for generalization to unseen sensors. CTTP (Rodriguez et al., 2025) aligns paired cross-sensor observations through contrastive learning, while SITR (Gupta et al., 2025) leverages shared contact geometry to learn sensor-invariant representations for zero-shot transfer to unseen sensors. CTSRL (Zhang et al., 2026) further integrates cross-sensor self-supervision with semantic alignment to improve generalization across heterogeneous tactile sensors.

However, general-purpose and cross-sensor approaches predominantly treat each sensor as a static domain with a fixed observation distribution. In practice, the same sensor evolves throughout its lifecycle due to illumination drift, material aging, surface wear, contamination, and damage. To address this issue, LifelongTouch extends tactile representation learning beyond static cross-sensor generalization by addressing variations across sensor identities and their evolving lifecycle states.

3 METHOD

We consider *lifelong tactile perception*, where a robot is expected to maintain stable tactile perception throughout the sensor's operational lifetime without storing historical data or continuously updating the model, relying only on information about the sensor's current state. Formally, let s, u, and q denote a physical tactile sensor, its lifecycle state, and the underlying physical contact, respectively. Our goal is to learn a generalizable tactile representation $\mathbf{F}^u_{s,q}$ that preserves information relevant to the physical contact q while suppressing variations caused by sensor identity and lifecycle state. For the same physical contact observed under different sensors and lifecycle states, the learned representations should satisfy $\mathbf{F}^u_{s,q} \approx \mathbf{F}^{u'}_{s',q}$, thereby enabling consistent tactile perception across sensors and throughout their lifecycles.

Inspired by the human tactile system, which maintains stable perception despite individual differences and gradual changes in sensory conditions, we seek to interpret tactile observations relative to the sensor's current state rather than relying on a fixed sensor appearance. As illustrated in Fig. 2, LifelongTouch realizes this idea through three complementary components: *reference-conditioned physical canonicalization*, *physical geometry supervision*, and *sensor-state physical alignment*.

3.1 REFERENCE-CONDITIONED PHYSICAL CANONICALIZATION

A straightforward approach is to subtract the no-contact reference from the contact image to suppress sensor-dependent appearance variations (Gupta et al., 2025; Zhang et al., 2026). However, the contact response of a VBTS is nonlinearly modulated by sensor-state factors such as illumination, material properties, and camera characteristics, making simple pixel-level subtraction insufficient for effective disentanglement. Therefore, LifelongTouch learns a nonlinear reference-conditioned canonicalization in feature space to obtain a more stable physical tactile representation.

Specifically, given a contact image $\mathbf{X}^u_{s,q}$ and its no-contact reference $\mathbf{B}^u_s$, dense patch features are extracted using a shared encoder $\mathcal{E}$:

$$\mathbf{F}_c = \mathcal{E}(\mathbf{X}^u_{s,q}), \qquad \mathbf{F}_r = \mathcal{E}(\mathbf{B}^u_s). \tag{1}$$

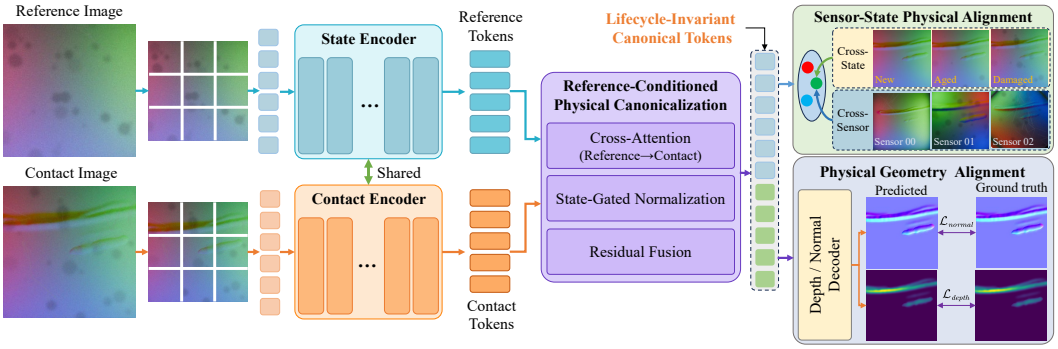

Figure 2: **Overview of LifelongTouch.** Given a contact image and its corresponding no-contact reference, LifelongTouch learns lifecycle-invariant tactile representations through reference-conditioned canonicalization, physical geometry supervision, and sensor-state physical alignment.

To condition the contact representation on the current sensor state, the reference feature is mapped to spatially varying modulation parameters:

$$(\boldsymbol{\gamma}_s^u, \boldsymbol{\beta}_s^u) = \mathcal{M}(\mathbf{F}_r), \qquad \widetilde{\mathbf{F}}_c = (1 + \boldsymbol{\gamma}_s^u) \odot \mathrm{LN}(\mathbf{F}_c) + \boldsymbol{\beta}_s^u. \tag{2}$$

This feature-wise modulation adaptively compensates for local sensor-state variations.

Beyond modulation, cross-attention captures spatial correspondences between the contact and reference features:

$$\mathbf{F}_{r\to c} = \mathrm{CrossAttn}\left(\widetilde{\mathbf{F}}_c, \mathbf{F}_r, \mathbf{F}_r\right). \tag{3}$$

Their complementary information is then encoded into a relational feature:

$$\mathbf{R} = \left[\widetilde{\mathbf{F}}_c, \mathbf{F}_{r\to c}, \widetilde{\mathbf{F}}_c - \mathbf{F}_{r\to c}, \widetilde{\mathbf{F}}_c \odot \mathbf{F}_{r\to c}\right]. \tag{4}$$

Finally, a lightweight Transformer Φ maps the relational feature to a canonical correction, which is added residually to the original contact feature:

$$\mathbf{F}_{s,q}^u = \mathbf{F}_c + \Phi(\mathbf{R}). \tag{5}$$

This residual formulation preserves contact information from the pretrained representation while correcting sensor-state-dependent variations.

3.2 Shared Physical Geometry Supervision

Although tactile appearance can vary substantially across sensors and lifecycle states, the underlying contact geometry remains largely consistent for the same physical interaction. Based on this property, LifelongTouch employs surface normals and depth as complementary geometric supervision to anchor the canonicalized representation in a stable physical space. Surface normals characterize local surface orientation and shape, while depth captures the spatial structure and magnitude of contact deformation, jointly encouraging physically consistent representations across sensors and states.

Specifically, given the canonicalized representation $\mathbf{F}_{s,q}^u$, two lightweight geometry decoders predict the surface-normal and depth fields:

$$\widehat{\mathbf{N}}_q = \mathcal{D}_N\left(\mathbf{F}_{s,q}^u\right), \qquad \widehat{\mathbf{Z}}_q = \mathcal{D}_Z\left(\mathbf{F}_{s,q}^u\right), \tag{6}$$

which are optimized with the joint geometry objective:

$$\mathcal{L}_{\mathrm{geo}} = \mathcal{L}_{\mathrm{normal}} + \lambda_d \mathcal{L}_{\mathrm{depth}}. \tag{7}$$

Here, $\mathcal{L}_{\mathrm{normal}}$ encourages consistency between the predicted and ground-truth surface-normal fields, while $\mathcal{L}_{\mathrm{depth}}$ supervises depth estimation over the valid contact region. The scalar coefficient λ_d balances the two geometric objectives.

3.3 SENSOR-STATE PHYSICAL ALIGNMENT

While geometric supervision anchors the representation to physical contact information, it does not explicitly enforce invariance to sensor- and lifecycle-dependent observation shifts. As shown in Fig. 3, our lifecycle-aligned simulation engine provides a stronger constraint: the same physical contact q can be rendered under different sensor identities and lifecycle states while preserving exact spatial correspondence. These physically equivalent observations therefore provide direct supervision for learning state-invariant representations.

For K observations of the same physical contact q, their representations $\mathbf{F}_{s_k,q}^{u_k}$ are projected into a shared normalized space:

$$\mathbf{Z}_{s_k,q}^{u_k} = \mathrm{Norm}\left(\mathcal{G}\left(\mathbf{F}_{s_k,q}^{u_k}\right)\right), \qquad k = 1, \ldots, K, \tag{8}$$

where (s_k, u_k) correspond to different lifecycle states of the same sensor or to different physical sensors. Since the simulation preserves spatial correspondence across these observations, alignment is performed directly at the patch-token level rather than on globally pooled features.

The shared alignment center is defined as:

$$\overline{\mathbf{Z}}_q = \mathrm{sg}\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{Z}_{s_k,q}^{u_k}\right), \tag{9}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator.

The cross-state physical alignment loss is then computed over the P spatial patch tokens:

$$\mathcal{L}_{\mathrm{align}} = \frac{1}{KP}\sum_{k=1}^{K}\sum_{p=1}^{P}\left[1 - \left(\mathbf{Z}_{s_k,q,p}^{u_k}\right)^{\top}\overline{\mathbf{Z}}_{q,p}\right]. \tag{10}$$

This objective encourages representations of the same physical contact to remain consistent across lifecycle states and sensor identities.

Finally, the overall learning objective is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{geo}} + \lambda_{\mathrm{align}}\mathcal{L}_{\mathrm{align}}, \tag{11}$$

where λ_{align} controls the contribution of cross-state physical alignment.

4 DATASET

4.1 SYNTHETIC PRE-TRAINING DATASET

Collecting long-term tactile data across diverse physical sensors is costly and difficult to scale. As illustrated in Fig. 3, a lifecycle-aligned data engine is built upon physics-based VBTS rendering (Agarwal et al., 2021; Si et al., 2024; Gupta et al., 2025). The generation process factorizes each tactile observation into *sensor identity*, *lifecycle state*, and *physical contact*, enabling the same contact to be rendered across different sensor states with aligned depth and surface-normal supervision.

Sensor and lifecycle diversity. Sensor diversity is introduced by varying camera field of view, sensing geometry, elastomer properties, and illumination configurations, covering a broad range of GelSight-like designs (Yuan et al., 2017a; Dong et al., 2017; Taylor et al., 2022; Wang et al., 2021; Lambeta et al., 2020). For each virtual sensor, lifecycle trajectories are generated by progressively varying illumination, camera response, gel properties, coating, contamination, and damage, together with stochastic events such as LED failure, camera displacement, cleaning, and gel-pad replacement.

Physical interactions. Physical contacts are independently sampled by varying object geometry, contact position, indentation depth, and interaction motion, including sliding, rotation, pressing, and their combinations. Decoupling physical interaction from sensor and lifecycle variations provides explicit physical correspondences across different sensor states.

The resulting pre-training dataset contains 100 virtual sensors, each with 8 lifecycle trajectories of 16 states, yielding 12,800 sensor-lifecycle states and 870,400 tactile observations at 224×224 resolution. Each observation is accompanied by depth and surface-normal ground truth. Detailed simulation settings are provided in Appendix A.2.1.

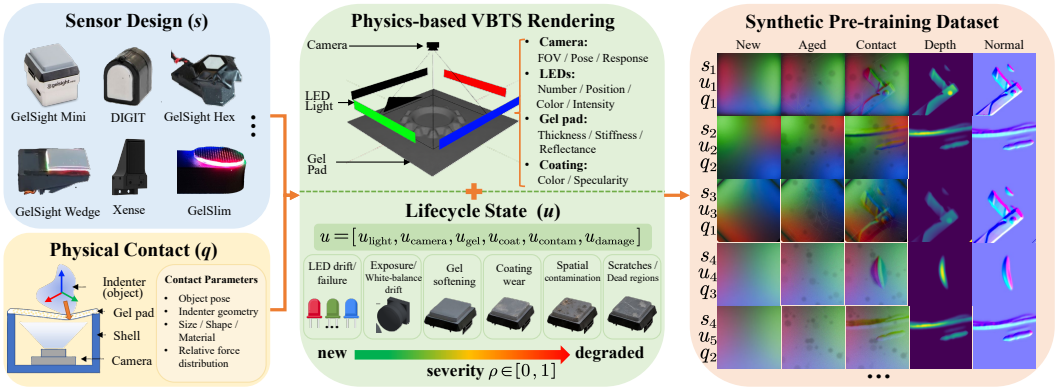

Figure 3: **Lifecycle-aligned data engine.** The same physical contact q is rendered across sensor designs s and lifecycle states u, providing aligned observations and geometric supervision.

4.2 Real-World Lifetime Benchmark

To evaluate the generalization of LifelongTouch on real tactile observations from complementary perspectives, we construct two benchmarks, LifetimeSITR and LifetimeFEATS, based on SITR-Dataset (Gupta et al., 2025) and FEATS-Dataset (Helmut et al., 2025), respectively. The two benchmarks emphasize different generalization dimensions. LifetimeSITR primarily evaluates cross-sensor generalization, with controlled pseudo-lifecycle degradation serving as an additional robustness test against sensor-state variations. In contrast, LifetimeFEATS exploits physically collected variations across different gel conditions and primarily evaluates generalization across real lifecycle states. Further construction details are provided in Appendix A.2.2.

LifetimeSITR evaluates object classification and contact pose estimation across different tactile sensors. SITR-Dataset contains observations from multiple real sensors but does not explicitly capture the long-term evolution of an individual sensor over its lifetime. We therefore use cross-sensor transfer as the primary evaluation objective of this benchmark, while introducing controlled pseudo-lifecycle degradation as an auxiliary stress test. The underlying physical contacts and task labels are preserved, and spatially consistent perturbations are applied to approximate sensor-state variations such as illumination drift, surface wear, and persistent contamination. Five conditions are considered: *New*, *Light-drift*, *Wear*, *Stain*, and *Mixed*, under *No Transfer*, *Intra-Sensor Transfer*, and *Inter-Sensor Transfer* settings.

LifetimeFEATS, in contrast, focuses on dense force-distribution estimation under real lifecycle variations. All tactile observations in FEATS-Dataset are collected using GelSight Mini sensors, while their gels exhibit different physically realized conditions associated with usage and degradation. It therefore enables evaluation of lifecycle-induced distribution shifts without relying on synthetic image degradation. The dataset additionally contains variations in contacted objects and pressing locations, allowing us to examine generalization when lifecycle changes coexist with contact variation. We consider four settings: *No Transfer* (*NoT*), *Unseen Sensor with New Gel* (*US-New*), *Unseen Sensor with Aged Gel* (*US-Aged*), and *Unseen Object Press* (*UO-Press*), where *US-Aged* provides the primary evaluation of generalization to real lifecycle states.

5 Experiments

Baselines. We compared LifelongTouch with ImageNet-pretrained ViT (Dosovitskiy et al., 2021), general-purpose tactile encoders including Sparsh (Higuera et al., 2025), AnyTouch (Feng et al., 2025), and T3 (Zhao et al., 2025), and cross-sensor encoders including CTTP (Rodriguez et al., 2025), SITR (Gupta et al., 2025), and CTSRL (Zhang et al., 2026).

Implementation Details. LifelongTouch used a DINOv2-initialized ViT-B/14 (Oquab et al., 2024) and was pre-trained on the synthetic dataset for 30 epochs with AdamW. Results were averaged over three random seeds. Further details are provided in Appendix A.1.

Table 1: **Object classification accuracy (%, ↑) on LifetimeSITR.** *Aged Avg.* denotes the average over Light-drift, Wear, Stain, and Mixed lifecycle conditions.

	New			*Light-drift*			*Wear*			*Stain*			*Mixed*			*Aged Avg.*		
Method	NoT	Inter	Intra	NoT	Inter	Intra	NoT	Inter	Intra	NoT	Inter	Intra	NoT	Inter	Intra	NoT	Inter	Intra
CTSRL	90.14	32.69	45.34	58.11	18.97	25.37	91.27	32.75	45.15	89.96	32.60	45.21	80.21	29.68	35.24	79.89	28.50	37.74
T3-Medium	83.62	19.33	28.62	81.13	19.89	25.50	84.00	18.83	28.96	83.50	19.26	28.70	82.95	19.18	27.78	82.90	19.29	27.74
ViT-Base	96.68	56.16	73.15	93.49	51.76	66.99	97.15	**59.06**	77.55	96.66	56.26	73.36	96.23	52.31	76.60	95.88	54.85	73.63
Sparsh	59.56	8.18	13.23	54.29	8.04	10.92	61.74	8.71	13.63	60.27	8.22	13.41	59.95	7.82	13.07	59.06	8.20	12.76
AnyTouch	76.23	13.65	30.76	60.79	13.23	23.73	77.51	14.16	31.34	76.28	13.67	30.66	71.57	14.61	28.28	71.54	13.92	28.50
CTTP	39.21	13.82	15.71	36.71	13.53	15.51	39.76	14.02	15.80	39.08	13.82	15.58	37.44	13.64	15.89	38.25	13.75	15.70
SITR	97.46	56.61	80.42	90.87	51.06	66.97	97.53	57.48	80.80	**97.38**	56.52	**80.21**	**97.11**	**55.41**	78.85	95.72	55.12	76.71
LifelongTouch	**98.36**	**57.42**	**82.77**	**97.68**	**57.28**	**79.08**	**98.28**	55.58	**81.62**	95.70	**56.82**	79.67	96.48	54.18	**79.37**	**97.04**	**55.97**	**79.94**

Table 2: **Contact-pose estimation RMSE (mm, ↓) on LifetimeSITR.**

	New		Light-drift		Wear		Stain		Mixed		Aged Avg.	
Method	NoT	Inter	NoT	Inter	NoT	Inter	NoT	Inter	NoT	Inter	NoT	Inter
CTSRL	0.5402	0.9880	0.6176	1.0192	0.5378	0.9980	0.5413	0.9931	0.5471	1.0060	0.5610	1.0041
T3-Medium	0.5333	1.1164	0.5510	1.1519	0.5321	1.1086	0.5353	1.1158	**0.5360**	1.1216	0.5386	1.1245
ViT-Base	1.6452	1.6840	1.6510	1.6737	1.6415	1.6812	1.6445	1.6836	1.6473	1.6781	1.6461	1.6792
Sparsh	0.7312	1.4301	0.7653	1.4469	0.7129	1.4199	0.7289	1.4267	0.7199	1.4306	0.7318	1.4310
AnyTouch	0.7288	1.5355	1.0364	1.5991	0.6841	1.5067	0.7247	1.5280	0.7802	1.5367	0.8064	1.5426
CTTP	0.6846	1.1889	0.6985	1.1901	0.7005	1.2155	0.6988	1.1997	0.7151	1.2072	0.7032	1.2031
SITR	0.5438	**0.8903**	0.6735	0.9693	0.5380	**0.8675**	0.5443	**0.8894**	0.5519	**0.9011**	0.5769	**0.9068**
LifelongTouch	**0.3334**	0.9193	**0.4160**	**0.9173**	**0.4479**	0.9468	**0.4083**	0.9347	0.6034	1.0063	**0.4689**	0.9513

5.1 Cross-Sensor and Cross-Lifecycle Transfer

To evaluate the cross-sensor and cross-lifecycle generalization of LifelongTouch, all downstream task heads were trained using data from a single source sensor in the *New* state and directly transferred to other sensors and lifecycle states without any additional adaptation.

Object classification. Tab. 1 reports the object classification results on LifetimeSITR. Overall, LifelongTouch achieves the strongest performance across different lifecycle states. Averaged over the four aged conditions, including Light-drift, Wear, Stain, and Mixed, LifelongTouch reaches accuracies of 97.04%, 55.97%, and 79.94% under the NoT, Inter, and Intra settings, respectively, consistently outperforming SITR at 95.72%, 55.12%, and 76.71%. In particular, under Light-drift, LifelongTouch achieves an Inter-Sensor accuracy of 57.28%, substantially higher than SITR at 51.06%, while also improving the Intra-Sensor accuracy from 66.97% to 79.08%. These results indicate that conditioning on the current sensor state effectively suppresses lifecycle-induced appearance shifts while preserving discriminative contact information. LifelongTouch also maintains stable performance under Wear, Stain, and Mixed conditions, and achieves the best aged-state average across all three evaluation settings.

Contact-pose estimation. Tab. 2 further evaluates metric contact geometry. LifelongTouch achieves the best No-Transfer performance under most lifecycle conditions, with an aged-state average RMSE of 0.4689 mm. For Inter-Sensor transfer, SITR remains stronger overall, achieving an aged-state average of 0.9068 mm versus 0.9513 mm for LifelongTouch. Notably, LifelongTouch performs best under Light-drift, reducing the RMSE from 0.9693 mm to 0.9173 mm, indicating improved robustness to illumination-related lifecycle shifts.

Overall, the LifetimeSITR experiments show that LifelongTouch consistently improves cross-lifecycle object recognition while maintaining competitive cross-sensor pose transfer. It should be noted that methods such as SITR and T3 adopt a "predicted surface normals $\rightarrow$ ResNet" downstream pipeline, whose final performance is jointly influenced by the specific geometric intermediate representation and task-head design, making it difficult to directly assess the general representation capability of the encoder itself. We therefore further conduct a unified frozen-encoder linear-probe evaluation in Appendix A.4.1. Under the same task heads, LifelongTouch exhibits stronger transferability on both classification and pose estimation, indicating that its encoded features more consistently preserve discriminative and geometric information shared across sensors and lifecycle states.

Table 3: **Force distribution estimation on LifetimeFEATS.** RMSE evaluates the dense three-axis force field and MAE evaluates integrated total force, both in newtons. *Transfer Avg.* is averaged over US-New, US-Aged, and UO-Press.

		NoT		*US-New*		*US-Aged*		*UO-Press*		*Transfer Avg.*	
Input	Method	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓
Contact-only protocol											
Contact	FEATS U-Net	0.01675	1.580	0.01629	1.105	0.01629	1.231	0.01508	1.267	0.01589	1.201
Contact	T3-Medium	**0.00841**	0.890	**0.01007**	**1.003**	**0.00988**	1.556	**0.00769**	0.968	**0.00921**	1.176
Contact	ViT-Base	0.00895	**0.715**	0.01069	1.147	0.01004	**0.940**	0.00936	0.731	0.01003	**0.939**
Contact	AnyTouch	0.00901	1.060	0.01036	1.110	0.01034	1.519	0.00787	**0.687**	0.00952	1.105
Reference-assisted protocol											
Contact+Ref.	Sparsh	0.01044	1.131	0.01195	3.216	0.01082	1.975	0.00958	0.893	0.01078	2.028
Contact+Ref.	CTSRL	0.00917	0.876	0.00921	**0.747**	0.00952	**0.741**	0.00872	0.790	0.00915	0.759
Contact+Ref.	CTTP	0.01283	1.442	0.01253	1.440	0.01240	1.071	0.01247	1.325	0.01247	1.279
Contact+Ref.	SITR	0.01079	1.157	0.01199	1.225	0.01174	0.873	0.00941	0.752	0.01105	0.950
Contact+Ref.	**LifelongTouch**	**0.00737**	**0.333**	**0.00908**	0.805	**0.00879**	0.778	**0.00675**	**0.349**	**0.00820**	**0.644**

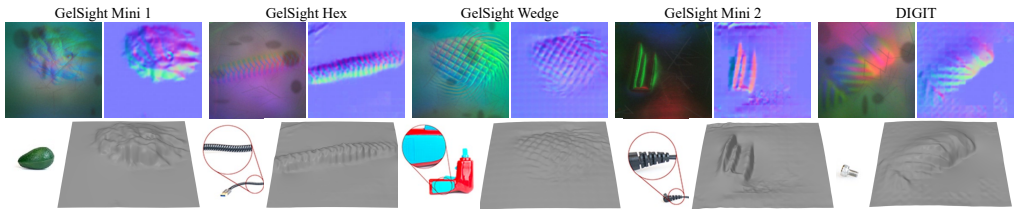

Figure 4: **Zero-shot surface reconstruction.** LifelongTouch consistently reconstructs fine-grained contact geometry across different sensors and lifecycle states.

5.2 Dense Force-Distribution Transfer

Since object classification and contact pose estimation rely heavily on contact geometry, we further evaluate whether LifelongTouch preserves richer local contact physics through dense force-distribution estimation on LifetimeFEATS. The prediction target is a $24 \times 32 \times 3$ force field containing two shear components and one normal-force component. Tab. 3 groups methods by their input protocol (Contact and Contact+Ref.), with comparisons made primarily within the same protocol.

Under the reference-assisted setting, LifelongTouch achieves the best overall force-transfer performance, with the lowest Transfer Avg. RMSE of 0.00820 N and MAE of 0.644 N. It also obtains the best RMSE across all evaluation settings, including NoT, US-New, US-Aged, and UO-Press, while achieving particularly low MAE on NoT (0.333 N) and UO-Press (0.349 N). Compared with CTSRL, which reaches 0.00915 N RMSE and 0.759 N MAE on Transfer Avg., LifelongTouch shows stronger generalization of dense physical information across unseen sensors, aged states, and contact conditions. Together with the LifetimeSITR results, these findings show that LifelongTouch can transfer the physical invariances learned from synthetic lifecycle variations to real tactile observations and real sensor degradation, effectively narrowing the sim-to-real gap while preserving semantic, geometric, and dense physical information.

5.3 Zero-Shot Surface Reconstruction

Finally, qualitative zero-shot surface reconstruction was used to further evaluate the dense contact geometry preserved in the learned representation. Surface normals were predicted from the frozen tactile representation and integrated to reconstruct continuous height fields. As shown in Fig. 4, LifelongTouch preserves the principal contact geometry and local structures across different sensors and lifecycle states, further demonstrating its ability to retain spatial physical information.

5.4 Ablation Study

Tab. 4 evaluates the key design choices of LifelongTouch by addressing several questions. First, **does stronger pixel-level normal supervision improve representation transferability?** Adding

Table 4: **Ablation study on LifelongSITR.** *Aged Avg.* denotes the average over Light-drift, Wear, Stain, and Mixed lifecycle conditions.

	New			*Aged Avg.*		
Variant	Class Inter ↑	Class Intra ↑	Pose Inter ↓	Class Inter ↑	Class Intra ↑	Pose Inter ↓
LifelongTouch (Full)	**58.72**	**80.08**	**0.9202**	**57.36**	**77.58**	**0.9308**
w/o $\mathcal{L}_{\text{align}}$	56.12	71.08	0.9213	55.05	68.64	0.9497
w/ Normal $\mathcal{L}_1$	52.78	75.48	0.9688	51.70	72.64	0.9921
w/o Reference	55.60	71.21	1.0155	54.45	68.78	1.0406
Modulation Only	56.88	77.40	0.9330	55.84	74.85	0.9560
Attention Only	54.90	76.68	0.9759	53.70	73.71	1.0083
w/o Contact Residual	56.03	76.21	0.9979	54.15	73.05	1.0229

a normal $\mathcal{L}_1$ loss degrades both classification and pose estimation performance, suggesting that pixel-level normal reconstruction may overfit the representation to simulation-specific geometric targets and reduce its transferability. Second, **is the no-contact reference necessary?** Removing the reference leads to consistent performance degradation across all metrics, particularly for intra-sensor classification and pose estimation, confirming the importance of sensor-state information for disentangling contact-induced changes from sensor-specific appearance variations.

Furthermore, **how should the reference be incorporated into the network?** Using either modulation or attention alone performs consistently worse than combining both, indicating that the two mechanisms provide complementary benefits for feature conditioning and interaction modeling. **Is it still important to preserve contact features?** Removing the contact residual substantially increases pose estimation error, showing that contact features remain important for retaining local geometric information. Finally, **what is the role of representation alignment?** Removing the alignment loss causes a more pronounced drop in classification performance, while having a relatively smaller effect on pose estimation, suggesting that alignment mainly improves representation invariance and transferability across sensors and states. More details are provided in Appendix A.4.2.

6 Discussion

Our qualitative and quantitative results show that LifelongTouch learns stable and transferable tactile representations across sensors and lifecycle states, with consistent performance on object classification, contact pose estimation, force distribution estimation, and surface reconstruction. This demonstrates that jointly modeling sensor-state references and contact observations mitigates variations caused by sensor heterogeneity and long-term degradation.

LifelongTouch currently focuses on vision-based tactile sensors and common lifecycle changes such as illumination drift, gel wear, coating degradation, and contamination. Future work may extend the framework to more complex sensor geometries and other tactile modalities through shared physical representations of contact geometry, deformation, motion, and force. Another promising direction is to combine lifecycle-invariant representations with lightweight adaptation or continual learning to handle more severe and non-stationary degradation. Overall, LifelongTouch extends cross-sensor tactile representation learning toward cross-lifecycle generalization for long-term tactile perception.

7 Conclusion

We proposed LifelongTouch, a lifecycle-invariant tactile representation framework for heterogeneous tactile sensors and their evolving lifecycle states. By jointly modeling current sensor-state references and contact observations, LifelongTouch disentangles appearance variations caused by sensor-state changes from contact-induced physical information. We further developed a lifecycle-aligned simulation data engine that provides geometric supervision for representation learning across diverse sensors and degradation states. Experimental results showed that LifelongTouch generalizes effectively to unseen sensors and lifecycle states on object classification, contact pose estimation, force distribution estimation, and surface reconstruction, while maintaining stable performance without continual calibration or model updates. These results demonstrate the effectiveness of lifecycle-invariant tactile representations for long-term robotic tactile perception.

AI USE STATEMENT

In this work, we used GPT-based tools to assist with parts of the code implementation and manuscript preparation. For code development, the data processing pipeline, core model architecture, and evaluation protocols were independently designed and implemented by the authors. AI tools were used only to assist with selected function-level implementations, all of which were subsequently reviewed, tested, and validated by the authors. For manuscript preparation, all figures, core technical content, experimental design and analysis, and references were produced by the authors, while AI tools were used only for language polishing, improving clarity, and enhancing readability.

REPRODUCIBILITY STATEMENT

Detailed implementation settings, dataset construction, downstream evaluation protocols, and additional experiments are provided in the Appendix. The simulation and experimental code required to reproduce the results is packaged with the supplementary material. The complete codebase and datasets will be publicly released upon acceptance.

REFERENCES

Alexander C. Abad and Anuradha Ranasinghe. Visuotactile Sensors With Emphasis on GelSight Sensor: A Review. *IEEE Sensors Journal*, 20(14):7628–7638, 2020.

Arpit Agarwal, Timothy Man, and Wenzhen Yuan. Simulation of Vision-based Tactile Sensors using Physics based Rendering. In *ICRA*, pp. 1–7, 2021.

Roberto Calandra, Andrew Owens, Dinesh Jayaraman, Justin Lin, Wenzhen Yuan, Jitendra Malik, Edward H. Adelson, and Sergey Levine. More Than a Feeling: Learning to Grasp and Regrasp Using Vision and Touch. *IEEE Rob. Autom. Lett.*, 3(4):3300–3307, 2018.

Guanqun Cao, Jiaqi Jiang, Danushka Bollegala, and Shan Luo. Learn from Incomplete Tactile Data: Tactile Representation Learning with Masked Autoencoders. In *IROS*, pp. 10800–10805, 2023.

Xingyu Chen, Jialei Shi, Helge Wurdemann, and Thomas George Thuruthel. Vision-based Tip Force Estimation on a Soft Continuum Robot. In *ICRA*, pp. 7621–7627, 2024.

Mauro Comi, Alessio Tonioni, Jonathan Tremblay, Max Yang, Valts Blukis, Yijiong Lin, Nathan F. Lepora, and Laurence Aitchison. Snap-it, Tap-it, Splat-it: Tactile-Informed 3D Gaussian Splatting for Reconstructing Challenging Surfaces. In *3DV*, 2025.

F Richard Cottrell, Megha H Tippur, and Edward H Adelson. A Durable Vision-Based Tactile Fingertip for Robotic Manipulation. *arXiv preprint arXiv:2608.24242*, 2026.

Siyuan Dong, Wenzhen Yuan, and Edward H. Adelson. Improved GelSight Tactile Sensor for Measuring Geometry and Slip. In *IROS*, pp. 137–144, 2017.

Siyuan Dong, Daolin Ma, Elliott Donlon, and Alberto Rodriguez. Maintaining Grasps within Slipping Bounds by Monitoring Incipient Slip. In *ICRA*, pp. 3818–3824, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021.

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. In *ICLR*, 2025.

Ruoxuan Feng, Yuxuan Zhou, Siyu Mei, Dongzhan Zhou, Pengwei Wang, Shaowei Cui, Bin Fang, Guocai Yao, and Di Hu. AnyTouch 2: General Optical Tactile Representation Learning For Dynamic Tactile Perception. In *ICLR*, 2026.

Letian Fu, Gaurav Datta, Huang Huang, William Chung-Ho Panitch, Jaimyn Drake, Joseph Ortiz, Mustafa Mukadam, Mike Lambeta, Roberto Calandra, and Ken Goldberg. A Touch, Vision, and Language Dataset for Multimodal Alignment. In *ICML*, 2024.

Harsh Gupta, Yuchen Mo, Shengmiao Jin, and Wenzhen Yuan. Sensor-Invariant Tactile Representation. In *ICLR*, 2025.

Irmak Guzey, Ben Evans, Soumith Chintala, and Lerrel Pinto. Dexterity from Touch: Self-Supervised Pre-Training of Tactile Representations with Robotic Play. In *CoRL*, volume 229, pp. 3142–3166, 2023.

Erik Helmut, Luca Dziarski, Niklas Funk, Boris Belousov, and Jan Peters. Learning Force Distribution Estimation for the GelSight Mini Optical Tactile Sensor Based on Finite Element Analysis. In *IROS*, pp. 8553–8560, 2025.

Carolina Higuera, Akash Sharma, Chaithanya Krishna Bodduluri, Taosha Fan, Patrick Lancaster, Mrinal Kalakrishnan, Michael Kaess, Byron Boots, Mike Lambeta, Tingfan Wu, and Mustafa Mukadam. Sparsh: Self-supervised Touch Representations for Vision-based Tactile Sensing. In *CoRL*, volume 270, pp. 885–915, 2025.

Justin Kerr, Huang Huang, Albert Wilcox, Ryan Hoque, Jeffrey Ichnowski, Roberto Calandra, and Ken Goldberg. Self-Supervised Visuo-Tactile Pretraining to Locate and Follow Garment Features. In *RSS*, 2023.

Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian 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. *IEEE Rob. Autom. Lett.*, 5(3):3838–3845, 2020.

Shoujie Li, Tong Wu, Jianle Xu, Yan Huang, Zongwen Zhang, Hongfa Zhao, Qinghao Xu, Zihan Wang, Linqi Ye, Yang Yang, Chuqiao Lyu, Shilong Mu, Xueqian Wang, Zhaoqian Xie, Changsheng Wu, Xinge Yu, and Wenbo Ding. Biomimetic Multimodal Tactile Sensing Enables Human-Like Robotic Perception. *Nat. Sens.*, 1:52–62, 2026.

Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting Touch and Vision via Cross-Modal Prediction. In *CVPR*, pp. 10601–10610, 2019.

Dantong Niu, Zhuoyang Liu, Zekai Wang, Boning Shao, Zhao-Heng Yin, Anirudh Pai, Yuvan Sharma, Stefano Saravalle, Ruijie Zheng, Jing Wang, Ryan Punamiya, Mengda Xu, Yuqi Xie, Yunfan Jiang, Letian Fu, Konstantinos Kallidromitis, Matteo Gioia, Junyi Zhang, Jiaxin Ge, Haiwen Feng, Fabio Galasso, Wei Zhan, David M. Chan, Yutong Bai, Roei Herzig, Jiahui Lei, Fei-Fei Li, Ken Goldberg, Jitendra Malik, Pieter Abbeel, Yuke Zhu, Danfei Xu, Jim, Fan, and Trevor Darrell. T-Rex: Tactile-Reactive Dexterous Manipulation. *arXiv preprint arXiv:2606.17055*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *TMLR*, 2024. ISSN 2835-8856.

Samanta Rodriguez, Yiming Dou, William van den Bogert, Miquel Oller, Kevin So, Andrew Owens, and Nima Fazeli. Contrastive Touch-to-Touch Pretraining. In *ICRA*, pp. 5857–5863, 2025.

Zilin Si, Gu Zhang, Qingwei Ben, Branden Romero, Zhou Xian, Chao Liu, and Chuang Gan. DIFFTACTILE: A Physics-based Differentiable Tactile Simulator for Contact-rich Robotic Manipulation. In *ICLR*, 2024.

Sudharshan Suresh, Haozhi Qi, Tingfan Wu, Taosha Fan, Luis Pineda, Mike Lambeta, Jitendra Malik, Mrinal Kalakrishnan, Roberto Calandra, Michael Kaess, Joseph Ortiz, and Mustafa Mukadam. NeuralFeels with Neural Fields: Visuotactile Perception for In-hand Manipulation. *Science Robotics*, 9(96):eadl0628, 2024.

Ian H. Taylor, Siyuan Dong, and Alberto Rodriguez. GelSlim 3.0: High-Resolution Measurement of Shape, Force and Slip in a Compact Tactile-Sensing Finger. In *ICRA*, pp. 10781–10787, 2022.

Shaoxiong Wang, Yu She, Branden Romero, and Edward Adelson. GelSight Wedge: Measuring High-Resolution 3D Contact Geometry with a Compact Robot Finger. In *ICRA*, pp. 6468–6475, 2021.

Zhengtong Xu, Raghava Uppuluri, Xinwei Zhang, Cael Fitch, Philip Glen Crandall, Wan Shou, Dongyi Wang, and Yu She. UniT: Data Efficient Tactile Representation With Generalization to Unseen Objects. *IEEE Rob. Autom. Lett.*, 10(6):5481–5488, 2025.

Han Xue, Jieji Ren, Wendi Chen, Gu Zhang, Yuan Fang, Guoying Gu, Huazhe Xu, and Cewu Lu. Reactive Diffusion Policy: Slow-Fast Visual-Tactile Policy Learning for Contact-Rich Manipulation. In *RSS*, 2025.

Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and Go: Learning from Human-Collected Vision and Touch . In *NeurIPS*, 2022.

Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, and Alex Wong. Binding Touch to Everything: Learning Unified Multimodal Tactile Representations. In *CVPR*, pp. 26330–26343, 2024.

Kelin Yu, Yunhai Han, Qixian Wang, Vaibhav Saxena, Danfei Xu, and Ye Zhao. MimicTouch: Leveraging Multi-modal Human Tactile Demonstrations for Contact-rich Manipulation. In *CoRL*, volume 270, pp. 4844–4865, 2025.

Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. GelSight: High-Resolution Robot Tactile Sensors for Estimating Geometry and Force. *Sensors*, 17(12):2762, 2017a.

Wenzhen Yuan, Shaoxiong Wang, Siyuan Dong, and Edward Adelson. Connecting Look and Feel: Associating the Visual and Tactile Properties of Physical Materials. In *CVPR*, pp. 4494–4502, 2017b.

Wenzhen Yuan, Yuchen Mo, Shaoxiong Wang, and Edward H. Adelson. Active Clothing Material Perception Using Tactile Sensing and Deep Learning. In *ICRA*, pp. 4842–4849, 2018.

Guanlan Zhang, Yipai Du, Hongyu Yu, and Michael Yu Wang. DelTact: A Vision-Based Tactile Sensor Using a Dense Color Pattern. *IEEE Rob. Autom. Lett.*, 7(4):10778–10785, 2022.

Yan Zhang, Zheng WANG, Pengpeng Zeng, Xing Xu, Jingkuan Song, and Heng Tao Shen. Cross-Tactile Sensor Representation Learning. In *ICML*, 2026.

Jialiang Zhao, Yuxiang Ma, Lirui Wang, and Edward Adelson. Transferable Tactile Transformers for Representation Learning Across Diverse Sensors and Tasks. In *CoRL*, volume 270, pp. 3766–3779, 2025.

A APPENDIX

CONTENTS

A.1 Implementation Details 13
A.1.1 Pre-training of LifelongTouch 13
A.1.2 Baselines 14
A.2 Dataset Details 16
A.2.1 Lifecycle-aligned Data Engine 16
A.2.2 Real-World Lifetime Benchmark 16
A.3 Downstream Evaluation Details 20
A.3.1 Object Classification 20
A.3.2 Contact Pose Estimation 20
A.3.3 Dense Force Distribution Estimation 21
A.3.4 Surface Reconstruction 22
A.4 Additional Experiments 23
A.4.1 Unified Linear-Probe Evaluation of Frozen Tactile Encoders 23
A.4.2 Detailed Ablation Analysis 24
A.4.3 Qualitative Results 25

A.1 IMPLEMENTATION DETAILS

A.1.1 PRE-TRAINING OF LIFELONGTOUCH

All tactile RGB images are resized to 224×224. Each training sample consists of a no-contact reference image and a contact image acquired from the same sensor and lifecycle state. Both images are normalized to $[-1, 1]$ and subsequently transformed using standard ImageNet normalization before being fed into the DINOv2 backbone.

During pre-training, identical photometric augmentations are applied to the paired reference and contact images to avoid introducing artificial discrepancies between them. Specifically, the intensity gain is uniformly sampled from $[0.85, 1.15]$, and the brightness offset from $[-0.04, 0.04]$. Gaussian blur is additionally applied with a probability of 0.3, with the blur radius sampled from $[0.1, 1.0]$. Ground-truth surface normals and depth maps are resized to 224×224, and the surface normals are normalized to unit vectors. Reference conditioning is randomly disabled with a probability of 0.1 during training to reduce over-reliance on the reference observation.

The pre-training objective combines dense geometric supervision with lifecycle-invariant feature alignment:

$$\mathcal{L} = \mathcal{L}_{\text{normal}} + 0.1\mathcal{L}_{\text{depth}} + 0.2\mathcal{L}_{\text{align}}, \tag{12}$$

where the surface-normal objective consists of a cosine loss and a component-wise ℓ_1 loss:

$$\mathcal{L}_{\text{normal}} = \mathcal{L}_{\cos} + \lambda_{\text{n-L1}}\mathcal{L}_{\text{n-L1}}. \tag{13}$$

In the full LifelongTouch model, we set $\lambda_{\text{n-L1}} = 0$. To investigate whether improving the accuracy of surface-normal prediction leads to more generalizable tactile representations, we additionally set $\lambda_{\text{n-L1}} = 1$ in the ablation study reported in Tab. 4.

Let $\hat{\mathbf{n}}_i \in \mathbb{R}^3$ and $\mathbf{n}_i \in \mathbb{R}^3$ denote the predicted and ground-truth surface normals at pixel i, respectively. Normal supervision is applied only to the valid pixel set:

$$\Omega_n = \{i \mid m_i^n = 1,\ \|\mathbf{n}_i\|_2 > 0.5\}, \tag{14}$$

where m_i^n denotes the provided normal-validity mask. The cosine loss is defined as:

$$\mathcal{L}_{\cos} = 1 - \frac{1}{|\Omega_n|} \sum_{i \in \Omega_n} \frac{\hat{\mathbf{n}}_i^\top \mathbf{n}_i}{|\hat{\mathbf{n}}_i|_2 |\mathbf{n}_i|_2}, \tag{15}$$

while the component-wise normal ℓ_1 loss is:

$$\mathcal{L}_{\text{n-L1}} = \frac{1}{3|\Omega_n|} \sum_{i \in \Omega_n} \sum_{c \in x,y,z} |\hat{n}_{i,c} - n_{i,c}| \,. \tag{16}$$

Similarly, given the predicted depth $\hat{d}_i$ and ground-truth depth d_i, the depth loss is computed as a masked pixel-wise ℓ_1 loss over the valid depth region Ω_d:

$$\mathcal{L}_{\text{depth}} = \frac{1}{|\Omega_d|} \sum_{i \in \Omega_d} \left| \hat{d}_i - d_i \right| . \tag{17}$$

For representation alignment, observations corresponding to the same physical contact are grouped according to their contact identity and frame index. For each physical contact, four observations from different sensor or lifecycle configurations are sampled. Their normalized projected token features are averaged to construct a stop-gradient group center, and $\mathcal{L}_{\text{align}}$ is applied densely over all spatial token locations.

For dataset partitioning, sensors 0–79 and lifecycle trajectories 0–5 are used for pre-training. Sensors $80-89$ from trajectory 6 are used for validation, while sensors $90-99$ and trajectory 7 are reserved for final testing and are not involved in either parameter optimization or checkpoint selection.

Each mini-batch contains four physical contacts, with four sensor/lifecycle views sampled for each contact, resulting in a global batch of 16 reference-contact pairs. We train the model for 30 epochs with 1,500 mini-batches per epoch using DistributedDataParallel on two NVIDIA RTX 3090 GPUs. AdamW is used for optimization. The DINOv2 backbone is optimized with a learning rate of 2×10^{-5}, while the newly initialized modules use a learning rate of 2×10^{-4}. Weight decay is set to 0.05, and the gradient norm is clipped to 1.0.

A.1.2 Baselines

We compare LifelongTouch with representative tactile representation learning methods covering general-purpose, cross-sensor, and geometry-oriented tactile representations. Unless otherwise specified, we use the official released checkpoints and freeze the pre-trained representation networks during downstream evaluation. We retain the original input formulation, normalization scheme, and spatial feature organization of each baseline whenever possible, while adapting only the downstream task heads to the corresponding evaluation task.

SITR. We use the officially released SITR-B18 checkpoint under the $K = 0$ setting. The calibration-specific branch is removed, and no contact calibration tokens are provided during downstream evaluation. We retain the representation and task interfaces defined by the official implementation. SITR serves as the primary geometry-oriented cross-sensor baseline because it learns sensor-invariant tactile representations using simulation-based geometric supervision.

CTSRL. We use the official ResNet-50-based tactile encoder together with its sensor modulation mechanism. Since the real LifetimeSITR evaluation sensors are not associated with sensor-specific embeddings provided during CTSRL pre-training, we use the generic sensor embedding for evaluation. The released representation network is frozen for all downstream tasks.

T3. We adopt the released T3-Medium model. Following the official architecture, the three blocks of the sensor-specific Mini stem and nine blocks of the shared Transformer trunk are combined to form the tactile encoder. Spatial patch tokens from the shared trunk are retained for downstream prediction.

Sparsh. We evaluate the released Sparsh DINOv2-Base tactile representation. Because the LifetimeSITR object-classification benchmark consists of static single-frame tactile observations whereas the released Sparsh interface expects paired tactile views, we duplicate the current frame to construct the required input without introducing additional observations.

AnyTouch. We use the officially released AnyTouch tactile encoder based on CLIP ViT-L/14. A fixed generic sensor token is used for all LifetimeSITR sensors. Since the released model expects a short temporal tactile clip, each static LifetimeSITR observation is repeated across three frames to match the original input format.

CTTP. We use the released ResNet-50 encoder and discard the contrastive projection head after pre-training. Because CTTP adopts its own tactile-image normalization convention, the LifetimeSITR input normalization is first inverted and the resulting RGB image is subsequently normalized using the statistics adopted by the official CTTP implementation.

ImageNet-pretrained ViT. To quantify the contribution of tactile-specific pre-training, we additionally include an ImageNet-pretrained ViT-B/16 as a generic visual representation baseline. Its pre-trained backbone is frozen during downstream evaluation in the same manner as the tactile-specific representation models.

For methods requiring a no-contact reference or paired input, the reference is restricted to information observable at inference time. In particular, for real LifetimeSITR experiments, only the no-contact background image is exposed as the reference, and no additional contact calibration samples are used. Methods that natively operate on a single tactile observation receive only the contact image.

A.2 Dataset Details

A.2.1 Lifecycle-aligned Data Engine

Degradation Type	Parameters	Variation
LED Illumination	intensity, color, direction, failed LED	severity $\rho \in [0, 1]$
Camera	exposure, gain, white balance, blur	severity $\rho \in [0, 1]$
Gel Pad	stiffness, thickness, hysteresis	severity $\rho \in [0, 1]$
Coating	albedo, roughness, missing regions	area / severity
Contamination	spots, film, particles, occlusion	area / opacity
Damage	scratches, cuts, dead regions	width / area / depth

Table 5: Lifecycle degradation factors modeled in our simulation engine. The six categories capture variations in the optical, imaging, material, and surface conditions of VBTSs during long-term operation.

Vision-based tactile sensors (VBTSs) convert physical contact into image observations through the joint effects of illumination, imaging, elastomer deformation, and surface appearance. During long-term use, changes in these components can continuously shift the tactile observation distribution even when the underlying physical contact remains unchanged. Based on the sensing and imaging mechanisms of VBTSs, we categorize lifecycle variations into six types: *LED illumination*, *camera*, *gel pad*, *coating*, *contamination*, and *physical damage*, as summarized in Tab. 5. These variations include both gradual state drift, such as illumination attenuation, camera response changes, and material aging, and spatial degradation, such as coating wear, contamination, scratches, and local damage.

To model these variations, we develop a lifecycle-aligned data engine that independently parameterizes individual degradation factors and composes them into long-term lifecycle trajectories. Virtual VBTSs with diverse designs are instantiated by randomizing sensor-specific parameters, including camera field of view, sensing area, elastomer surface properties, and illumination configurations. For each sensor, multiple independent lifecycle trajectories are generated, along which illumination, camera, elastomer, coating, contamination, and physical damage progressively evolve at different severities. We further introduce stochastic events, including LED failure and recovery, camera displacement, contamination and cleaning, and elastomer pad replacement, to capture both gradual and non-monotonic state changes encountered during real-world operation.

In addition, physical interactions are sampled independently of sensor design and lifecycle state, with variations in object size, contact position, indentation depth, and four motion types: *sliding*, *rotation*, *pressing*, and *combined motion*. For the same physical interaction, varying the sensor design or lifecycle state produces aligned tactile observations that share the same contact geometry and motion while exhibiting different appearances and degradation patterns. These explicit correspondences enable the model to distinguish sensor-state-induced observation changes from stable physical contact information, providing supervision for learning representations invariant across sensors and lifecycle states.

Detailed configurations are provided in Tab. 6, with additional rendering examples shown in Fig. 5. In total, we instantiate 100 virtual VBTSs, each with 8 lifecycle trajectories of 16 states, yielding 12,800 sensor-lifecycle states. Each state contains 4 anchor contacts and 8 interaction sequences of 8 frames, resulting in 870,400 tactile observations at 224×224 resolution, including 51,200 anchor-contact observations and 819,200 interaction frames. All observations are accompanied by depth and surface-normal ground truth. The resulting dataset spans diverse sensor designs, lifecycle states, and physical interactions, providing large-scale aligned data for geometry-supervised tactile representation learning across sensors and their lifecycles.

A.2.2 Real-World Lifetime Benchmark

LifetimeSITR Dataset. The SITR dataset (Gupta et al., 2025) contains tactile observations collected using multiple real tactile sensors, together with object-category and contact-pose annotations, but does not capture lifecycle variations that may arise within the same sensor during long-term use. To address this limitation, we apply controlled pseudo-lifecycle degradation to the real tac-

Table 6: Configuration of our lifecycle-aligned data engine. Sensor identity parameters characterize cross-sensor diversity, while lifecycle parameters model long-term sensor-state variations.

Category	Parameter	Range / Setting
Dataset	Sensors	100
	Trajectories / sensor	8
	Lifecycle states	16
	Anchor contacts / state	4
	Stream sequences / state	8×8 frames
Contact	Object size	$[10, 50]$ mm
	Contact position (x, y)	$[-7, 7]$ mm
	Indentation depth	$[0.6, 1.8]$ mm
	Press displacement	$[0.3, 1.0]$ mm
	Sliding displacement	$[-3.5, 3.5]$ mm
	Rotation	$[-0.50, 0.50]$ rad
Sensor	Camera FOV	$[20, 60]^\circ$
	Sensing half-length	$[7.5, 12.5]$ mm
	Smoothing iterations	$[15, 100]$
	Surface roughness	$[0.4, 1.0]$
	Light scale	$[0.25, 1.0]$
	Light strength (R/G/B)	40–100/80–120/70–100
Lifecycle	LED decay	$[0.05, 0.45]$
	Color drift	$[-0.15, 0.15]$
	Roughness increment	$[0, 0.28]$
	Smoothness ratio	$[-0.25, 0.30]$
	FOV ratio	$[-0.03, 0.03]$
	Camera shift	$[-0.035, 0.035]$
	Camera rotation	$[-0.035, 0.035]$ rad
	Exposure drift	$[-0.65, 0.65]$ EV
	Blur / noise std.	$[0, 1.5]$ px / $[0, 0.035]$
Artifacts	Contamination coverage	$[0, 0.12]$
	Scratches	$[4, 28]$
	Random-walk scale	0.018
Rendering	Resolution	224×224
	Renderer	Cycles (64 samples)
	Geometry labels	Depth + surface normal

tile observations while preserving the underlying physical contact states and task labels, simulating sensor-state changes caused by factors such as optical drift, surface wear, and contamination. Unlike frame-wise random image augmentation, these degradations are designed to be temporally continuous and spatially consistent: sensor states evolve smoothly over time, while localized artifacts such as wear and stains persist across consecutive observations. This construction more closely reflects the lifecycle variations encountered by tactile sensors during long-term operation.

Specifically, we model a real tactile observation at lifecycle state t as:

$$I_t = \mathcal{R}(S_t; \theta_t) + A_t + \epsilon_t,$$

where S_t denotes the underlying physical contact state, θ_t represents the slowly varying sensor state, including global changes in illumination, imaging response, and surface properties, A_t denotes fixed or semi-persistent spatial artifacts caused by contamination, scratches, and local wear, and ϵ_t represents stochastic imaging noise. The sensor state θ_t evolves smoothly over the lifecycle, while the spatial artifacts A_t remain spatially consistent over consecutive lifecycle states. This construction avoids the temporally incoherent degradation patterns introduced by independently augmenting individual frames.

Based on the above model, we construct five lifecycle evaluation conditions. All degradations are applied only to the sensor observations while preserving the original contact states and downstream task labels, enabling direct and controlled comparison across different lifecycle conditions:

- **New:** The original real tactile observations are retained without additional degradation, serving as the undegraded baseline.
- **Light-drift:** Simulates global appearance drift caused by variations in LED intensity, color, and imaging response, evaluating robustness to long-term changes in optical conditions.
- **Wear:** Simulates wear and local degradation of the elastomer and surface coating caused by long-term use, evaluating robustness to changes in the tactile surface condition.
- **Stain:** Simulates contamination and stains accumulated during long-term use by introducing spatially persistent artifacts with varying locations and coverage.
- **Mixed:** Combines multiple degradation factors, including illumination variation, surface wear, and contamination, to simulate more complex long-term sensor conditions.

In addition, based on the seven real tactile sensors included in SITR Dataset, we define three evaluation settings according to the degree of sensor variation. GelSight Mini 1–4 share the same optical design but use different gel pads, resulting in variations in brightness and color due to manufacturing differences, surface coatings, and illumination conditions. In contrast, GelSight Mini, GelSight Wedge, GelSight Hex, and DIGIT employ substantially different camera, illumination, and optical configurations, leading to larger domain shifts in their tactile observations. Accordingly, we consider the following three settings:

- **No Transfer (No Trans.):** Training and testing are performed on the same tactile sensor, providing a baseline without sensor transfer.
- **Intra-Sensor Transfer (Intra):** Transfer is performed among GelSight Mini 1–4, which share the same optical design but use different gel pads. This setting evaluates generalization to variations caused by manufacturing differences, gel pads, and surface coatings.
- **Inter-Sensor Transfer (Inter):** Transfer is performed across sensors with substantially different optical designs, including GelSight Mini 1, GelSight Wedge, GelSight Hex, and DIGIT. This setting evaluates cross-sensor generalization under larger variations in sensor design and tactile appearance.

Fig. 6 presents representative samples from the constructed LifetimeSITR dataset. From the lifecycle perspective, different lifecycle states exhibit substantial observation shifts while preserving the underlying physical contact states and task labels, including illumination and color drift, surface wear, localized contamination, and combinations of multiple degradation factors. Meanwhile, different tactile sensors can produce substantially different observations for the same or similar physical contacts due to variations in optical design, gel pads, surface coatings, and imaging characteristics. By covering both cross-lifecycle and cross-sensor variations, LifetimeSITR enables systematic and controlled evaluation of model generalization to sensor-state and sensor-domain shifts without requiring long-term collection of physically degraded sensor data.

LifetimeFEATS Dataset. The original FEATS dataset (Helmut et al., 2025) provides tactile observations paired with dense force-distribution ground truth, but does not explicitly provide a non-contact reference for each contact sample. To construct the reference–contact pairs required by LifelongTouch, we first derive contact masks from the force-distribution labels and identify non-contact observations whose force labels are zero. For each contact sample, we then perform image-similarity matching over the non-contact region within the same sensor configuration and retrieve the non-contact observation that best matches the current sensor state as its reference. This procedure reorganizes the original FEATS data into $\langle$reference, contact, force distribution$\rangle$ triplets while preserving the original sensor and gel-pad information.

Based on the sensor identities, gel-pad conditions, and object configurations in FEATS, we construct LifetimeFEATS to evaluate the generalization of force-distribution estimation across sensors, lifecycle states, and object-contact conditions. As illustrated in Fig. 7, we show representative contact observations and their corresponding three-axis force-distribution labels under different evaluation settings. LifetimeFEATS contains five tactile sensors and seven gel conditions and is divided into four evaluation settings:

- **No Transfer (No Trans.):** Test samples use sensor and gel conditions similar to those observed during training, providing a baseline under similar sensor states.

- **Unseen Sensor with New Gel (US-New):** Test samples are collected using unseen sensors equipped with new gel pads, evaluating cross-sensor generalization.
- **Unseen Sensor with Aged Gel (US-Aged):** Test samples are collected using unseen sensors equipped with aged gel pads, evaluating generalization under the joint effects of sensor variation and lifecycle changes.
- **Unseen Object Press (UO-Press):** Test samples are generated by objects unseen during training, evaluating generalization to novel objects and contact conditions.

Rather than predicting only a global force vector, we focus on dense local force-distribution estimation, which preserves fine-grained spatial interaction information within the contact region and provides a richer description of local contact states. LifetimeFEATS therefore evaluates whether the learned tactile representation can preserve transferable contact information under variations in sensor identity, gel lifecycle state, and object-contact conditions.

A.3 Downstream Evaluation Details

Unless otherwise specified, all pre-trained tactile encoders are frozen during downstream evaluation, and only the task-specific prediction heads are optimized. We preserve the native feature organization and downstream interface of each representation whenever possible, such that the evaluation reflects the practical transferability of each pre-trained model rather than forcing heterogeneous representations into an identical decoder design.

A.3.1 Object Classification

We evaluate object recognition on the real LifetimeSITR dataset using the 16 official object categories. For each source sensor, an independent classifier is trained using only the corresponding New training split, while the pre-trained tactile encoder remains frozen.

- **SITR.** We retain the official downstream formulation that jointly exploits the predicted surface-normal map and the 128-dimensional contrastive representation. The normal map is processed by an ImageNet-pre-trained ResNet-18 to obtain a 512-dimensional feature. The contrastive feature is projected using a linear layer, and the two branches are fused for the final 16-class prediction.
- **CTSRL.** The 2048-dimensional global tactile feature is processed by an MLP classifier, $2048 \to 256 \to 128 \to 16$, with LayerNorm, ReLU activations, and dropout with probability 0.2.
- **T3-Medium.** The CLS token is discarded, and the 14×14 spatial patch tokens are reshaped into a $768 \times 14 \times 14$ feature map. The resulting feature map is processed by a ResNet-18 adapter initialized without ImageNet pre-training. Its first convolution uses a 3×3 kernel with stride 1, the max-pooling layer is removed, and the final fully connected layer outputs 16 classes.
- **ImageNet-pretrained ViT-B/16.** The global encoder representation is fed into a single linear classifier.
- **Sparsh.** The spatial patch tokens are mean-pooled and processed by a single linear classification layer.
- **AnyTouch.** Auxiliary CLS or sensor tokens are removed before mean pooling, and the resulting global tactile representation is processed by a linear classifier.
- **CTTP.** The 2048-dimensional feature produced by the ResNet-50 encoder is directly processed by $\mathrm{Linear}(2048, 16)$.
- **LifelongTouch.** Both the tactile encoder and the pre-trained normal decoder are frozen during downstream evaluation. The predicted $3 \times 224 \times 224$ normal map is processed by an ImageNet-pre-trained ResNet-18, with the final fully connected layer replaced by a 16-class classifier.

All classification heads are optimized using cross-entropy loss.

A.3.2 Contact Pose Estimation

Contact pose estimation is evaluated on the real LifetimeSITR pose benchmark. Given two tactile observations, the task predicts their 3D relative translation. For each source sensor, the downstream regressor is trained exclusively on New data and is subsequently frozen during all transfer evaluations.

- **SITR.** The predicted normal maps from the two observations are concatenated along the channel dimension to form a $6 \times 224 \times 224$ input. The paired normal maps are processed by an ImageNet-pre-trained ResNet-18. Its first convolution is replaced with a six-channel input layer, and the final fully connected layer is replaced with $\mathrm{Linear}(512, 3)$.
- **CTSRL.** The ResNet-50 layer-3 feature map is extracted independently from each frame. Each 1024-channel feature map is first projected to three channels using a 1×1 convolution followed by BatchNorm and ReLU. The two projected feature maps are concatenated into a six-channel representation and processed by a ResNet-18 regressor. Its first convolution

uses a 3×3 kernel with stride 1, the max-pooling layer is removed, and the final layer outputs three values.

- **T3-Medium.** Corresponding spatial patch features from the two observations are concatenated along the feature dimension, yielding a $1536 \times 14 \times 14$ feature map. This representation is processed by a ResNet-18 adapter initialized without ImageNet pre-training, with the final fully connected layer predicting the three-dimensional relative displacement.
- **ImageNet-pretrained ViT-B/16.** The two RGB images are concatenated along the input channel dimension and jointly processed by the encoder, followed by a linear three-dimensional regression head.
- **Sparsh and AnyTouch.** The two tactile observations are independently encoded and mean-pooled into z_0 and z_1. Their pairwise representation is constructed as $[z_0, z_1, z_1 - z_0]$, and processed by$\mathrm{Linear}(3D, D) \rightarrow \mathrm{GELU} \rightarrow \mathrm{Linear}(D, 3)$, where D denotes the feature dimension of the corresponding encoder.
- **CTTP.** Two 2048-dimensional global features are independently extracted and combined as $[z_0, z_1, z_1 - z_0] \in \mathbb{R}^{6144}$. The resulting representation is processed by $6144 \rightarrow 2048 \rightarrow 3$ with GELU activation.
- **LifelongTouch.** The pre-trained tactile encoder and normal decoder remain frozen. The two predicted surface-normal maps are concatenated along the channel dimension to form a six-channel input and processed by an ImageNet-pre-trained ResNet-18 regressor, whose final fully connected layer outputs the three-dimensional relative displacement.

All relative pose regressors are optimized using mean squared error.

A.3.3 Dense Force Distribution Estimation

We evaluate dense physical information on the LifetimeFEATS force-distribution benchmark. All pre-trained tactile encoders remain frozen, and only the force prediction head is optimized.

For token-based encoders, including ImageNet ViT-B/16, T3, SITR, Sparsh, AnyTouch, and LifelongTouch, we use a common token-based force decoder. Spatial tokens are first normalized using LayerNorm and reshaped into their two-dimensional token grid. They are then processed by,

$$\mathrm{Conv}_{3\times3}(D, 256) \rightarrow \mathrm{GELU} \rightarrow \mathrm{Conv}_{3\times3}(256, 128) \rightarrow \mathrm{GELU} \rightarrow \mathrm{Conv}_{1\times1}(128, 3).$$

The resulting three-channel force map is bilinearly interpolated to 24×32.

For CTSRL and CTTP, the convolutional feature maps are processed by a lightweight convolutional decoder with channel dimensions $C \rightarrow 256 \rightarrow 256 \rightarrow 128 \rightarrow 3$, using 1×1, 3×3, 3×3, and 1×1 convolutions, respectively, with GELU activations between layers. The output is bilinearly interpolated to 24×32.

The official FEATS baseline is evaluated using its released U-Net rather than a frozen representation head. We also train the same U-Net architecture from scratch as an additional reference.

The x- and y-axis force components are normalized to $[-1, 1]$ according to the training-set ranges, while the z-axis force is normalized to $[0, 1]$. The force decoder is optimized using mean squared error on the normalized force field.

Evaluation is performed under four settings: in-distribution testing, unseen sensor with a new gel pad, unseen sensor with an aged gel pad, and unseen indenter. We report pixel-wise force-field MAE and RMSE together with integrated total-force MAE in Newtons. Following the FEATS evaluation protocol, total-force metrics are computed only for samples satisfying,

$$f_z > -40\,\mathrm{N}.$$

For methods that natively require a reference observation, including CTTP, Sparsh, SITR, CTSRL and LifelongTouch, the corresponding reference image is provided according to the same paired-input protocol used throughout our experiments. Single-frame baselines receive only the contact observation.

A.3.4 SURFACE RECONSTRUCTION

To qualitatively examine the dense geometric information preserved by LifelongTouch, we convert the predicted surface-normal maps into continuous height fields for visualization. Given a unit normal field $\mathbf{n}(x, y) = (n_x, n_y, n_z)$, the corresponding surface gradients are computed as,

$$p(x, y) = \frac{\partial z}{\partial x} = -\frac{n_x}{n_z}, \qquad q(x, y) = \frac{\partial z}{\partial y} = -\frac{n_y}{n_z},$$

where n_z is clamped to a small positive value for numerical stability. Mild Gaussian smoothing is applied to the normal field before gradient computation to reduce local noise.

Because small normal errors over nearly planar regions can accumulate into noticeable low-frequency drift after integration, we apply a lightweight bias correction to the estimated gradient field. Approximately flat background regions are identified according to the gradient magnitude,

$$g(x, y) = \sqrt{p(x, y)^2 + q(x, y)^2},$$

and their median gradients are used to estimate and remove the global gradient bias. Large gradient outliers are additionally clipped using a high-percentile threshold.

The corrected gradient field is integrated using the Frankot–Chellappa algorithm, which projects the estimated gradients onto an integrable surface in the frequency domain. Given the Fourier transforms P and Q of the two gradient components, the surface spectrum is computed as,

$$Z(\omega_x, \omega_y) = \frac{-j\omega_x P(\omega_x, \omega_y) - j\omega_y Q(\omega_x, \omega_y)}{\omega_x^2 + \omega_y^2},$$

with the DC component set to zero. The height field is then recovered by inverse Fourier transformation. To further suppress residual global tilt, we fit a plane to the estimated background region and subtract it from the reconstructed surface.

For qualitative visualization, the recovered height field is converted into a dense 3D mesh using the original image-plane coordinates. The predicted normal map is retained for surface shading rather than recomputing normals from the reconstructed mesh, allowing fine local geometric variations encoded by the tactile representation to remain visible. All samples are rendered using the same camera and lighting configuration for consistent comparison.

This normal-to-surface conversion is used only for qualitative visualization and does not participate in representation learning or quantitative evaluation.

Table 7: Object classification accuracy (%, ↑) of common frozen-encoder linear probes on LifetimeSITR. All encoders are frozen and final spatial features are mean-pooled without trainable parameters. Values are mean ± standard deviation over three random seeds.

	New			Light-drift			Wear			Stain			Mixed			Aged Avg.		
Method	NoT	Inter	Intra	NoT	Inter	Intra	NoT	Inter	Intra	NoT	Inter	Intra	NoT	Inter	Intra	NoT	Inter	Intra
ViT-Base ImageNet	97.85±0.07	44.31±0.81	60.42±0.32	90.35±0.25	41.72±0.45	54.93±1.16	96.84±0.13	45.49±0.59	59.13±0.33	97.92±0.02	44.42±1.00	60.71±0.35	91.44±0.29	39.46±0.88	58.25±1.37	94.14	42.77	58.26
ViT-Large ImageNet	99.05±0.02	68.49±0.41	77.33±0.47	97.33±0.13	64.08±1.00	73.32±0.79	99.03±0.05	71.60±0.53	76.49±0.31	99.11±0.02	68.57±0.69	77.58±0.45	97.71±0.09	63.39±0.93	74.19±0.37	98.30	66.91	75.40
T3-Medium	56.79±0.24	11.28±0.15	13.61±0.22	45.32±0.26	10.57±0.01	12.88±0.06	54.68±0.18	11.42±0.10	13.79±0.27	56.50±0.26	11.27±0.11	13.48±0.18	51.70±0.49	10.65±0.12	13.29±0.20	52.05	10.98	13.36
Sparsh	71.17±0.07	7.90±0.22	9.63±0.22	55.09±0.20	8.05±0.25	9.26±0.15	69.92±0.26	8.36±0.28	9.63±0.16	71.37±0.15	8.03±0.23	9.61±0.23	60.55±0.26	8.09±0.20	9.85±0.23	64.23	8.13	9.59
AnyTouch	79.30±0.20	11.75±0.13	20.99±0.16	33.53±0.63	12.16±0.34	13.76±0.26	62.01±0.85	11.74±0.20	19.27±0.51	78.87±0.17	11.73±0.08	20.98±0.26	64.41±1.14	12.05±0.16	18.97±0.16	59.71	11.92	18.25
CTTP	49.61±0.32	13.78±0.02	13.94±0.25	41.82±0.18	13.40±0.03	13.52±0.16	46.45±0.26	13.77±0.14	13.97±0.18	48.77±0.39	13.82±0.04	13.92±0.25	43.59±0.18	13.73±0.13	14.30±0.15	45.16	13.68	13.93
CTSRL	**91.27**±0.21	25.55±0.09	34.65±0.25	43.08±1.14	17.35±0.24	15.79±0.17	71.82±0.96	25.93±0.17	30.31±0.31	**90.95**±0.29	25.36±0.15	34.39±0.21	65.44±1.07	24.12±0.42	26.44±0.16	67.82	23.19	26.73
SITR	61.26±0.92	18.87±0.71	17.35±0.07	37.58±0.47	17.80±0.83	12.94±0.33	55.12±0.58	17.72±0.63	16.12±0.34	60.49±0.84	18.83±0.68	17.36±0.06	55.51±0.52	17.94±0.65	16.12±0.17	52.18	18.07	15.64
LifelongTouch	79.85±0.56	**33.21**±0.18	**45.95**±0.28	**78.27**±0.23	**34.40**±0.27	**43.03**±0.12	**77.59**±0.41	**30.88**±0.17	**45.02**±0.38	76.92±0.83	**33.47**±0.14	**44.43**±0.35	**75.84**±0.60	**32.98**±0.14	**44.50**±0.29	**77.16**	**32.93**	**44.25**

Table 8: Contact-pose estimation RMSE (mm, ↓) of common frozen-encoder linear probes on LifetimeSITR. All encoders are frozen and the pose head is trained only on New state. Values are mean ± standard deviation over three random seeds.

	New		Light-drift		Wear		Stain		Mixed		Aged Avg.	
Method	NoT	Inter	NoT	Inter	NoT	Inter	NoT	Inter	NoT	Inter	NoT	Inter
CTSRL	0.6717±0.0019	1.7400±0.0095	1.2266±0.0276	1.7489±0.0332	0.8245±0.0027	1.7450±0.0062	0.6836±0.0018	1.7386±0.0095	0.8849±0.0016	1.7168±0.0053	0.9049	1.7373
T3-Medium	1.4706±0.0003	1.5956±0.0003	1.4892±0.0003	1.5992±0.0004	1.4696±0.0003	1.5937±0.0003	1.4715±0.0004	1.5954±0.0003	1.4756±0.0003	1.5956±0.0003	1.4765	1.5960
ViT-Base ImageNet	0.7614±0.0003	1.7838±0.0105	0.9775±0.0090	1.7581±0.0082	0.8330±0.0006	1.7671±0.0153	0.7756±0.0003	1.7777±0.0114	0.9725±0.0044	1.7635±0.0036	0.8897	1.7666
ViT-Large ImageNet	0.6534±0.0009	1.7052±0.0012	0.8395±0.0041	1.7111±0.0044	0.7295±0.0092	1.7318±0.0068	0.6632±0.0015	1.7057±0.0023	0.8698±0.0073	1.7355±0.0062	0.7755	1.7210
Sparsh DINOv2-Base	0.8000±0.0044	1.5360±0.0067	0.8401±0.0036	1.5458±0.0072	0.7816±0.0043	1.5281±0.0062	0.7966±0.0043	1.5307±0.0064	0.7931±0.0042	1.5312±0.0066	0.8029	1.5340
AnyTouch	0.7254±0.0016	1.6360±0.0182	1.0610±0.0080	1.6483±0.0165	0.7121±0.0030	1.6298±0.0191	0.7258±0.0018	1.6315±0.0186	0.8164±0.0036	1.6475±0.0165	0.8288	1.6393
CTTP	0.6052±0.0009	1.3546±0.0019	0.6428±0.0017	1.3317±0.0037	0.6467±0.0009	1.3806±0.0008	0.6343±0.0010	1.3703±0.0022	0.6556±0.0006	1.3676±0.0013	0.6449	1.3626
SITR	1.1994±0.0043	1.4392±0.0018	1.3489±0.0032	1.4861±0.0013	1.1833±0.0047	1.4340±0.0019	1.1984±0.0044	1.4340±0.0019	1.1531±0.0046	1.4179±0.0020	1.2209	1.4430
LifelongTouch	0.6455±0.0019	**0.8905±0.0009**	0.7547±0.0024	**0.9820±0.0027**	0.6710±0.0018	**0.9021±0.0016**	0.6930±0.0018	**0.9124±0.0010**	0.8044±0.0034	**0.9941±0.0030**	0.7308	**0.9477**

A.4 Additional Experiments

A.4.1 Unified Linear-Probe Evaluation of Frozen Tactile Encoders

We observe that the original downstream evaluation protocols of different methods vary substantially in decoder design and supervision, and therefore do not necessarily provide a direct comparison of encoder representation quality. For example, methods such as SITR and CTSRL explicitly learn surface normals during pretraining and subsequently use the predicted normals together with a ResNet decoder for object classification and pose estimation. Since these downstream tasks are themselves highly dependent on contact geometry, such task-specific decoding pipelines may introduce additional geometric priors and decoder capacity, making it difficult to isolate the intrinsic cross-sensor and cross-lifecycle generalization ability of the encoder.

To address this issue, we further adopt a unified frozen-encoder linear-probe protocol on LifetimeSITR to enable a more direct comparison across tactile representations. Specifically, all pretrained encoders are kept frozen, their final spatial features are mean-pooled without trainable parameters, and only lightweight linear prediction heads are trained on New split *train_set*. This setup minimizes the influence of task-specific decoder architectures and focuses the evaluation on the linear accessibility of the information preserved by the encoder and its ability to generalize across domains. The complete object-classification and contact-pose estimation results under different lifecycle conditions are reported in Tables 7 and 8, respectively.

For object classification, generic ImageNet-pretrained ViTs, particularly ViT-Large, exhibit strong linear separability under both clean and degraded conditions, indicating that high-capacity visual representations can preserve rich object-discriminative information from tactile images. In contrast, several tactile-specific representations, such as SITR, show more pronounced performance variations across transfer settings and lifecycle states. Excluding the ImageNet-pretrained ViTs, our LifelongTouch achieves the best overall classification performance under the unified linear-probe protocol. Compared with SITR, it improves the average accuracy by 24.98, 14.86, and 28.61 percentage points under the NoT, Inter, and Intra settings, respectively, demonstrating stronger discriminability and stability across sensors and lifecycle variations. Meanwhile, the strong performance of ImageNet-pretrained ViTs also suggests that object classification can benefit substantially from high-level visual appearance cues. Therefore, classification accuracy alone may not provide a complete measure of whether a tactile representation truly captures transferable physical contact information.

A different trend is observed for contact-pose estimation. While several baselines achieve competitive performance in the no-transfer setting, LifelongTouch consistently obtains the lowest inter-sensor RMSE across all lifecycle conditions, including New, Light-drift, Wear, Stain, and Mixed. Its inter-sensor pose error remains below 1 mm for most degradation types and achieves an aged-condition average of 0.9477 mm, substantially outperforming the other frozen representations under the same linear-probe protocol. These results indicate that LifelongTouch preserves contact geome-

try in a form that is more transferable across sensors and more robust to lifecycle-induced appearance changes.

Together with the classification results, the common-probe analysis further shows that strong semantic separability and transferable physical geometry are not necessarily aligned, and supports the design goal of LifelongTouch to prioritize lifecycle-invariant contact information rather than sensor-specific visual appearance.

A.4.2 Detailed Ablation Analysis

To analyze the contribution of each component in LifelongTouch, we consider the following ablation variants:

- **w/ $\mathcal{L}_1$ normal loss:** Add an element-wise L1 loss on surface normals to examine whether explicit pixel-level normal supervision benefits representation learning.
- **w/o reference:** Remove the non-contact reference input to evaluate the contribution of sensor-state information.
- **Modulation only:** Keep reference modulation while removing reference attention to isolate the effect of modulation-based conditioning.
- **Attention only:** Keep reference attention while removing reference modulation to isolate the contribution of attention-based conditioning.
- **w/o contact residual:** Remove the residual connection of contact features to assess whether directly preserving contact information improves representation stability.
- **w/o alignment:** Remove the representation alignment objective to evaluate its role in learning invariant representations across sensors and lifecycle states.

Due to computational cost, ablations use a controlled reduced-budget screening setting, intended to compare relative trends rather than reproduce the final fully trained numbers. All variants are trained for 15 epochs with 300 batches per epoch using the same training/validation split, DINOv2 initialization, optimizer, learning rate, batch construction, and random seed.

The detailed results in Tables 9 and 10 allow us to further examine the key design choices of LifelongTouch through the following questions.

Does stronger pixel-level normal supervision improve representation transferability? We first examine whether stronger pixel-level geometric supervision necessarily improves the learned tactile representation. Adding the normal $\mathcal{L}_1$ loss consistently reduces transfer performance. On Aged Avg., classification accuracy decreases from 57.36% to 51.70% for Inter transfer and from 77.58% to 72.64% for Intra transfer. Similarly, the Inter pose RMSE increases from 0.9308 to 0.9921. This trend is also consistent across Light, Wear, Stain, and Mixed conditions. These results suggest that enforcing element-wise normal reconstruction may encourage the encoder to fit simulation-specific geometric details, whereas removing this constraint allows the representation to retain more transferable physical information.

Is the no-contact reference necessary? Removing the no-contact reference consistently degrades the learned representation across lifecycle conditions. On Aged Avg., the Inter and Intra classification accuracies decrease from 57.36% and 77.58% to 54.45% and 68.78%, respectively. The effect is particularly evident for pose estimation, where the NoT and Inter RMSE increase from 0.5468 and 0.9308 to 0.6355 and 1.0406. Similar degradation is observed under all individual aging conditions. This confirms that the reference provides important sensor-state information that helps distinguish contact-induced physical changes from sensor-specific appearance and lifecycle variations.

How should the reference be incorporated? We further compare the two reference-conditioning mechanisms by retaining either modulation or attention alone. Both variants are generally inferior to the full model, while modulation-only performs better than attention-only in most pose-estimation settings. For example, on Aged Avg., modulation-only achieves an Inter pose RMSE of 0.9560, compared with 1.0083 for attention-only and 0.9308 for the full model. For classification, modulation-only and attention-only achieve Intra accuracies of 74.85% and 73.71%, respectively,

Table 9: Full object classification ablation results on LifelongSITR under different lifecycle conditions. Classification performance is reported as accuracy (%). Aged Avg. reports the average over Light-drift, Wear, Stain, and Mixed conditions.

	New			Light-drift			Wear			Stain			Mixed			Aged Avg.		
Variant	NoT	Inter	Intra	NoT	Inter	Intra	NoT	Inter	Intra	NoT	Inter	Intra	NoT	Inter	Intra	NoT	Inter	Intra
LifelongTouch (Full)	98.16	58.72	80.08	96.96	59.70	76.14	98.08	56.84	78.81	97.55	57.68	78.31	95.94	55.23	77.06	97.13	57.36	77.58
w/ Normal $\mathcal{L}_1$	98.53	52.78	75.48	98.11	53.76	72.17	98.12	51.05	74.34	97.94	51.45	72.66	96.77	50.54	71.38	97.74	51.70	72.64
w/o Reference	97.55	55.60	71.21	96.36	56.85	67.84	97.47	53.93	69.76	95.31	54.52	69.62	96.25	52.51	67.89	96.35	54.45	68.78
Modulation Only	97.20	56.88	77.40	96.36	57.45	73.33	97.11	55.26	76.71	96.43	56.29	75.40	95.91	54.34	73.96	96.45	55.84	74.85
Attention Only	97.82	54.90	76.68	97.49	55.43	73.55	97.66	53.36	75.15	96.01	53.93	74.12	96.68	52.06	72.01	96.96	53.70	73.71
w/o Contact Residual	97.73	56.03	76.21	96.72	57.19	71.54	97.61	52.81	74.77	96.63	54.72	74.54	96.22	51.90	71.36	96.80	54.15	73.05
w/o $\mathcal{L}_{\text{align}}$	98.06	56.12	71.08	97.76	57.43	67.55	97.67	53.71	69.85	96.78	55.37	69.78	95.89	53.68	67.36	97.02	55.05	68.64

Table 10: Full contact-pose estimation ablation results on LifelongSITR under different lifecycle conditions. Pose estimation performance is reported as RMSE (mm, ↓).

	New		Light-drift		Wear		Stain		Mixed		Aged Avg.	
Variant	NoT	Inter	NoT	Inter	NoT	Inter	NoT	Inter	NoT	Inter	NoT	Inter
LifelongTouch (Full)	0.5392	0.9202	0.5473	0.9211	0.5424	0.9265	0.5463	0.9325	0.5513	0.9430	0.5468	0.9308
w/ Normal $\mathcal{L}_1$	0.5371	0.9688	0.5529	0.9645	0.5374	0.9741	0.5452	0.9952	0.6026	1.0346	0.5595	0.9921
w/o Reference	0.6115	1.0155	0.6252	1.0088	0.6187	1.0197	0.6184	1.0502	0.6797	1.0835	0.6355	1.0406
Modulation Only	0.5580	0.9330	0.5794	0.9459	0.5606	0.9353	0.5650	0.9598	0.6100	0.9830	0.5788	0.9560
Attention Only	0.5935	0.9759	0.6124	0.9812	0.5969	0.9830	0.6053	0.9978	0.7439	1.0712	0.6396	1.0083
w/o Contact Residual	0.5884	0.9979	0.6118	1.0080	0.5920	1.0037	0.6000	1.0271	0.6682	1.0529	0.6180	1.0229
w/o $\mathcal{L}_{\text{align}}$	0.5675	0.9213	0.5780	0.9236	0.5697	0.9294	0.5720	0.9467	0.6508	0.9993	0.5926	0.9497

both below the 77.58% of LifelongTouch. These results indicate that modulation and attention capture complementary aspects of reference conditioning: modulation adjusts contact features according to the current sensor state, whereas attention enables explicit interaction between reference and contact features.

Is preserving direct contact information important? Removing the contact residual leads to a clear degradation, particularly for geometric prediction. On Aged Avg., the Inter pose RMSE increases from 0.9308 to 1.0229, while the NoT RMSE increases from 0.5468 to 0.6180. Classification performance also decreases, with the Intra accuracy dropping from 77.58% to 73.05%. The degradation becomes more evident under stronger lifecycle variations such as Mixed, where the Inter pose RMSE increases from 0.9430 to 1.0529. This suggests that directly preserving contact features is important for retaining local geometric information that may otherwise be weakened by reference-conditioned feature transformation.

What is the role of representation alignment? Finally, we remove the representation alignment objective to investigate its contribution to sensor- and lifecycle-invariant feature learning. The largest degradation is observed in classification transfer: on Aged Avg., Intra accuracy decreases from 77.58% to 68.64%, while Inter accuracy decreases from 57.36% to 55.05%. In contrast, the effect on pose estimation is relatively smaller, with the Inter RMSE increasing from 0.9308 to 0.9497. This difference suggests that representation alignment primarily encourages invariance across sensor identities and lifecycle states, which is particularly important for semantic transfer, while local geometric information required by pose estimation can still be partially preserved without explicit alignment.

Overall, the no-contact reference, complementary reference-conditioning mechanisms, contact residual, and representation alignment contribute to different aspects of transferability and geometric preservation, while removing the pixel-level normal $\mathcal{L}_1$ constraint yields a more transferable tactile representation across sensors and lifecycle conditions.

A.4.3 Qualitative Results

We further qualitatively analyze LifelongTouch from both physical prediction and representation perspectives. As shown in Fig. 8, LifelongTouch predicts coherent surface-normal fields across different sensors and lifecycle conditions, indicating that local contact geometry is consistently pre-

served despite substantial appearance variations. Similarly, Fig. 9 shows that the predicted force distributions retain the major spatial contact patterns under different sensor and degradation settings.

We then examine the learned encoder features directly. Fig. 10 compares spatial feature responses from Sparsh, T3, SITR, and LifelongTouch. While baseline representations are more affected by degradation-related appearance changes, LifelongTouch places stronger emphasis on the physical contact region and suppresses irrelevant variations outside the contact area. At the global representation level, the t-SNE visualization in Fig. 11 shows more compact intra-class grouping and clearer separation between object categories across lifecycle conditions. Consistently, the cosine-similarity retrieval results in Fig. 12 show that LifelongTouch retrieves more samples from the same object class under the challenging Mixed condition. Overall, these qualitative results support that LifelongTouch retains contact-relevant physical structure while reducing sensitivity to sensor- and lifecycle-dependent appearance variations.

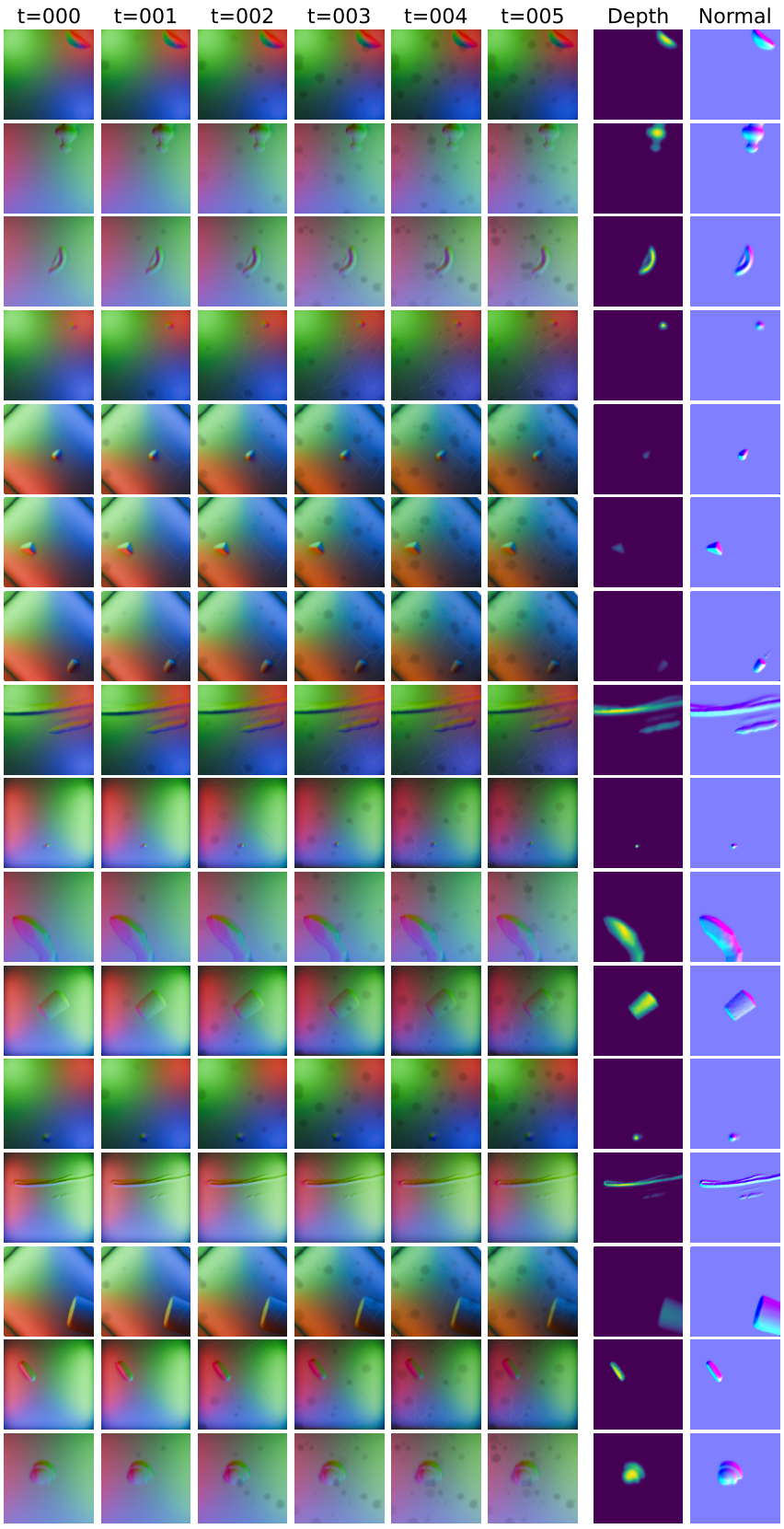

Figure 5: **Samples from the simulation dataset.**

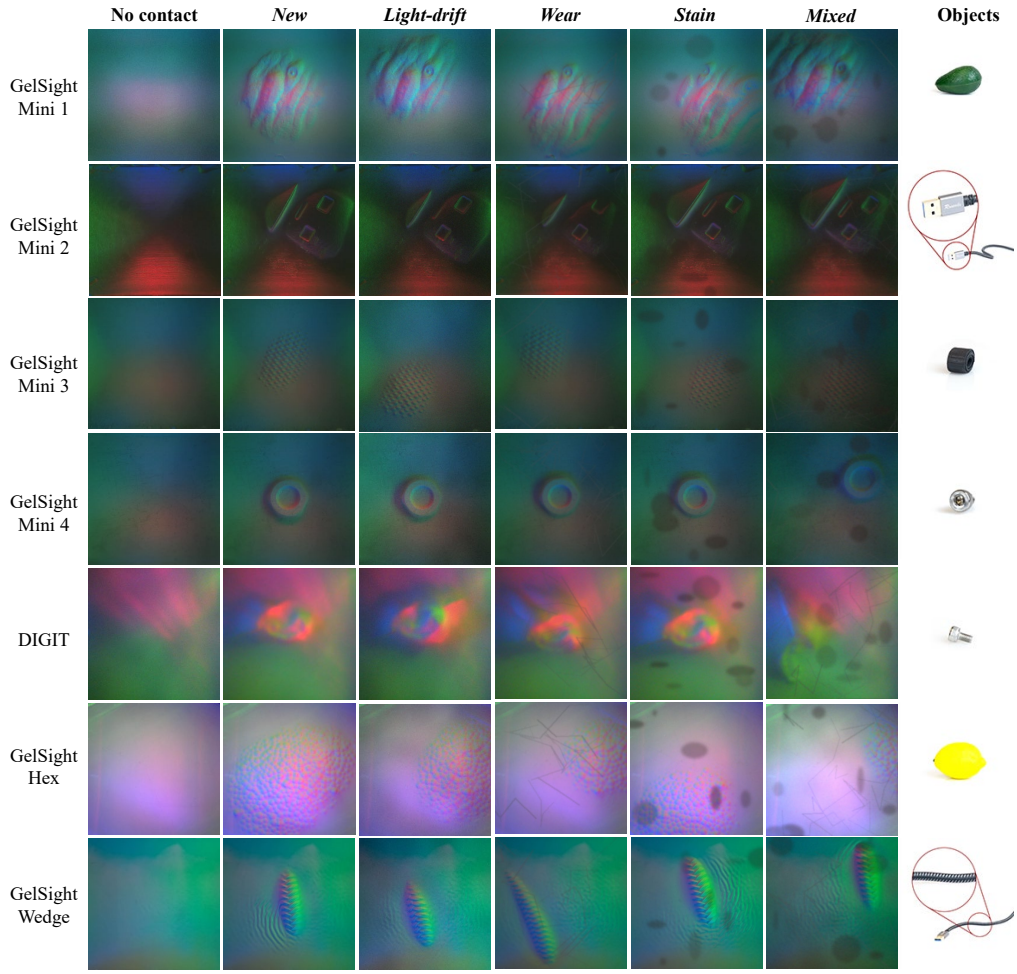

Figure 6: **Pseudo-lifecycle examples from LifetimeSITR.** Representative tactile observations under *New*, *Light-drift*, *Wear*, *Stain*, and *Mixed* conditions. Different degradation states are applied while preserving the underlying physical contact and task labels, providing controlled evaluation of robustness to lifecycle-induced observation shifts.

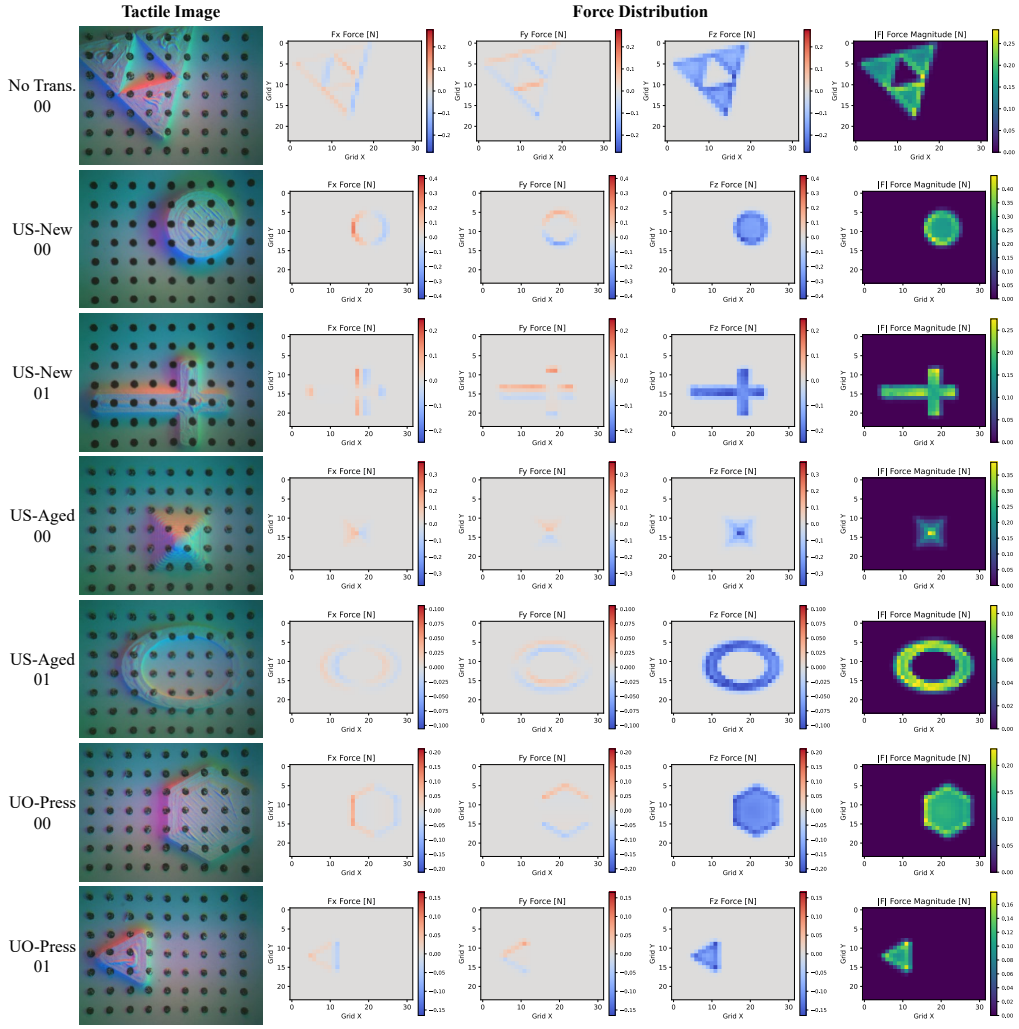

Figure 7: **Representative samples from LifetimeFEATS.** Examples from the four evaluation settings: *No Transfer (No Trans.)*, *Unseen Sensor with New Gel (US-New)*, *Unseen Sensor with Aged Gel (US-Aged)*, and *Unseen Object Press (UO-Press)*. Each example consists of a contact observation, and the corresponding three-axis force-distribution ground truth. These settings evaluate generalization across sensor identities, gel lifecycle states, and object-contact conditions.

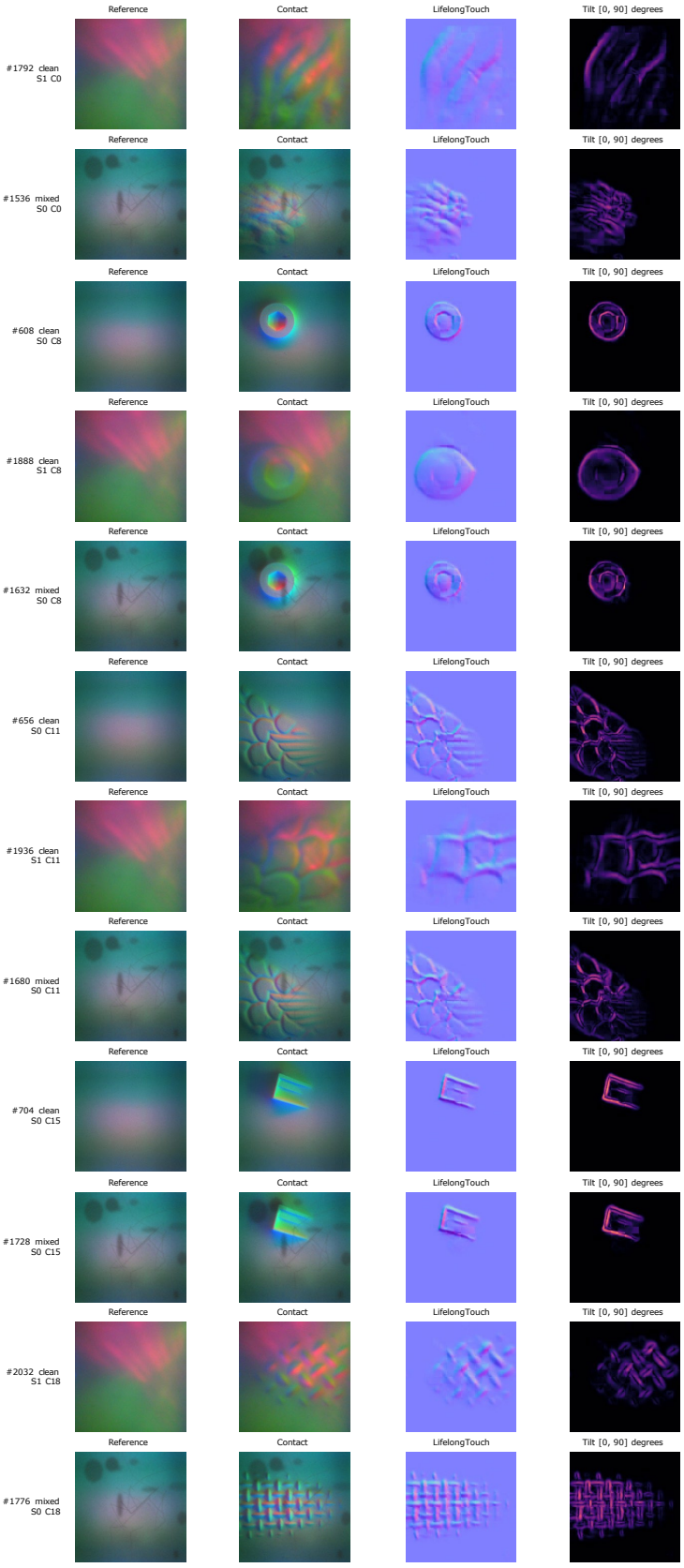

Figure 8: **Qualitative surface-normal prediction results.** Surface normals predicted by Lifelong-Touch under different sensor and lifecycle conditions.

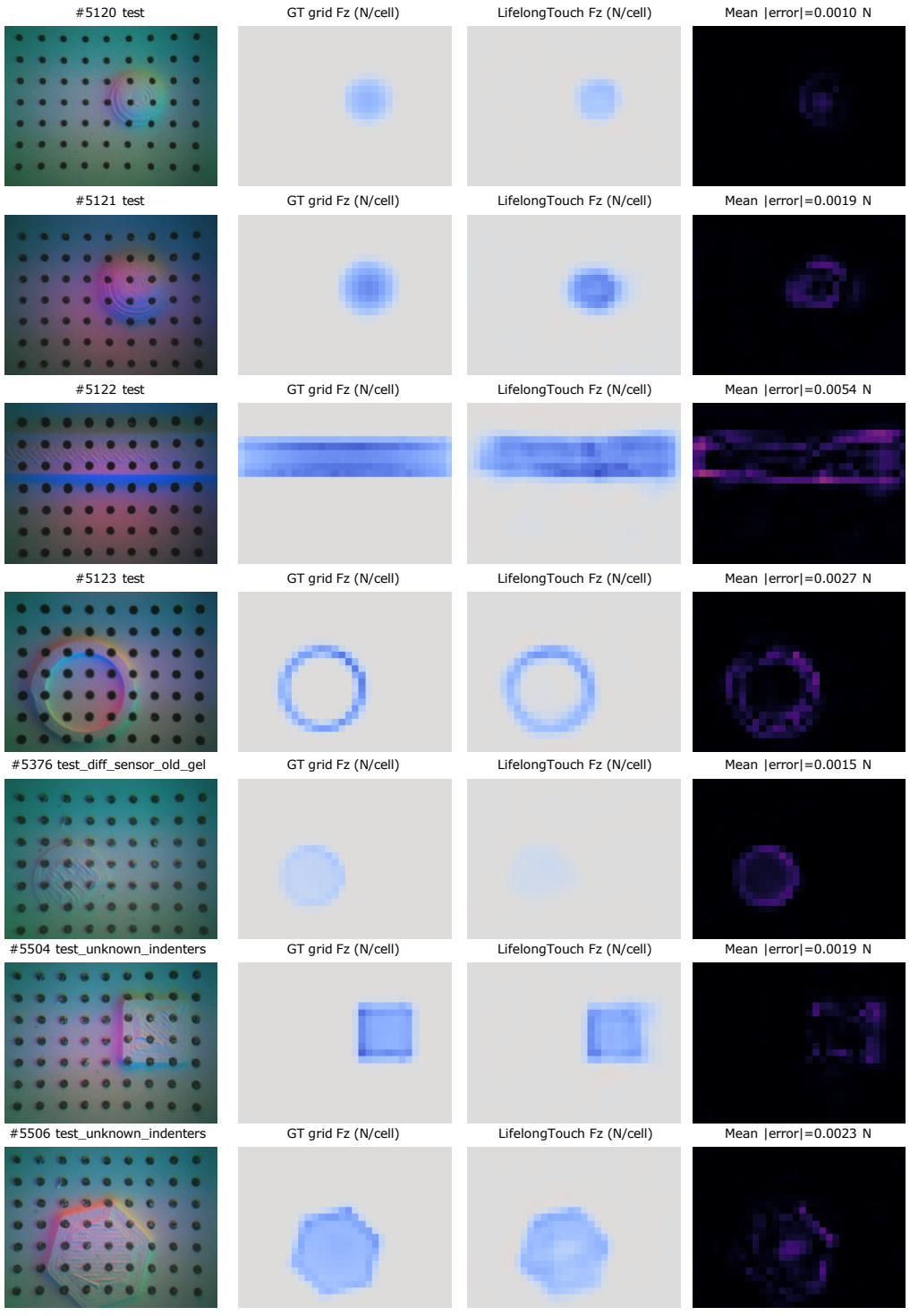

Figure 9: **Qualitative results on LifetimeFEATS.** Visualization of force-distribution estimation by LifelongTouch under different sensor and lifecycle conditions, demonstrating consistent recovery of the spatial contact-force patterns across domain variations.

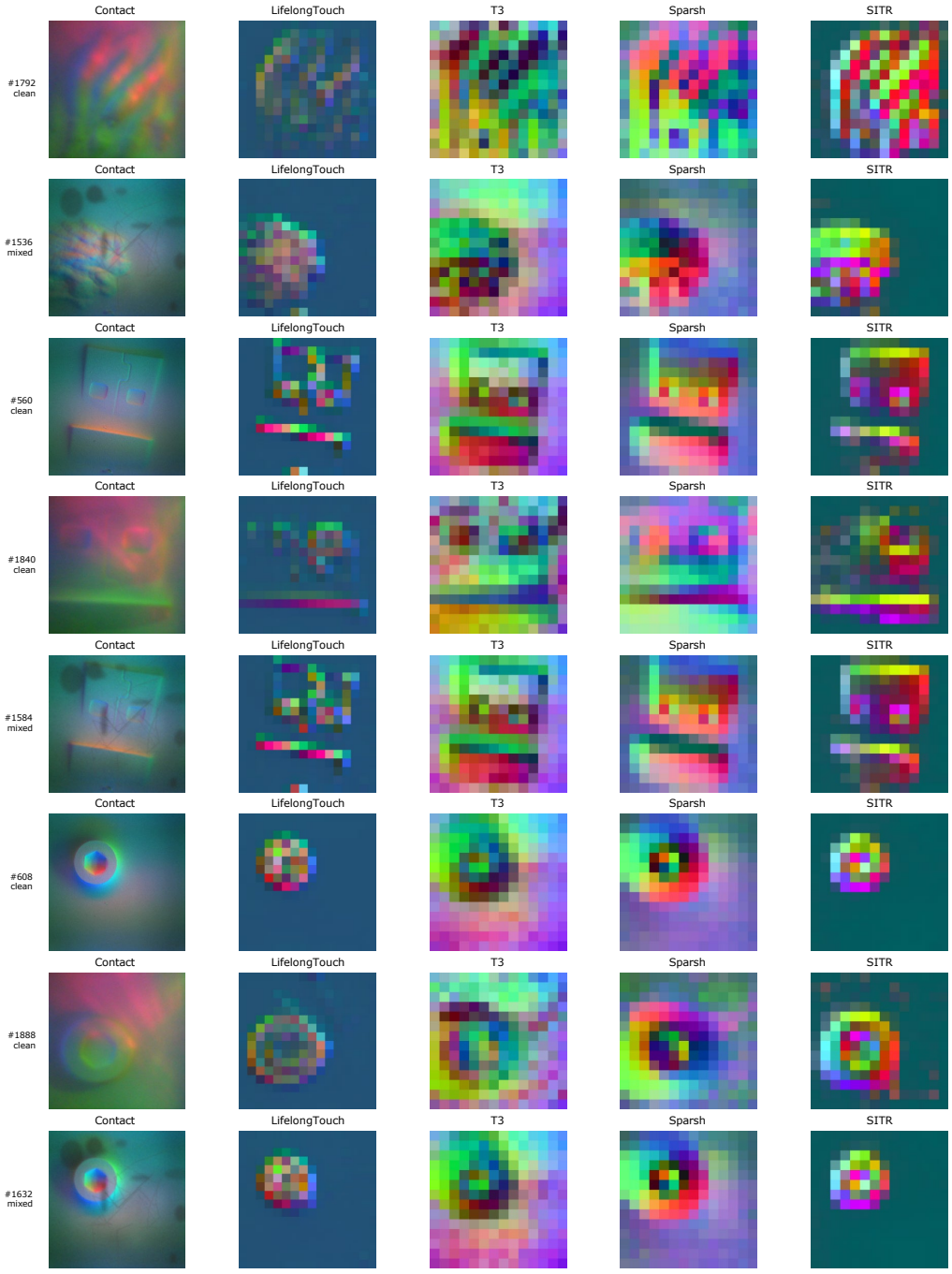

Figure 10: **Qualitative comparison of encoder features.** PCA visualizations of features from Sparsh, T3, SITR, and LifelongTouch under different lifecycle degradations. Compared with existing tactile representations, LifelongTouch more effectively suppresses degradation-induced appearance variations and concentrates its feature responses on the physical contact region, yielding more stable and contact-relevant representations across lifecycle states.

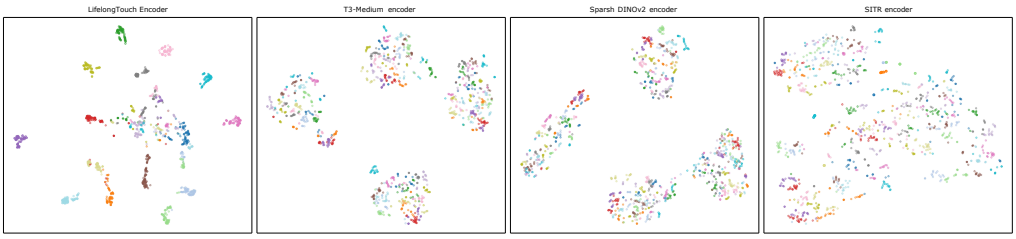

Figure 11: **t-SNE visualization of object-class representations.** Comparison of feature distributions from Sparsh, T3, SITR, and LifelongTouch on the LifetimeSITR object-classification task. LifelongTouch exhibits more compact intra-class clusters and clearer inter-class separation across lifecycle conditions, indicating that its representation better preserves object-discriminative information while reducing degradation-induced feature variations.

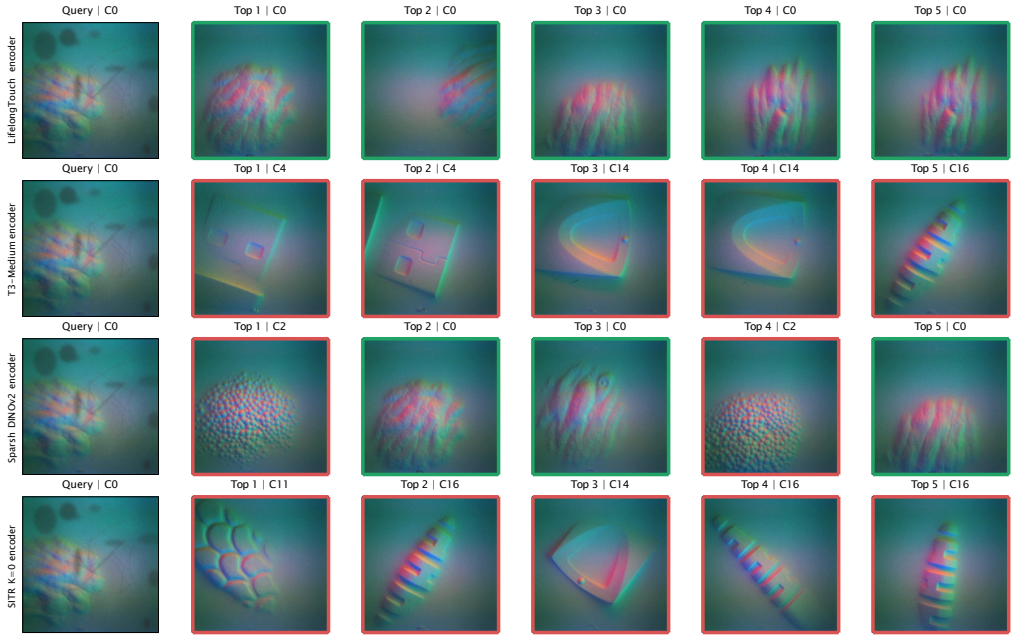

Figure 12: **Feature retrieval under lifecycle degradation.** Retrieval results based on cosine similarity between encoder features in the Mixed lifecycle condition. Given each query sample, the most similar tactile observations are retrieved in the learned feature space, with green boxes indicating samples from the same object class as the query.