TAC-DINO: LEARNING TACTILE FEATURES WITH PATCH ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Multisensory perception is essential for robotic manipulation, with tactile sensing providing critical information for fine-grained, contact-rich interactions. Although prior work integrates tactile and spatial signals to improve perception, vision-tactile learning commonly relies on object-level alignment. Addressing this limitation requires systematically aligned 3D vision-tactile data. We develop a data collection system and introduce **Touch3D**, a dataset comprising **505 real-world objects** and **20,025 tactile contacts**. Building on this dataset, we establish the **Contact-Centric Multimodal Tactile Representation Benchmark (CoMT-Bench)**, which expands evaluation from object-level retrieval and tactile properties to include patch-level contact evaluation. We further propose **Geometry-Relation Distilled Tactile-Centric Cross-Modal Learning (GeoTAC)**, which explores tactile learning across all multisensory data. Experiments on the Touch3D test set, patch-level alignment improves the average score by **5.32** over object-level alignment, with GeoTAC's geometric supervision improving **4.48** more. In robotic manipulation, GeoTAC improves TP by **19.14** and **11.34** over DINOv2 (Vision-only) and DINOv2 (Vision-Tactile), respectively.

1 INTRODUCTION

Human perceive the world through multiple senses, including vision, hearing, and touch (Hedger et al., 2025). Inspired by this, multi-directional breakthroughs have emerged in various artificial intelligence applications. Computer vision now excels at object understanding, driven by the emergence of numerous 2D image datasets (Lin et al., 2014; Li et al., 2023b; 2024). 3D vision has overcome data bottlenecks with the rise of Objaverse (Deitke et al., 2023) and Objaverse-XL (Deitke et al., 2024). Recently, tactile sensing become a more critical modality especially for robotics. The development of diverse sensors (Taylor et al., 2022; Bhirangi et al., 2024) and hardware systems has yielded substantial amounts of paired vision-tactile data. Here different modalities share intrinsic physical properties, such as geometric shape, material composition, and texture. Therefore, it is vital to connecting these corresponding modalities for a joint understanding of objects and environment, particularly for robotics, virtual reality and next generation physical intelligence.

Multimodal large language models have achieved substantial success in question-answering tasks (Liu et al., 2023), with similar progress in vision-language-action (VLA) models for generalizable manipulation (Kim et al., 2024). Despite these advances, limited physical and spatial understanding can still lead to hallucinations and interaction inconsistencies (Li et al., 2023a). Existing work incorporates tactile (Li et al., 2019; Huang et al., 2025b) and 3D representations (Wang et al., 2024) to enhance contact-rich manipulation and spatial perception. The efficiency of incorporating these modalities suggests that next-generation models will continue to incorporate more modalities toward human-like intelligence. However, current research lacks both a formal definition of the relationships between tactile sensing and other modalities and a unified approach that integrates all relevant modalities for tactile learning. The main difficulty is the lack of large-scale 3D vision-tactile datasets collected under a formally defined and systematic protocol. ObjectFolder Real closely matches these requirements, but its limited scale of 100 objects (Gao et al., 2023) constrains broader methodological exploration.

To this end, we first develop a data collection system and collect a 3D vision-tactile dataset, **Touch3D**, comprising **505 objects** and **20,025 tactile samples**, together with scanned object geom-

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.

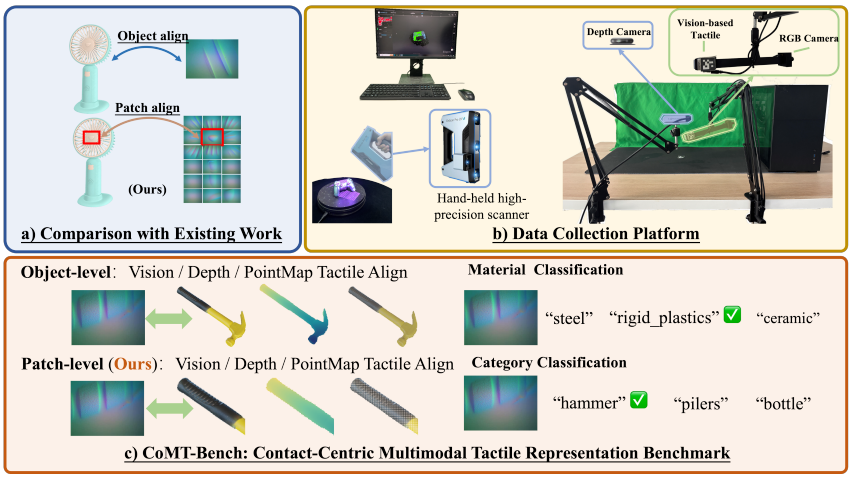

Figure 1: Overview of the Tac-DINO framework. (a) We focus on patch-level alignment rather than the object-level alignment used in existing work. (b) We develop a data collection platform for 3D vision-tactile data. (c) Building on existing work, we propose CoMT-Bench to evaluate tactile representations using multisensory data.

etry, contact-location annotations, and material labels (Fig. 1(b)). We then introduce the **Contact-Centric Multimodal Tactile Representation Benchmark (CoMT-Bench)** (Fig. 1(c)), which formally associates each tactile observation with a physically localized region in RGB, depth, and pointmap data, rather than an object image with an undefined scale and contact correspondence (Fig. 1(a)). CoMT-Bench evaluates both object-level and patch-level alignment between tactile and other perceptual modalities, as well as tactile-based material and object-category classification. Together, these evaluations provide a common basis for comparing existing methods and studying tactile-centric representation learning.

Furthermore, to investigate tactile representation learning with patch-aligned multi-sensory data, we propose **Geometry-Relation Distilled Tactile-Centric Cross-Modal Learning (GeoTAC)**. GeoTAC combines DINO local-to-global self-distillation, tactile-centric cross-modal alignment, and geometry distillation that transfers the geometric relationships between tactile observations and object surfaces. Through extensive experiments, we investigate three questions: (1) Does patch-level alignment improve tactile performance compared with object-level? (2) Can depth and point clouds, when incorporated through appropriate methods improve tactile representation learning? (3) Do the learned tactile representations consistently improve performance in downstream robotic manipulation tasks? Our experiments lead to the following conclusions:

Our main contributions are threefold: **(1)** We formally define the spatial correspondence between tactile observations and visual and geometric modalities, and develop a data-collection system for constructing a large-scale 3D vision-tactile dataset. **(2)** We introduce CoMT-Bench, which evaluates tactile representations through object-level alignment, patch-level multimodal alignment, and tactile property recognition, including material and object-category classification. **(3)** We propose GeoTAC, a tactile-centric cross-modal representation learning method. GeoTAC improves over object-level alignment by 5.32, gains a further 4.48 by incorporating depth and pointmaps. In robotics mainpulation task, it improves average TP by 19.14 over DINOv2 (Vision-only), and by 11.34 over DINOv2 (Vision-Tactile), respectively.

2 Related Work

2.1 Multi-source Tactile Data

Tactile sensors provide contact cues essential for dexterous manipulation. Representative technologies include visuo-tactile (Yuan et al., 2017a; Lambeta et al., 2020; Sipos et al., 2024; Ren et al., 2023), capacitive (Glauser et al., 2019; Wu et al., 2020; Xu et al., 2016), resistive (Sundaram et al.,

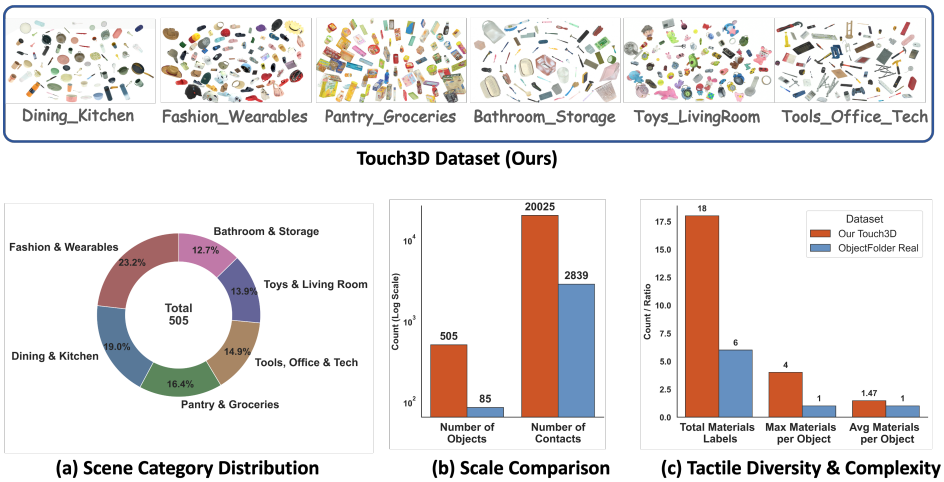

Figure 2: Data distribution and labeling comparison.

2019; Bhattacharjee et al., 2013; Stassi et al., 2014), and magnetic (Bhirangi et al., 2025) sensing. Visuo-tactile sensors are particularly prevalent because they are inexpensive and easy to fabricate. ObjectFolder (Gao et al., 2021; 2022; 2023) provides visual and acoustic signals, X-Capture (Clarke et al., 2025b) includes material annotations, and Touch100K (Cheng et al., 2024) introduces multi-granularity language labels.

2.2 Vision-Tactile Learning

Recent vision-tactile learning tackles representation integration and sensor heterogeneity across three main directions. The first is learning composite representations through imitation learning (Heng et al., 2025; Romero et al., 2024; Lin et al., 2024; Pattabiraman et al., 2024), which updates parameters via action alignment. The second transfers tactile perception into intermediate representations, for example, by converting it into point clouds (Huang et al., 2025a; Yuan et al., 2024) or reconstructing it with NeRF (Suresh et al., 2024; Dou et al., 2024). The last is learning tactile features using self-supervised methods (Feng et al., 2025; 2026; Rodriguez et al., 2025; Guzey et al., 2023; Cheng et al., 2026; Higuera et al., 2026). Despite these advances, tactile signals are typically captured with only coarse alignment to vision, neglecting the precise alignment between the tactile data and the contact patch area. In this work, we propose to study vision-tactile patch alignment feature learning.

2.3 Representation Learning

Robust representation learning from multimodal data is critical for robotic perception, driving the adoption of self-supervised learning (SSL) paradigms like DINO (Caron et al., 2021), CLIP (Radford et al., 2021a), V-JEPA (Bardes et al., 2024), MAE (He et al., 2021), and GAN (Goodfellow et al., 2014). Among these, DINO distinctly excels in discriminative feature extraction. Unlike methods focused on reconstruction, generation, or cross-modal alignment, DINO's self-distillation mechanism captures both fine-grained local details and holistic global semantics. DINOv2 (Oquab et al., 2023) further scales this framework to massive corpora. Since the local-to-global paradigm perfectly aligns with the requirement of capturing local tactile contacts within a global scene context, we utilize DINOv2 as our base model to investigate vision-tactile patch alignment learning.

3 Touch3D Dataset

Multisensory learning is critical for representation learning in fine-grained robotic manipulation. Tactile, depth, and pointmap observations are commonly integrated with vision to address the limitations of vision-only systems in contact-rich interaction and spatial perception. However, collecting

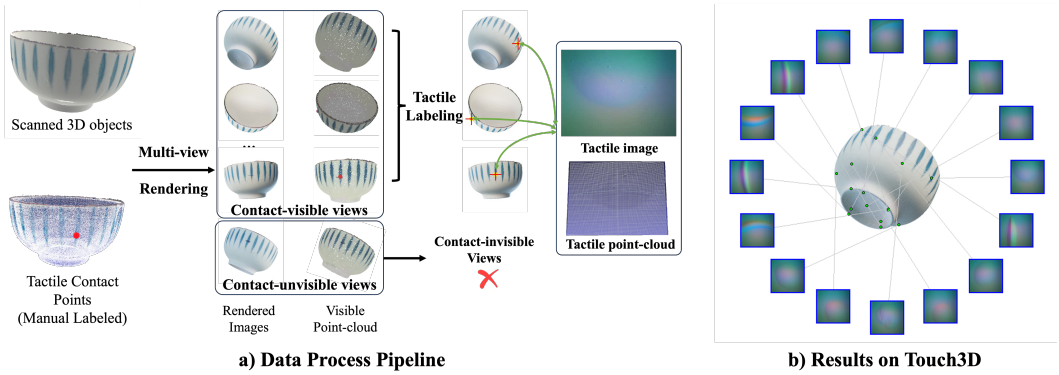

Figure 3: **Contact-region data curation.** Scanned shapes, registered views, and contact-center annotations associate a tactile contact with a local visual region.

real-world multisensory data is expensive and time-consuming. ObjectFolder Real (Gao et al., 2023) collects data from 100 real-world objects [1] to mitigate this challenge. However, its limited object diversity restricts its suitability for large-scale representation learning, making it more appropriate for evaluation than training. The dataset was collected using a Franka Emika Panda robot arm equipped with a GelSight visuo-tactile sensor (Dong et al., 2017; Yuan et al., 2017b). This setup requires carefully controlled physical interactions and is difficult to scale because data collection relies on time-consuming teleoperation. Therefore, to address data scarcity and support multisensor-enhanced tactile learning, we introduce a movable, time-efficient data collection system capable of handling arbitrary real-world objects. We then propose **Touch3D**, a dataset comprising **505** real objects and **20,025** tactile contacts.

3.1 3D Shape Data Collection

We use an EinScan Pro 2X 2020 handheld 3D scanner to collect a high-quality 3D mesh that includes point locations and corresponding color textures for each object. The scanner utilizes structured light scanning, which involves three steps to construct high-quality 3D features: projecting grating patterns, capturing the deformed grating, and calculating the 3D point cloud. The minimum point distance of the produced 3D point cloud reaches 0.2 mm, enabling a high-definition and precise reconstruction of the physical form. For each object, we scan over two faces, some complex objects scan $4 \sim 5$ faces, and then combine them into one 3D shape. After merging, we utilize the official scanning software to filter out noise points, as shown in Fig. 1 b).

3.2 Vision-Tactile Data Collection

We utilize the GelSight Mini as our tactile sensor to construct an vision-tactile data collection platform. The platform consists of an omnidirectional stand mounted on a desktop, a custom 3D-printed holder attached to the stand, and a GelSight Mini sensor alongside an RGB camera fixed at opposite ends of this holder. Additionally, a depth camera is mounted on a separate omnidirectional stand to record the collection process, providing third-view data. For each object, we first use the tactile sensor to record the pressing process, accompanied by a third-person perspective video captured by the depth camera. Next, we rotate the 3D-printed sensor holder by 180 degrees while keeping all other components stationary. Once a paired data sample is collected, the data collection system displays the scanned 3D shape, and the operator manually selects the contact location. After completing these three steps to gather a set of paired image-tactile data, we move to the next contact location to continue the collection process.

[1] Data obtained from `https://objectfolder.stanford.edu/`; after filtering out incompletely labeled data, 85 objects with 2,839 contacts remained usable.

Table 1: Comparison of multisensory datasets in terms of collection environments, numbers of objects and tactile contacts, and included modalities.

Dataset	Env	Obj.	Tac.	Tactile Sensor	Correlated Modalities	
					RGB-D	Scanned 3D
Touch and Go (Yang et al., 2022)	Real	3,971	13.9k	Gelsight	✗	✗
Objectfolder 2.0 (Gao et al., 2022)	Simulation	1,000	-	Gelsight	✓	✓
X-Capture (Clarke et al., 2025a)	Real	500	3,000	DIGIT	✓	✗
ObjectFolder Real (Gao et al., 2023)	Real	100	2,995	Gelsight	✓	✓
Touch3D (Ours)	Real	**505**	**20,025**	**Gelsight Mini**	✓	✓

3.3 Human Annotation and Sanity Check

To label the contact locations and tactile materials, as well as to double-check the quality of the scanned meshes, tactile, and vision data, we developed a data labeling system. The system includes features to load captured data, add or select tactile material labels, and show the contact location annotations. When the scanned 3D shape includes noisy points or the contact location exhibits bias, we utilize MeshLab (Cignoni et al., 2011) to filter out the noise or correct the labels. When vision or tactile data is damaged or missing, we filter it out and then renumber the remaining data.

3.4 Data Statistics

Using our data collection and labeling system, we referenced the object selection from PACE (You et al., 2024) and collected **505** objects with **20,025** tactile contacts. The overall time cost was over **700** hours, including **200** hours to scan 3D shapes at a rate of 2-3 objects per hour, and **500** hours to collect vision-tactile data.To the best of our knowledge, our **Touch3D** dataset, is the largest 3D Vision-Tactile dataset in the community, as shown in Table 1. A comparison with ObjectFolder Real (Gao et al., 2023) can be found in Fig. 2. The ObjectFolder Real features kitchen objects. Our dataset includes 6 different scenarios: *dining_kitchen, fashion_wearables, pantry_groceries, bathroom_storage, toys_livingroom, and tools_office_tech*. Each scenario has a balanced number of objects. After eliminating the visual indivisibility tactile material label, our dataset includes 18 materials, compared to 6 in ObjectFolder Real. In ObjectFolder Real, each object has only one material. Our dataset includes more diverse objects, with each object having at most 4 materials, a mean of 1.47 material labels.

4 Methods

Multi-sensory learning enhances perceptual understanding, while integrating tactile sensing with depth and point maps is critical for contact-rich manipulation and spatial understanding in robotics. The effectiveness of joint multimodal learning encourages next-generation models to incorporate an increasing number of modalities. Motivated by these observations, we first formally define the relationship between tactile sensing and other modalities. We then introduce the Contact-Centric Multimodal Tactile Representation Benchmark (**CoMT-Bench**) (Sec. 4.1), which extends tactile evaluation from object-level alignment to patch-level alignment. Finally, to investigate tactile-centric representation learning from patch-aligned multisensory data, we propose Geometry-Relation Distilled Tactile-Centric Cross-Modal Learning (**GeoTAC**) (Sec. 4.2).

4.1 CoMT-Bench: Contact-Centric Multimodal Tactile Representation Benchmark

Existing tactile learning benchmarks commonly evaluate object-level cross-modal alignment and material or category classification using manually annotated labels, without a standardized definition of the visual region corresponding to each tactile observation. We extend this evaluation from object-level to patch-level alignment to comprehensively evaluate tactile features.

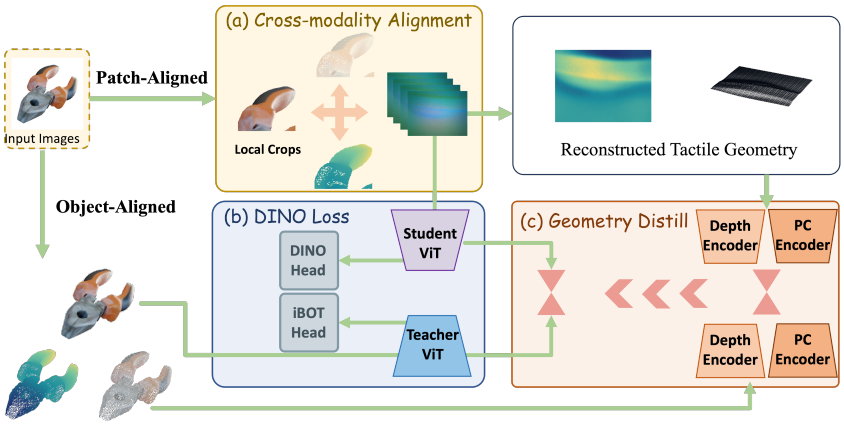

Figure 4: Overview of the proposed GeoTAC framework.

For each tactile contact at $\mathbf{p}$, we extract aligned observations from RGB, depth, and pointmap inputs: $X^m_{\text{local}} = \text{Crop}(X^m, \pi_m(\mathbf{p})), \quad m \in \{V, D, P\}$, where π_m projects the contact location into modality m. Given a local tactile query T_{local}, object-level evaluation retrieves a global observation from modality m: $G^*_m = \arg\max_{G \in \mathcal{G}_m} \text{Sim}(E_T(T_{\text{local}}), E_m(G))$, where $\mathcal{G}_m$ is the global candidates. Object-level retrieval is evaluated using object-category kNN. Patch-level evaluation replaces $\mathcal{G}_m$ with the local candidate set and treats only the observation from the same contact as positive. We additionally evaluate material prediction and object-category classification using the tactile representation alone.

4.2 GeoTAC: Geometry-Relation Distilled Tactile-Centric Cross-Modal Learning

We learn tactile representations from patch-aligned RGB, depth, and pointmap observations. As shown in Figure 4, GeoTAC consists of tactile-centered cross-modal alignment, the retained DINOv2 objective, and geometry distillation.

Tactile-Centered Cross-Modal Alignment. We align local RGB, depth, and pointmap observations with the tactile representation of the same contact: $\mathcal{L}_{\text{align}} = \frac{1}{3}[\mathcal{L}_c(V_l, T_l) + \mathcal{L}_c(D_l, T_l) + \mathcal{L}_c(P_l, T_l)]$, where $\mathcal{L}_c$ is the bidirectional contact-level contrastive loss. It uses soft region matching rather than comparing patches with identical indices. The external RGB, depth, and point encoders are frozen, while the tactile encoder, temporal head, and projection heads are optimized.

Local-to-Global Distillation. Following DINOv2 (Oquab et al., 2023), the teacher processes global RGB crops, while the student processes the corresponding local tactile observations:

$$\mathcal{L}_{\text{lg}} = -\frac{1}{N_g N_l} \sum_{g=1}^{N_g} \sum_{l=1}^{N_l} P_{\text{teacher}}(V_g)^{\mathsf{T}} \log P_{\text{student}}(T_l), \tag{1}$$

where P_{teacher} and P_{student} are the DINO-head output distributions. The original global-crop DINO, iBOT, and KoLeo objectives are retained, giving $\mathcal{L}_{\text{DINO}} = \mathcal{L}_{\text{lg}} + \mathcal{L}_{\text{gg}} + \mathcal{L}_{\text{iBOT}} + \mathcal{L}_{\text{KoLeo}}$, where $\mathcal{L}_{\text{gg}}$ denotes the global RGB DINO loss.

Geometry-Relation Distillation. We reconstruct tactile depth $\widetilde{D}_l$ and tactile point cloud $\widetilde{P}_l$ and compare them with the corresponding local depth D_l and pointmap P_l. Their equally weighted bidirectional similarity distributions form the geometry teacher: $R^r_{\text{geo}} = \frac{1}{2}(R^r_{D,\widetilde{D}} + R^r_{P,\widetilde{P}}), \quad r \in \{\rightarrow, \leftarrow\}$. The geometry relations are distilled into the local RGB–tactile matching distribution $R^r_{V,T}$:

$$\mathcal{L}_{\text{geo}} = \frac{1}{2} \sum_{r \in \{\rightarrow, \leftarrow\}} D_{\text{KL}} \left(\left[\text{softmax} \left(R^r_{\text{geo}} / \tau \right) \right] \middle\| \, \text{softmax} \left(R^r_{V,T} / \tau \right) \right). \tag{2}$$

The reconstructed tactile depth and point cloud are used only during training.

The complete GeoTAC objective is $\mathcal{L}_{\text{GeoTAC}} = \mathcal{L}_{\text{DINO}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}}$, where λ_{align} and λ_{geo} control the contributions of cross-modality alignment and geometry distill.

5 Experiments and Discussion

5.1 Experimental Settings

We systematically evaluate patch alignment using DINOv2 (Oquab et al., 2023) ViT-Large models. The collected data are split into training, validation, and test sets at 4:1:1. We also use ObjectFolder Real as an additional test set. We include general vision-tactile methods, Contrastive Learning (CL) (Radford et al., 2021b) and DAR (Zhen et al., 2019), as baselines. We further compare multisensory alignment methods, including ImageBind (Girdhar et al., 2023), UniTouch (Yang et al., 2024), and ViT-Lens (Lei et al., 2024), for tactile learning using all available multisensory data. For modalities without an encoder in these methods, we incorporate general-purpose encoders, such as PointBERT.

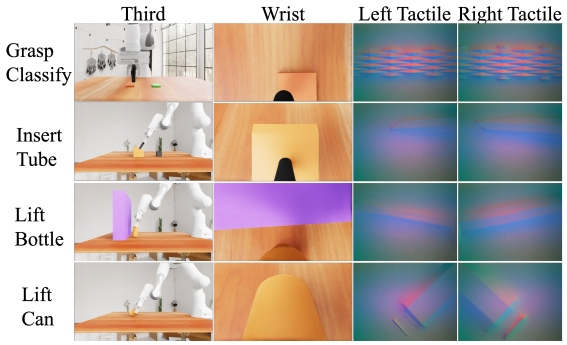

Figure 5: Visualization of the UniVTAC evaluation.

We also use the UniVTAC (Chen et al., 2026) simulator as a testbed for evaluating robotic manipulation, with the GelSight Mini tactile sensor matching that used in Touch3D. Following the official configuration, we select four tasks, as shown in Figure 5, and collect 100 demonstration episodes per task to train Diffusion Policy Chi et al. (2025) for 30,000 steps.

5.2 Metrics

Tactile Representation Evaluation. Object-level retrieval is measured by object-category k-NN Top-5 accuracy. Patch-level alignment is measured by exact-contact Top-5 retrieval accuracy (E@5), averaged over both retrieval directions. For tactile properties, we report k-NN mAP for multi-label material classification and probe accuracy for object-category classification. *Avg.* denotes the arithmetic mean of the reported metrics.

Robotic Manipulation. We initialize each policy with the pretrained encoders and run 40 rollouts using the official UniVTAC settings. We report success rate (SR) and average final task progress (TP, %). *Avg.* is the mean across the four tasks, computed before rounding.

5.3 Effectiveness of Patch-Aligned Tactile Learning

In this subsection, we investigate whether patch-level alignment is more effective than object-level alignment for tactile representation learning. We focus on the vision-tactile setting, comparing object-level vision-tactile alignment, patch-aligned tactile learning, and patch-aligned tactile learning combined with vision-tactile methods.

Table 2 summarizes the results of vision-tactile learning methods. From the results, we can draw the following conclusions: (1) Vision-tactile alignment is helpful for tactile feature learning (see all results compared with the Tac row). (2) Patch alignment improves performance compared with object-level alignment. The Tac + CL row shows lower patch-level performance than the Tac-only baseline. Patch Tac + CL achieves the best results. (3) The results on ObjectFolder Real are consistent with those on Touch3D, but overall performance is lower due to cross-sensor generalization.

Table 2: Tactile representation learning with object-, patch-level, and tactile-property evaluation.

Training configuration	Object-level $T_l \rightarrow V_g$ Top-5	Patch-level $T_l \leftrightarrow V_l$ E@5	Tactile properties Material mAP	Tactile properties Category Acc.	Avg. ↑
Touch3D					
Tac	27.49	4.32	25.94	56.73	28.62 (0.00)
Tac + CL (Radford et al., 2021b)	37.68	3.72	28.81	58.98	32.30 (+3.68)
Patch Tac	46.50	6.46	30.37	56.42	34.94 (+6.32)
Patch Tac + CL (Radford et al., 2021b)	42.94	13.95	34.50	59.08	37.62 (+9.00)
Patch Tac + DAR (Aytar et al., 2017)	47.39	7.62	31.72	56.81	35.89 (+7.27)
ObjectFolder Real					
Tac	26.41	1.92	14.97	0.07	10.84 (0.00)
Tac + CL (Radford et al., 2021b)	27.13	1.82	15.95	0.33	11.31 (+0.47)
Patch Tac	30.26	1.92	16.62	1.51	12.58 (+1.74)
Patch Tac + CL (Radford et al., 2021b)	30.77	3.85	21.71	2.13	14.62 (+3.77)
Patch Tac + DAR (Aytar et al., 2017)	33.08	2.05	17.85	3.12	14.03 (+3.18)

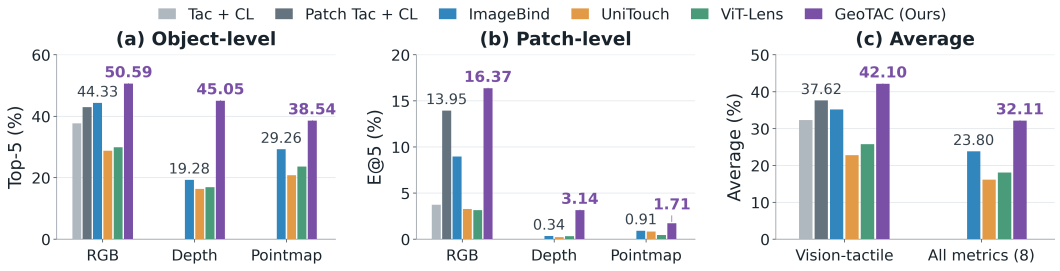

Figure 6: Visualization of multisensory learning results on CoMT-Bench.

5.4 Effectiveness of Tactile-Centric Cross-Modal Learning

In this subsection, we investigate two questions: (1) Does incorporating additional modalities improve performance over vision-tactile learning? (2) Is our proposed tactile-centric learning method more effective than treating all modalities equally during alignment?

Figure 6 presents the results using all multimodal sensory data. We observe that: (1) Incorporating more modalities requires stricter alignment, as all object-level results decline when additional modalities are involved. UniTouch's average score decreases by 9.55 compared with object-level alignment using CL. (2) Our proposed GeoTAC achieves the best results, improving the score by 4.48 over vision-tactile methods. (3) Correct labels are more important than complex alignment methods: Tac with contrastive learning (CL) and object-level multimodal alignment using ImageBind outperform other fine-grained methods. As seen in Figure 7, geometry aligns the correct patch area, and other object-level methods show global importance and incorrect areas.

5.5 Evaluation on Robotic Manipulation

In this subsection, we explore the utility of tactile representations learned under different training strategies for robotic manipulation tasks. We replace the vision and tactile encoders with our trained models and post-train the policies on robot demonstrations.

Table 4 presents the results on robotic manipulation tasks. We observe the following: (1) The real-world results are consistent with the representation benchmark. Comparing Table 4 and Figure 6, GeoTAC achieves the best results, UniTouch performs the worst, and the other methods show largely consistent rankings. (2) Poorly performing vision-tactile encoders yield worse results than vision-

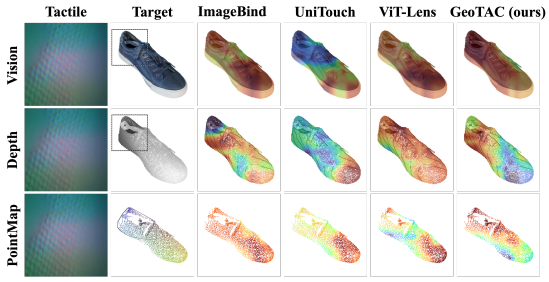

Figure 7: Attention visualization for a Touch3D sample across multisensory learning methods.

Table 3: Ablation study of GeoTAC on Touch3D.

Training configuration	Avg. ↑
Patch Tac	34.94
Patch Tac + CL	37.62
Patch Tac + CMA	40.42
Patch Tac + GD	39.40
GeoTAC	**42.10**

Table 4: Task performance in UniVTAC across multisensory learning algorithms. Parentheses show changes in average SR and TP relative to DINOv2 (Vision-only), in percentage points.

Encoder initialization	Lift Can		Lift Bottle		Grasp Classify		Insert Tube		Avg. (%)	
	SR	TP ↑	SR	TP ↑	SR	TP ↑	SR	TP ↑	SR ↑	TP ↑
DINOv2 (Vision-only)	2/40	7.34	0/40	40.01	24/40	73.02	0/40	88.08	16.25 (0.00)	52.11 (0.00)
DINOv2 (Vision-Tactile)	4/40	9.32	2/40	50.45	24/40	90.00	0/40	89.90	18.75 (+2.50)	59.92 (+7.81)
ImageBind (Girdhar et al., 2023)	5/40	9.78	4/40	55.07	40/40	100.00	1/40	85.81	31.25 (+15.00)	62.67 (+10.55)
UniTouch (Yang et al., 2024)	1/40	1.20	1/40	42.08	23/40	69.58	0/40	76.15	15.63 (−0.63)	47.25 (−4.86)
ViT-Lens (Lei et al., 2024)	2/40	5.78	2/40	49.62	24/40	71.95	0/40	77.70	17.50 (+1.25)	51.26 (−0.85)
Patch Tac + CL	6/40	10.08	4/40	65.51	40/40	100.00	8/40	90.77	36.25 (+20.00)	66.59 (+14.48)
GeoTAC (Ours)	6/40	20.01	5/40	72.38	40/40	100.00	12/40	92.63	39.38 (+23.13)	71.26 (+19.14)

only encoders. UniTouch and ViT-Lens perform worse than Tac + CL in Figure 6, and policies using their trained encoders perform worse than DINOv2 (Vision-only) in Table 4. (3) Patch-aligned multisensory learning improves policy performance in simulation: Patch Tac + CL and GeoTAC improve average TP by 6.67 and 11.34 percentage points over DINOv2 (Vision-Tactile), respectively.

5.6 Ablation Study

We conduct ablation studies on the utility of individual modules, specifically comparing performance when adding Cross-Modal Alignment (row **Patch Tac + CMA**) and Geometry Distillation (row **Patch Tac + GD**) separately. Table 3 reports the average vision-tactile alignment score (the same as *Avg.* in Table 2). Building on patch alignment with contrastive learning (row **Patch Tac + CL**), adding CMA and GD improves the score by 2.8 and 1.78, respectively, with CMA providing a larger gain than GD. Combining both modules yields an improvement of 4.48 over the baseline.

6 Limitations

Despite our extensive efforts to collect 3D vision-tactile data and explore multisensory learning methods, several limitations remain: (1) limited aligned multisensory data; (2) challenges in unified learning across sensors; and (3) evaluation of utility to real-world robotic tasks.

7 Conclusion

We formalize tactile alignment with other modalities and build a 3D vision-tactile collection system, acquiring over 20K tactile samples from 505 objects. We extend existing tactile benchmarks from object- to patch-level alignment and propose GeoTAC to learn tactile representations from aligned multisensory data. The proposed methods exhibit consistent improvements in representation learning and robotic tasks.

AI USE STATEMENT

In this work, we used generative AI to correct grammar, polish sentences, and assist with figure creation.

REFERENCES

Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*, 2017.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video, 2024. URL https://arxiv.org/abs/2404.08471.

Tapomayukh Bhattacharjee, Advait Jain, Sarvagya Vaish, Marc D Killpack, and Charles C Kemp. Tactile sensing over articulated joints with stretchable sensors. In *2013 World Haptics Conference (WHC)*, pp. 103–108. IEEE, 2013.

Raunaq Bhirangi, Venkatesh Pattabiraman, Enes Erciyes, Yifeng Cao, Tess Hellebrekers, and Lerrel Pinto. Anyskin: Plug-and-play skin sensing for robotic touch. *arXiv preprint arXiv:2409.08276*, 2024.

Raunaq Bhirangi, Venkatesh Pattabiraman, Enes Erciyes, Yifeng Cao, Tess Hellebrekers, and Lerrel Pinto. Anyskin: Plug-and-play skin sensing for robotic touch. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 16563–16570. IEEE, 2025.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021. URL https://arxiv.org/abs/2104.14294.

Baijun Chen, Weijie Wan, Tianxing Chen, Xianda Guo, Congsheng Xu, Yuanyang Qi, Haojie Zhang, Longyan Wu, Tianling Xu, Zixuan Li, et al. Univtac: A unified simulation platform for visuo-tactile manipulation data generation, learning, and benchmarking. *arXiv preprint arXiv:2602.10093*, 2026.

Ning Cheng, Changhao Guan, Jing Gao, Weihao Wang, You Li, Fandong Meng, Jie Zhou, Bin Fang, Jinan Xu, and Wenjuan Han. Touch100k: A large-scale touch-language-vision dataset for touch-centric multimodal representation, 2024. URL https://arxiv.org/abs/2406.03813.

Zhengxue Cheng, Yan Zhao, Keyu Wang, Li Song, and Hengdi ZHANG. Taco: A benchmark for lossless and lossy codecs of heterogeneous tactile data. In *The Fourteenth International Conference on Learning Representations*, 2026.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Paolo Cignoni, Guido Ranzuglia, M Callieri, M Corsini, F Ganovelli, N Pietroni, M Tarini, et al. Meshlab, 2011.

Samuel Clarke, Suzannah Wistreich, Yanjie Ze, and Jiajun Wu. X-capture: An open-source portable device for multi-sensory learning. *arXiv preprint arXiv:2504.02318*, 2025a.

Samuel Clarke, Suzannah Wistreich, Yanjie Ze, and Jiajun Wu. X-capture: An open-source portable device for multi-sensory learning, 2025b. URL https://arxiv.org/abs/2504.02318.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13142–13153, 2023.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024.

Siyuan Dong, Wenzhen Yuan, and Edward H Adelson. Improved gelsight tactile sensor for measuring geometry and slip. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 137–144. IEEE, 2017.

Yiming Dou, Fengyu Yang, Yi Liu, Antonio Loquercio, and Andrew Owens. Tactile-augmented radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26529–26539, 2024.

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, 2025. URL https://arxiv.org/abs/2502.12191.

Ruoxuan Feng, Yuxuan Zhou, Siyu Mei, Dongzhan Zhou, Pengwei Wang, Shaowei Cui, Bin Fang, Guocai Yao, and Di Hu. Anytouch 2: General optical tactile representation learning for dynamic tactile perception. *arXiv preprint arXiv:2602.09617*, 2026.

Ruohan Gao, Yen-Yu Chang, Shivani Mall, Li Fei-Fei, and Jiajun Wu. Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations. *arXiv preprint arXiv:2109.07991*, 2021.

Ruohan Gao, Zilin Si, Yen-Yu Chang, Samuel Clarke, Jeannette Bohg, Li Fei-Fei, Wenzhen Yuan, and Jiajun Wu. Objectfolder 2.0: A multisensory object dataset for sim2real transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10598–10608, 2022.

Ruohan Gao, Yiming Dou, Hao Li, Tanmay Agarwal, Jeannette Bohg, Yunzhu Li, Li Fei-Fei, and Jiajun Wu. The objectfolder benchmark: Multisensory learning with neural and real objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17276–17286, 2023.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190. IEEE, 2023.

Oliver Glauser, Daniele Panozzo, Otmar Hilliges, and Olga Sorkine-Hornung. Deformation capture via soft and stretchable sensor arrays. *ACM Transactions on Graphics (TOG)*, 38(2):1–16, 2019.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

Irmak Guzey, Ben Evans, Soumith Chintala, and Lerrel Pinto. Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play. *arXiv preprint arXiv:2303.12076*, 2023.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. URL https://arxiv.org/abs/2111.06377.

Nicholas Hedger, Thomas Naselaris, Kendrick Kay, and Tomas Knapen. Vicarious body maps bridge vision and touch in the human brain. *Nature*, 650(8100):173, 2025.

Liang Heng, Haoran Geng, Kaifeng Zhang, Pieter Abbeel, and Jitendra Malik. Vitacformer: Learning cross-modal representation for visuo-tactile dexterous manipulation, 2025. URL https://arxiv.org/abs/2506.15953.

Carolina Higuera, Sergio Arnaud, Byron Boots, Mustafa Mukadam, Francois Robert Hogan, and Franziska Meier. Visuo-tactile world models. *arXiv preprint arXiv:2602.06001*, 2026.

Binghao Huang, Yixuan Wang, Xinyi Yang, Yiyue Luo, and Yunzhu Li. 3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing, 2025a. URL https://arxiv.org/abs/2410.24091.

Jialei Huang, Shuo Wang, Fanqi Lin, Yihang Hu, Chuan Wen, and Yang Gao. Tactile-vla: unlocking vision-language-action model's physical knowledge for tactile generalization. *arXiv preprint arXiv:2507.09160*, 2025b.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

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 Robotics and Automation Letters*, 5(3):3838–3845, 2020. doi: 10.1109/LRA.2020.2977257.

Weixian Lei, Yixiao Ge, Kun Yi, Jianfeng Zhang, Difei Gao, Dylan Sun, Yuying Ge, Ying Shan, and Mike Zheng Shou. ViT-Lens: Towards omni-modal representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26647–26657, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Lei_ViT-Lens_Towards_Omni-modal_Representations_CVPR_2024_paper.html.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023a.

Yong-Lu Li, Yue Xu, Xinyu Xu, Xiaohan Mao, Yuan Yao, Siqi Liu, and Cewu Lu. Beyond object recognition: A new benchmark towards object concept learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20029–20040, 2023b.

Yong-Lu Li, Xiaoqian Wu, Xinpeng Liu, Zehao Wang, Yiming Dou, Yikun Ji, Junyi Zhang, Yixing Li, Xudong Lu, Jingru Tan, et al. From isolated islands to pangea: Unifying semantic space for human action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16582–16592, 2024.

Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting touch and vision via cross-modal prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10609–10618, 2019.

Haohong Lin, Radu Corcodel, and Ding Zhao. Generalize by touching: Tactile ensemble skill transfer for robotic furniture assembly. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9227–9233. IEEE, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Duc Huy Nguyen, Tim Schneider, Guillaume Duret, Alap Kshirsagar, Boris Belousov, and Jan Peters. Tacex: Gelsight tactile simulation in isaac sim–combining soft-body and visuotactile simulators. *arXiv preprint arXiv:2411.04776*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Venkatesh Pattabiraman, Yifeng Cao, Siddhant Haldar, Lerrel Pinto, and Raunaq Bhirangi. Learning precise, contact-rich manipulation through uncalibrated tactile skins. *arXiv preprint arXiv:2410.17246*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021a. URL https://arxiv.org/abs/2103.00020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021b.

Jieji Ren, Jiang Zou, and Guoying Gu. Mc-tac: Modular camera-based tactile sensor for robot gripper. In *International Conference on Intelligent Robotics and Applications*, pp. 169–179. Springer, 2023.

Samanta Rodriguez, Yiming Dou, William van den Bogert, Miquel Oller, Kevin So, Andrew Owens, and Nima Fazeli. Contrastive touch-to-touch pretraining. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5857–5863. IEEE, 2025.

Branden Romero, Hao-Shu Fang, Pulkit Agrawal, and Edward Adelson. Eyesight hand: Design of a fully-actuated dexterous robot hand with integrated vision-based tactile sensors and compliant actuation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1853–1860. IEEE, 2024.

Andrea Sipos, William van den Bogert, and Nima Fazeli. Gelslim 4.0: Focusing on touch and reproducibility, 2024. URL https://arxiv.org/abs/2409.19770.

Stefano Stassi, Valentina Cauda, Giancarlo Canavese, and Candido Fabrizio Pirri. Flexible tactile sensing based on piezoresistive composites: A review. *Sensors*, 14(3):5296–5332, 2014.

Subramanian Sundaram, Petr Kellnhofer, Yunzhu Li, Jun-Yan Zhu, Antonio Torralba, and Wojciech Matusik. Learning the signatures of the human grasp using a scalable tactile glove. *Nature*, 569 (7758):698–702, 2019.

Sudharshan Suresh, Haozhi Qi, Tingfan Wu, Taosha Fan, Luis Pineda, Mike Lambeta, Jitendra Malik, Mrinal Kalakrishnan, Roberto Calandra, Michael Kaess, et al. 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 *2022 International Conference on Robotics and Automation (ICRA)*, pp. 10781–10787. IEEE, 2022.

Chenxi Wang, Hongjie Fang, Hao-Shu Fang, and Cewu Lu. Rise: 3d perception makes real-world robot imitation simple and effective. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2870–2877. IEEE, 2024.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025.

Te-Yen Wu, Lu Tan, Yuji Zhang, Teddy Seyed, and Xing-Dong Yang. Capacitivo: Contact-based object recognition on interactive fabrics using capacitive sensing. In *Proceedings of the 33rd annual acm symposium on user interface software and technology*, pp. 649–661, 2020.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21469–21480, 2025.

Daniel Xu, Andreas Tairych, and Iain A Anderson. Stretch not flex: programmable rubber keyboard. *Smart Materials and Structures*, 25(1):015012, 2016.

Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. *arXiv preprint arXiv:2211.12498*, 2022.

Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, et al. Binding touch to everything: Learning unified multimodal tactile representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26340–26353, 2024.

Yang You, Kai Xiong, Zhening Yang, Zhengxiang Huang, Junwei Zhou, Ruoxi Shi, Zhou Fang, Adam W Harley, Leonidas Guibas, and Cewu Lu. Pace: A large-scale dataset with pose annotations in cluttered environments. In *European Conference on Computer Vision*, pp. 473–489. Springer, 2024.

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

Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017b.

Ying Yuan, Haichuan Che, Yuzhe Qin, Binghao Huang, Zhao-Heng Yin, Kang-Won Lee, Yi Wu, Soo-Chul Lim, and Xiaolong Wang. Robot synesthesia: In-hand manipulation with visuotactile sensing. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6558–6565. IEEE, 2024.

Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10394–10403, 2019.

A More Details on Touch3D

This section describes the annotation interface and object distribution of Touch3D, which contains 505 physical objects and 20,025 tactile contacts.

A.1 Data Inspection and Annotation

Figure 8 shows the interface used to inspect collected samples and assign annotations. The left column selects the object category, instance, and tactile file, provides navigation between samples, and records material and 3D-part labels. The center and right panels display the tactile frame and video, the corresponding RGB view, and the scanned 3D part. Viewing these observations together helps operators check contact correspondence and material labels. As described in the main paper, incorrect geometry or contact annotations are corrected, while samples with missing or damaged observations are removed.

A.2 Object Coverage and Distribution

Figure 9 compares the object names represented in Touch3D and ObjectFolder Real. The three groups contain 22 categories unique to ObjectFolder Real, 13 shared categories, and 83 categories unique to Touch3D. Shared categories include bowls, spoons, cups, and plates, while Touch3D additionally covers objects such as slippers, handbags, toy cars, and toothbrushes. These counts refer to distinct object names in the comparison, not individual physical instances.

Figure 10 complements this comparison with instance counts for each object category. Blue and orange bars represent ObjectFolder Real and Touch3D, respectively. Categories are ordered by their combined counts and divided into three panels with different horizontal scales. Both datasets have uneven category frequencies: a few categories account for many instances, while numerous others have only one or a few. Thus, the balanced coverage of the six collection scenarios described in the main paper does not imply balanced counts for every object category.

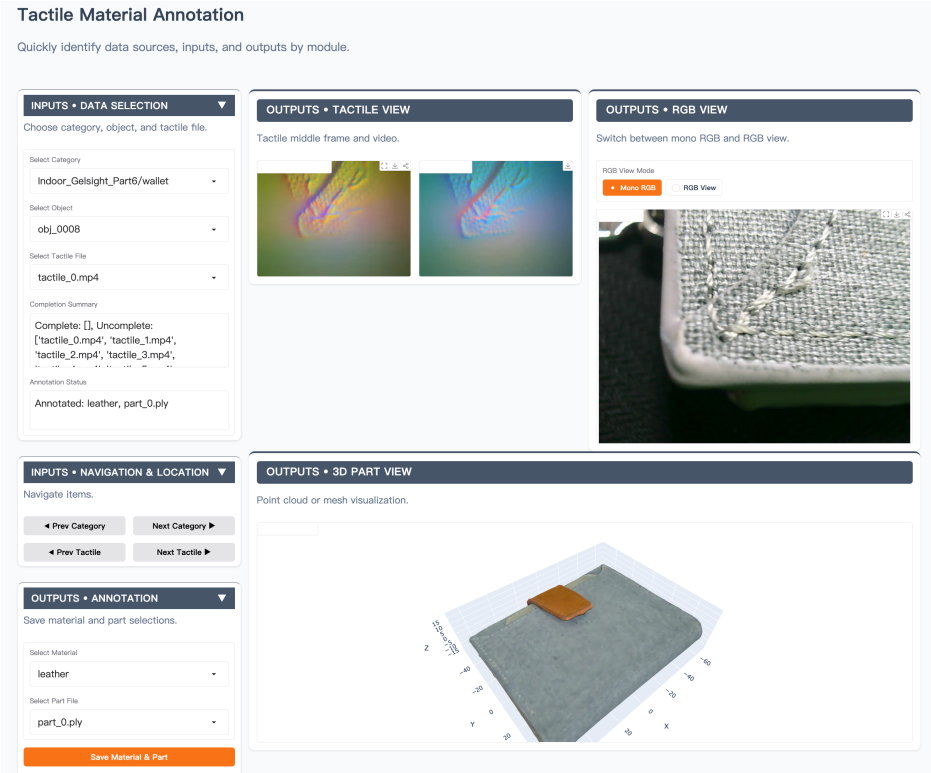

Figure 8: **Data inspection and annotation interface.** Sample selection, navigation, and annotation controls are shown on the left; tactile, RGB, and 3D-part views support joint inspection of each contact.

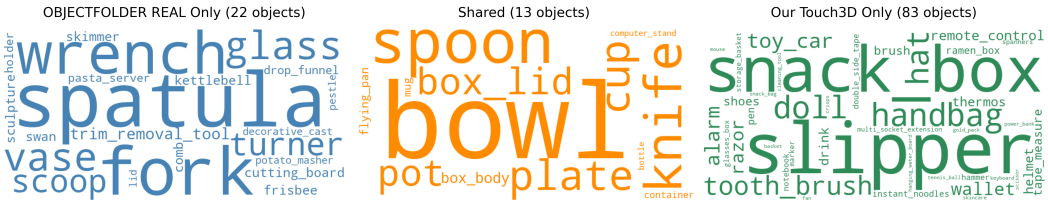

Figure 9: **Object Category Coverage.** From left to right: names unique to ObjectFolder Real, shared by both datasets, and unique to Touch3D. The panel counts describe distinct names rather than physical object instances.

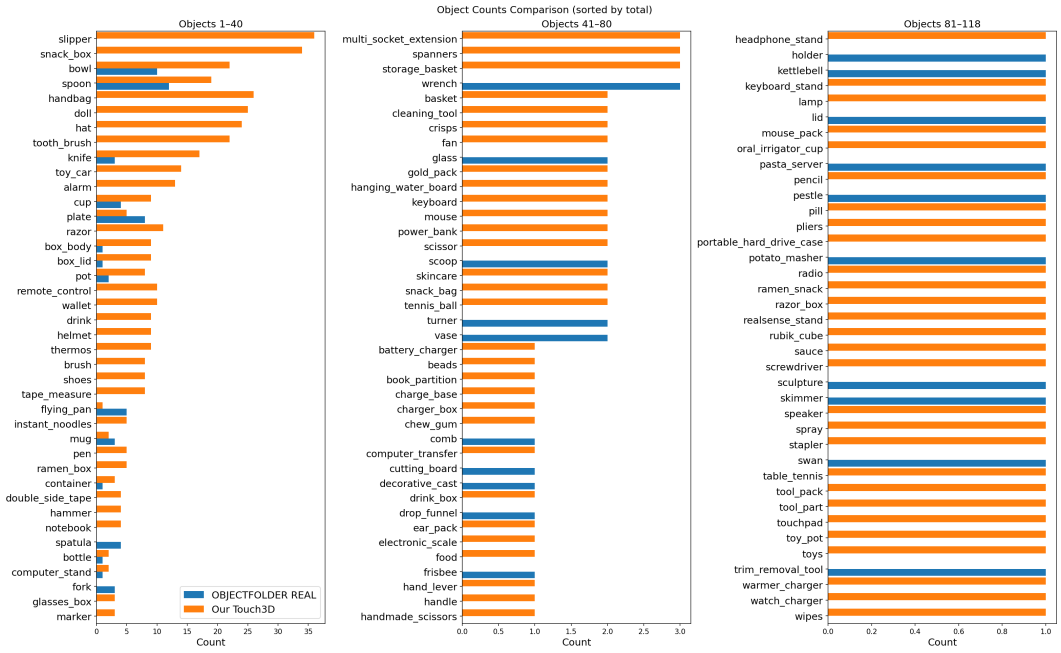

Figure 10: **Object-instance counts in Touch3D and ObjectFolder Real.** Categories are sorted by combined count across both datasets. The three panels use different horizontal scales to show frequent and infrequent categories.

B Additional Evaluation Details

B.1 Representation Learning and Evaluation

We use DINOv2 ViT-Large for the patch-alignment experiments. Touch3D is split into training, validation, and test sets at a ratio of 4:1:1, and ObjectFolder Real provides an additional test set. Representation training uses Touch3D; evaluation on ObjectFolder Real examines transfer to data collected with a different tactile sensor.

Object-level evaluation measures category correspondence between a local tactile query and global visual candidates using k-NN Top-5 accuracy. Patch-level evaluation instead requires the candidate to originate from the same contact and reports exact-contact Top-5 retrieval accuracy (E@5), averaged over both directions. Matching the object category alone is therefore insufficient for patch-level success. Material prediction uses k-NN mAP with multiple material labels, while object-category classification uses probe accuracy. The average in Table 2 and Table 3 is the unweighted mean of these four vision-tactile evaluation metrics.

B.2 Robotic Manipulation Evaluation

We evaluate transfer to Diffusion Policy in the UniVTAC simulator using Lift Can, Lift Bottle, Grasp Classify, and Insert Tube. Each task uses 100 demonstration episodes and 30,000 policy-training steps, followed by 40 evaluation rollouts. The vision-only baseline excludes tactile observations; the vision-tactile methods use both modalities with their corresponding encoder initializations.

In Table 4, task-level SR is the number of successful rollouts out of 40. TP is the mean final task progress across all 40 rollouts, including unsuccessful ones. Average SR and TP weight the four tasks equally and are computed before rounding. Parenthesized changes are absolute percentage-point differences from DINOv2 (Vision-only), not relative percentage gains. These results measure simulated manipulation rather than physical-robot deployment.

C Limitations and Future Directions

Data Scale and Alignment Quality. Collecting paired tactile observations and scanned geometry remains costly, and category coverage is uneven. Generated geometry (Wang et al., 2025; Xiang et al., 2025), relabeled object-level data, and tactile simulation (Nguyen et al., 2024) could expand the training set, but their contact correspondences require quality control before being used as patch-level supervision.

Cross-Sensor Generalization. Differences in sensor appearance, geometry, and response can limit transfer between datasets. Learning from multiple tactile sensors and harmonizing contact annotations are important directions toward a shared tactile representation.

Physical-Robot Transfer. The current manipulation evaluation is conducted in simulation. Physical-robot experiments are still needed to assess robustness to real sensor noise, calibration errors, and variations in contact. They would also help determine when improved representation metrics translate into more reliable manipulation.