TouchAnything: A Dataset and Framework for Bimanual Tactile Estimation from Egocentric Video

Jianyi Zhou[1], Ziteng Gao[1], Feiyang Hong[1], Zirui Liu[1], Guannan Zhang[1], Weisheng Dai[1], Ruichen Zhen[2], Chuqiao Lyu[3], Haotian Wu[2], Yinian Mao[2], Xushi Wang[1], Yuxiang Jiang[1], Wenbo Ding[3], Shuo Yang[1]✉

[1]Harbin Institute of Technology, Shenzhen
[2]Meituan Academy of Robotics
[3]Tsinghua Shenzhen International Graduate School, Tsinghua University

arXiv:2605.13083v1 [cs.RO] 13 May 2026

Abstract

Egocentric human video data, which captures rich human-environment interactions and can be collected at scale, has become a key driver of embodied intelligence research. However, existing egocentric datasets typically lack tactile sensing, a critical modality that provides direct cues about contact, force, and pressure in human-object interaction. Without such signals, models struggle to learn physically grounded representations of real-world interaction dynamics. While tactile sensors provide these cues, deploying high-quality tactile hardware at scale remains expensive and cumbersome. This raises a central question: can tactile feedback be inferred directly from visual observations, enabling scalable tactile supervision for egocentric video data and supporting physically grounded embodied learning? To enable research in this direction, we introduce **EgoTouch**, a large-scale multi-view egocentric dataset with dense tactile supervision for bimanual hand-object interaction. EgoTouch comprises 208 manipulation tasks spanning 1,891 episodes in diverse indoor and outdoor environments, with synchronized multi-view RGB (head-mounted egocentric and dual wrist-mounted cameras), bimanual 3D hand pose, and continuous pressure maps from wearable tactile sensors. Building on EgoTouch, we introduce **TouchAnything**, a baseline multi-view vision-to-touch prediction framework that uses the egocentric view as the primary input and flexibly leverages available wrist-mounted views at inference time. Experiments show that incorporating wrist-mounted views generally improves tactile prediction over egocentric-only input, achieving up to 5.0% relative improvement in Contact IoU and 6.1% relative improvement in Volumetric IoU. We will publicly release the dataset, code, and benchmark.

Date: May 14, 2026
Correspondence: shuoyang@hit.edu.cn
Project Page: https://jianyi2004.github.io/TouchAnything-Website/

1 Introduction

Egocentric video datasets have become a key source of scalable supervision for embodied intelligence, as they are relatively easy to collect, capture natural human-environment interactions, and cover diverse real-world manipulation behaviors [8, 10, 11, 14, 21, 35, 38, 39]. However, despite this progress, a critical modality remains largely missing: tactile sensing. While visual observations capture appearance and motion cues, they do not directly reveal the physical signals underlying successful manipulation, such as contact force and pressure distribution. Without access to such signals, embodied models lack direct supervision of the physical interaction dynamics that govern real-world manipulation [19, 32], limiting their ability to develop a deeper understanding of the physical world [27, 30]. Although tactile sensors can provide this information, collecting

Figure 1 EgoTouch combines egocentric and wrist-mounted views with synchronized 3D hand pose and dense tactile pressure maps, providing complementary visual evidence for learning contact-aware interactions.

large-scale tactile data with high-quality hardware is expensive, intrusive, and difficult to scale. This creates a fundamental bottleneck: the availability of large-scale visual data contrasts sharply with the scarcity of tactile supervision. This gap raises a key question: can tactile signals be inferred directly from visual observations, enabling scalable tactile supervision for large-scale egocentric data and supporting interaction-aware embodied learning?

Vision-to-touch prediction has emerged as a promising direction [18, 31, 34], but progress remains fundamentally constrained by data. Existing datasets either rely on single-view capture [34] or focus on relatively narrow interaction settings such as hand-surface contact or single-finger pressing [9]. As a result, they provide limited support for studying tactile prediction in realistic bimanual hand-object interactions, which involve diverse manipulation contexts and frequent occlusion of hand-object contact regions. This occlusion is a central challenge in egocentric vision-to-touch prediction: contact regions are often hidden by the hand itself or the manipulated object, making tactile signals only partially observable from the head-mounted view. This missing contact evidence introduces substantial ambiguity, especially in complex manipulation scenarios.

To address these challenges, we introduce **EgoTouch** (as shown in Figure 1), a large-scale multi-view egocentric dataset with dense tactile supervision for bimanual hand-object interaction. EgoTouch pairs egocentric and wrist-mounted videos with synchronized tactile pressure maps and bimanual hand pose, providing complementary views of realistic hand-object interactions. It contains 208 manipulation tasks, 1,891 episodes, and over 20 hours of interaction data across diverse indoor and outdoor environments, supporting cross-modal learning of physical interaction dynamics.

Building on EgoTouch, we establish **TouchAnything**, a baseline vision-to-touch prediction model that supports flexible inference with single or multiple camera views. A key design choice is view dropout: during training, wrist camera views are randomly dropped, forcing the model to learn robust representations that work with any subset of available views at inference time. This enables deployment in settings where only an egocentric camera is available, while gracefully leveraging additional views when present. Experiments show that adding wrist-mounted views improves tactile prediction on both seen and unseen objects: the full multi-view setting improves Contact IoU from 0.4792 to 0.5030 and Volumetric IoU from 0.4311 to 0.4575 on seen objects, and improves Contact IoU from 0.4396 to 0.4496 and Volumetric IoU from 0.3743 to 0.3852 on unseen objects.

In summary, our contributions are:

1. We introduce **EgoTouch**, a large-scale multi-view egocentric dataset for bimanual hand-object interaction, comprising 208 tasks, 1,891 episodes, synchronized RGB videos from one head-mounted camera and two

Table 1 Comparison of EgoTouch with existing hand interaction and tactile datasets. EgoTouch is the first to jointly provide multi-view video, bimanual hand pose, and real dense pressure data across diverse scenes.

Dataset	In-the-wild	Hand Pose	Contact	Wrist Views	Hands	Objects	Frames
GRAB [28]	✗	MoCap	Analytical	✗	Biman.	51	1.6M
ContactDB [3]	✗	✗	Thermal	✗	Biman.	50	375k
ARCTIC [7]	✗	MoCap	Analytical	✗	Biman.	11	2.1M
OakInk [33]	✗	MoCap	Analytical	✗	Single	100	230k
DexYCB [4]	✗	Est.	Analytical	✗	Single	20	582k
ActionSense [6]	✗	Glove	Pressure	✗	Biman.	21	521k
HOI4D [20]	✗	Est.	✗	✗	Single	800	2.4M
EgoPressure [9]	✗	Est.	Pressure	✗	Single	31	4.3M
EgoDex [14]	✗	Est.	Analytical	✗	Biman.	500	90M
OpenTouch [25]	✓	Glove	Pressure	✗	Single	800	∼ 500k
EgoTouch (Ours)	✓	**Glove + Est.**	**Pressure**	✓	**Biman.**	**∼1000**	**2.1M**

wrist-mounted cameras, bimanual 3D hand pose, and dense continuous pressure maps across diverse environments.

2. We establish **a multi-view vision-to-touch benchmark** on EgoTouch, with evaluation protocols for seen and unseen objects and different camera-view configurations, enabling systematic analysis of how complementary wrist views affect tactile prediction.

3. We propose **TouchAnything**, a baseline vision-to-touch model with cross-view fusion and view dropout training, supporting flexible inference with egocentric-only or multi-view inputs and improving tactile prediction when wrist views are available.

2 EgoTouch Dataset

The EgoTouch dataset contains 20 hours of multi-view egocentric video collected at 30 Hz, comprising 1,891 episodes across 208 diverse manipulation tasks. This amounts to approximately 2.1 million frames covering over 1,000 objects in both indoor and outdoor environments. The dataset provides rich and structured annotations, including synchronized multi-view RGB videos, bimanual 3D hand pose, and dense tactile pressure maps for both hands. All modalities are temporally aligned at the frame level to enable precise cross-modal learning. We compare EgoTouch with existing hand-object interaction and tactile datasets in Table 1.

2.1 Data Collection Setup

EgoTouch is collected with a synchronized wearable capture system that records complementary visual, kinematic, and tactile signals during natural bimanual manipulation (Figure 2). The setup includes a head-mounted RGB camera for global egocentric context, two wrist-mounted RGB cameras for close-up observations of hand-object contact regions, Rokoko motion-capture gloves for bimanual 3D hand pose, custom pressure-sensing gloves with 16×16 tactile arrays on each palm, and HTC Vive Trackers for 6-DoF head and wrist localization. All streams are synchronized onto a shared 30Hz timeline using timestamps and latest valid sensor snapshots. This produces frame-level synchronized multi-view RGB, hand pose, tactile pressure maps, and tracker poses. Figure 3 shows representative synchronized observations from the dataset. Additional hardware specifications, acquisition details, and synchronization strategy are provided in Appendix B.1.

2.2 Data Modalities

Each frame in EgoTouch is organized on a synchronized 30Hz timeline and contains the following modalities:

- **Multi-view RGB videos.** The dataset provides three egocentric RGB views: a head-mounted camera capturing the global manipulation scene ($V^h \in \mathbb{R}^{640\times480\times3}$), and two wrist-mounted fisheye cameras

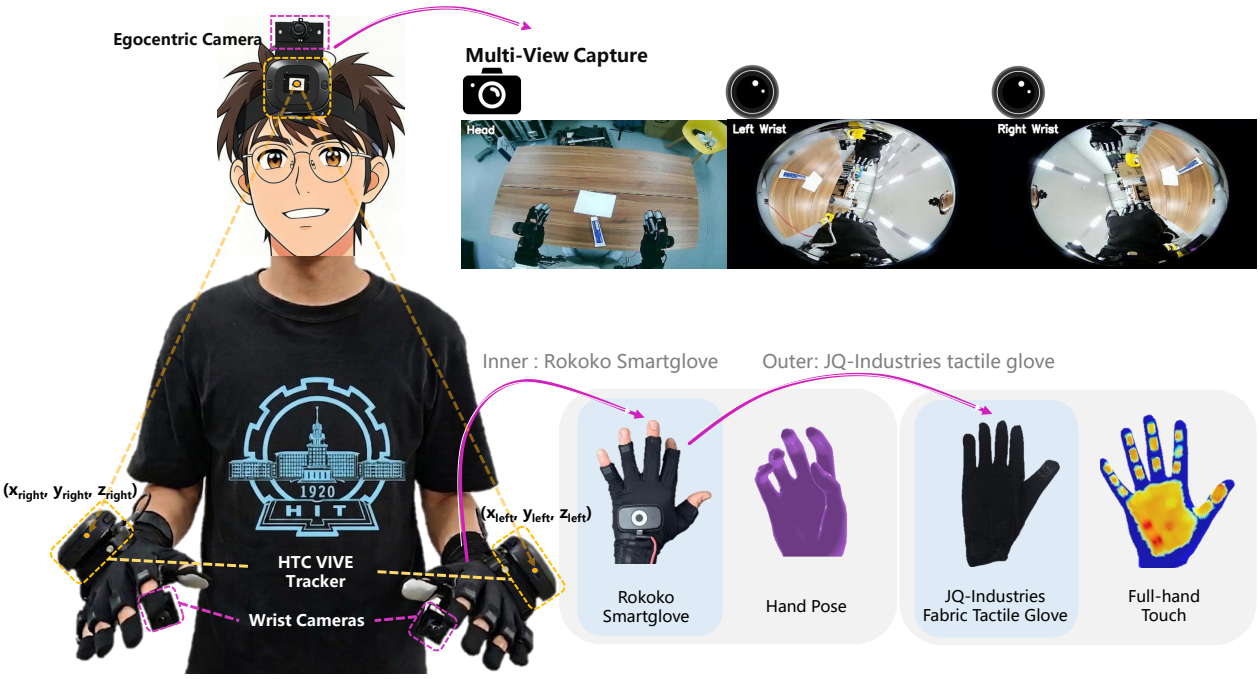


Figure 2 Data collection setup and example multi-modal data. The participant wears a head-mounted egocentric camera, two wrist-mounted cameras, and pressure-sensing gloves. All modalities are temporally synchronized.

($V^{wL}, V^{wR} \in \mathbb{R}^{640\times480\times3}$) providing close-up observations of hand-object contact regions during bimanual interactions.

- **Bimanual 3D hand pose.** Hand kinematics are represented by 42 three-dimensional joints ($\mathbf{P} \in \mathbb{R}^{42\times3}$), including wrist, finger, and fingertip joints for both hands.
- **Tactile pressure maps.** Dense tactile feedback is recorded as bilateral 16×16 raw pressure arrays ($\mathbf{M}_{raw} \in \mathbb{R}^{2\times16\times16}$), which are normalized and remapped into canonical 21×21 hand-shaped grids ($\mathbf{M} \in \mathbb{R}^{2\times21\times21}$) for training. Details are provided in Appendix B.2.
- **Tracker poses and metadata.** Each frame additionally includes 6-DoF poses from HTC Vive Trackers mounted on the head and wrists, together with metadata annotations such as task category, object category, scene description, and environment type.

All modalities are temporally aligned at the frame level to support cross-modal learning of physical interaction dynamics.

2.3 Task Taxonomy

The 208 tasks in EgoTouch are grouped into five environment-based categories that capture diverse real-world interaction patterns (Figure 4):

- **Home**: Everyday household interactions such as opening containers, pushing and pulling objects, pressing switches, wiping surfaces, folding clothes, and handling daily items.
- **Workbench**: Tool-based manipulation tasks including gripping and turning tools, sawing, drilling, sanding, cutting, clamping, and precision assembly interactions.
- **Office**: Workspace activities such as swiping cards, typing on keyboards, operating office tools, and manipulating books or stationery.

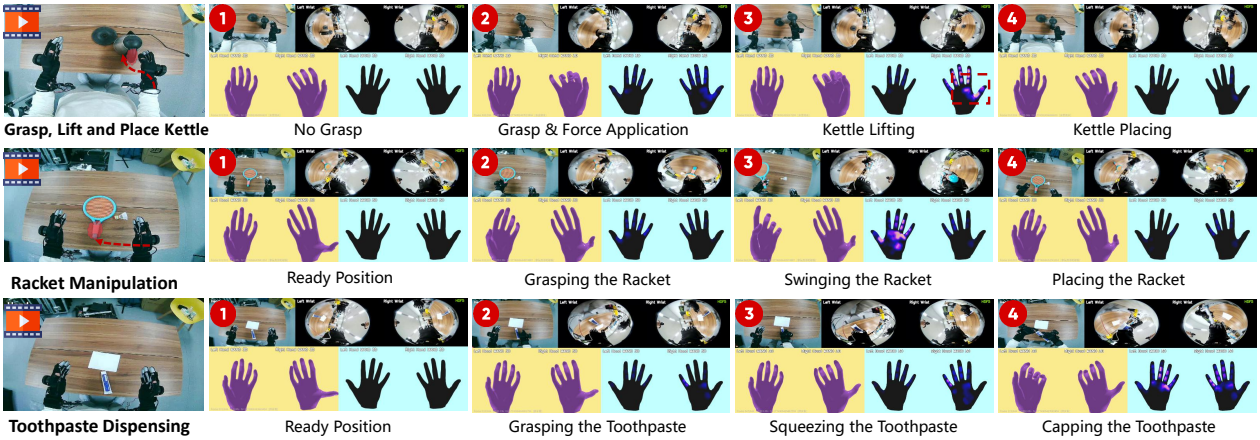


Figure 3 Example data from EgoTouch demonstrates that hardware-based tactile sensing and pose tracking reveal critical force, contact, and motion cues that vision alone cannot capture.

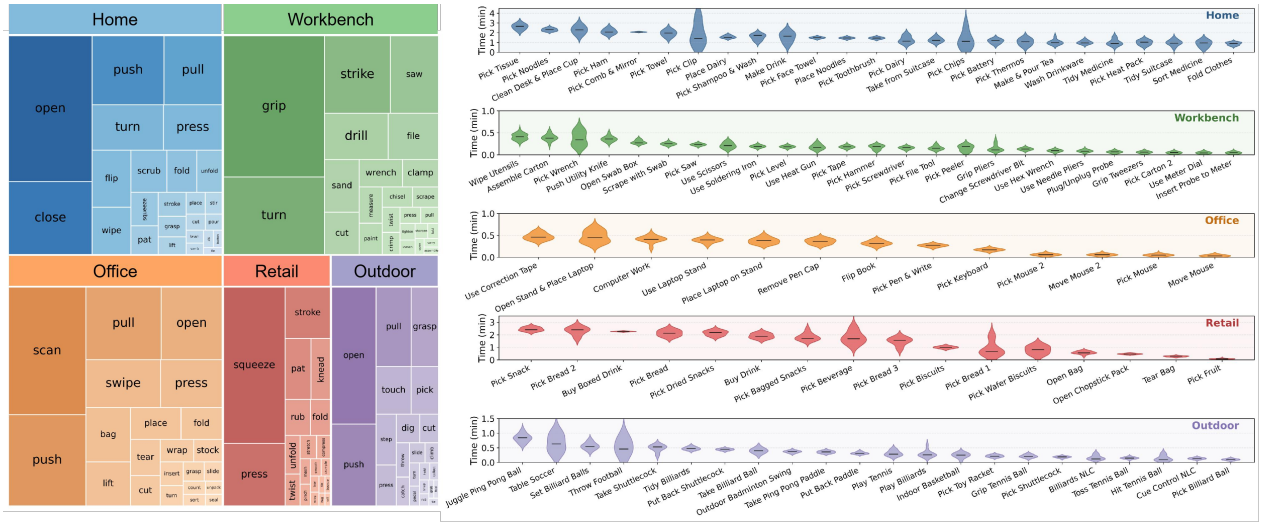


Figure 4 Dataset statistics and analysis. EgoTouch exhibits broad task coverage across five environment categories.

- **Retail**: Consumer interaction behaviors including squeezing products, pressing packaged items, folding goods, opening bags, and handling snacks or beverages.
- **Outdoor**: Dynamic open-environment interactions including ball games, racket sports, outdoor object handling, and other full-body coordinated manipulation activities.

3 METHOD: Tactile Prediction Framework

We introduce a baseline framework for multi-view tactile prediction that maps visual observations and hand pose to dense bimanual pressure maps. The framework is designed to (1) leverage complementary viewpoints to address occlusion, and (2) support flexible inference under missing views.

3.1 Problem Formulation

Given a video clip of T frames from a subset of views $\mathcal{V} \subseteq \{V^{ego}, V^{wL}, V^{wR}\}$ and the corresponding bimanual hand pose sequence $\mathbf{P} \in \mathbb{R}^{T\times 42\times 3}$, our goal is to predict bilateral tactile maps $\hat{\mathbf{M}} \in \mathbb{R}^{T\times 2\times 21\times 21}$ at each timestep, where the tactile maps are represented in a canonical 21×21 hand-shaped grid after preprocessing

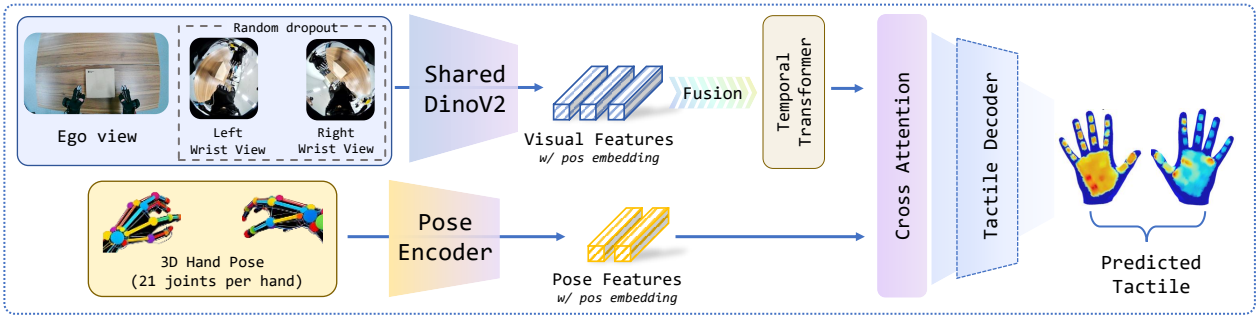


Figure 5 Architecture of the multi-view tactile prediction model. A shared backbone encodes each view with view embeddings. Cross-view attention and gated fusion produce unified visual features, which are combined with hand pose through pose-aware fusion and decoded into bilateral pressure maps.

and spatial remapping of the raw tactile sensor layout (Appendix B.2).

3.2 Multi-View Tactile Prediction Framework

Our framework consists of a shared visual encoder, a cross-view fusion module, a pose-aware fusion mechanism, and a tactile decoder, enabling joint modeling of appearance, geometry, and motion cues.

Each input view is processed independently using a shared visual backbone, followed by a learnable view embedding to encode camera identity. This allows the model to distinguish between egocentric and wrist-mounted perspectives while maintaining parameter efficiency. To integrate information across views, we apply a lightweight cross-view attention module over view-level features, enabling complementary reasoning across viewpoints. For example, wrist views can provide contact information that is occluded in the egocentric view. The fused representation is further aggregated using a gated mechanism that dynamically weighs the contribution of each view, ensuring robustness to missing or unreliable inputs.

To incorporate geometric information, we encode bimanual hand pose into joint-level features and fuse them with visual representations via cross-attention. Each joint attends to the most relevant visual regions, enabling spatially grounded reasoning about contact. This design allows the model to associate tactile signals with specific hand regions, which is critical for predicting structured pressure distributions. Temporal dependencies are modeled using a lightweight temporal module applied to the fused features, capturing interaction dynamics such as grasping and sliding.

The fused joint-level features are decoded into bilateral pressure maps for both hands. Each hand is predicted independently as a 21×21 pressure grid, representing normalized contact intensity. This formulation enables dense spatial supervision beyond binary contact prediction.

3.3 View Dropout and Training Objective

To support varying sensor configurations, we employ a view dropout strategy during training. The egocentric view is always retained, while wrist views are randomly dropped. This exposes the model to different input combinations and enables flexible inference at test time without architectural changes. The model can operate with only egocentric input, while benefiting from additional views when available.

The model is trained using a weighted regression loss that combines pixel-wise reconstruction with spatial regularization:

$$\mathcal{L} = \lambda_{mse}\mathcal{L}_{MSE} + \lambda_{l1}\mathcal{L}_{L1} + \lambda_{tv}\mathcal{L}_{TV}(\hat{\mathbf{M}}) \tag{1}$$

where $\mathcal{L}_{MSE}$ and $\mathcal{L}_{L1}$ measure pressure reconstruction error, and $\mathcal{L}_{TV}$ encourages spatial smoothness. To address the sparsity of tactile maps and prevent the model from collapsing to all-zero predictions, we apply higher loss weights to contact regions where pressure exceeds a threshold (0.1). We use $\lambda_{mse} = 1.0$, $\lambda_{l1} = 0.5$, $\lambda_{tv} = 0.01$, and a contact-region weight of 3.0.

4 Experiments

4.1 Experimental Setup

Dataset split. We split the dataset into training, validation, and test sets with a ratio of 80% / 10% / 10% at the episode level to avoid temporal data leakage. The test set is further divided into *seen-object* and *unseen-object* subsets to evaluate generalization to novel object instances. All splits cover a diverse set of interaction scenarios, ensuring variation in objects, tasks, and environments. This setup enables evaluation under both in-distribution and out-of-distribution settings.

Evaluation metric. We define the following benchmark tasks on EgoTouch: 1) Tactile prediction. Given multi-view video and hand pose, predict the bilateral pressure map. This is the primary benchmark. 2) Contact detection. A binarized version: predict whether each region is in contact (pressure $> \tau$). Derived from the tactile prediction output.

We evaluate with the following metrics, following PressureVision [34]:

- **Temporal Accuracy** ↑: Evaluates the temporal accuracy of contact onset and termination. If any contact is present in the estimated and ground truth contact maps, the frame is marked as in contact. A frame is marked correct if the presence of contact is consistent in estimated and ground truth frames.
- **Contact IoU** ↑: Evaluates the spatial and temporal accuracy of estimated contact by computing the intersection over union (IoU) between the binary contact images. This metric does not consider the magnitude of the estimated pressure, and is an upper bound on Volumetric IoU.
- **Volumetric IoU** ↑: Extends Contact IoU to evaluate the magnitudes of pressure estimates in addition to their spatial and temporal accuracy. Each 2D pressure image is converted into a 3D "pressure volume", where the height of the volume is equal to the amount of pressure at that pixel. The Volumetric IoU is calculated as:

$$IoU_{vol} = \frac{\sum^{i,j} min(P_{i,j}, \hat{P}_{i,j})}{\sum^{i,j} max(P_{i,j}, \hat{P}_{i,j})} \tag{2}$$

  where $P_{i,j}$ is the ground truth pressure at pixel (i, j) and $\hat{P}_{i,j}$ is the predicted pressure.
- **MAE** ↓: Mean absolute error over normalized pressure values. We calculate MAE over each pixel. As most of the dataset pressure images consist of zeros, these numbers are close to zero.

Implementation. We implement TouchAnything in PyTorch and train it on NVIDIA GPUs with distributed data parallelism. The visual encoder is a frozen DINOv2-Base (ViT-B/14) backbone initialized from a pretrained checkpoint. Each training sample consists of a clip of $T = 8$ frames sampled with a frame interval of 2 from three synchronized RGB views, including one egocentric view and two wrist-mounted views. All frames are resized to 224×224 and paired with 42 3D hand joints and bilateral 21×21 tactile maps normalized to $[0, 1]$. The model is trained for 25 epochs using AdamW with a learning rate of 5×10^{-5}, weight decay of 0.05, and a cosine learning-rate schedule with 10 warmup epochs. We optimize a contact-aware weighted reconstruction objective that combines MSE and L1 losses with a total-variation regularizer; tactile cells with pressure larger than 0.1 are assigned a weight of 3.0 to mitigate the sparsity of contact signals and discourage trivial all-zero predictions. During training, each wrist view is independently dropped with probability $p = 0.3$, while the egocentric view is always retained. This view-dropout strategy enables flexible inference with any available subset of camera views.

4.2 Main Results

Table 2 shows that wrist-mounted views generally improve tactile prediction over the egocentric-only baseline, particularly for Contact IoU, Volumetric IoU, and MAE. Overall, Ego + wL + wR improves Contact IoU from 0.4792 to 0.5030 and Volumetric IoU from 0.4311 to 0.4575 on seen objects, and from 0.4396 to 0.4496 and 0.3743 to 0.3852 on unseen objects, respectively. These gains indicate that wrist views provide complementary contact-region evidence, especially for localizing contact and estimating pressure magnitude under egocentric occlusion.

Table 2 Multi-view tactile prediction across five diverse scenarios. All methods use the same architecture and training recipe. Arrows (↑ / ↓) show relative percentage change vs. the Ego-only baseline.

Method	*Seen Objects*				*Unseen Objects*			
	T.Acc	C.IoU	V.IoU	MAE	T.Acc	C.IoU	V.IoU	MAE
Home Scenario								
Ego-only	0.8532	0.4869	0.4361	0.0470	0.8175	0.4250	0.3671	0.0604
Ego + wL	0.8659 ↑1.5%	0.5074 ↑4.2%	0.4598 ↑5.4%	0.0453 ↓3.6%	0.8278 ↑1.3%	0.4375 ↑2.9%	0.3779 ↑2.9%	0.0591 ↓2.2%
Ego + wR	0.8658 ↑1.5%	0.5072 ↑4.2%	0.4597 ↑5.4%	0.0453 ↓3.6%	0.8276 ↑1.2%	0.4372 ↑2.9%	0.3776 ↑2.9%	0.0592 ↓2.0%
Ego + wL + wR	**0.8659** ↑1.5%	**0.5075** ↑4.2%	**0.4598** ↑5.4%	**0.0452** ↓3.8%	**0.8267** ↑1.1%	**0.4371** ↑2.8%	**0.3775** ↑2.8%	**0.0590** ↓2.3%
Workbench Scenario								
Ego-only	0.7954	0.4131	0.3747	0.0420	0.8096	0.2932	0.2801	0.0599
Ego + wL	0.8250 ↑3.7%	0.4422 ↑7.0%	0.4080 ↑8.9%	0.0395 ↓6.0%	0.7525 ↓7.1%	0.3016 ↑2.9%	0.2841 ↑1.4%	0.0579 ↓3.3%
Ego + wR	0.8242 ↑3.6%	0.4408 ↑6.7%	0.4070 ↑8.6%	0.0395 ↓6.0%	0.7539 ↓6.9%	0.3017 ↑2.9%	0.2840 ↑1.4%	0.0579 ↓3.3%
Ego + wL + wR	**0.8241** ↑3.6%	**0.4414** ↑6.9%	**0.4078** ↑8.8%	**0.0395** ↓6.0%	**0.7538** ↓6.9%	**0.3015** ↑2.8%	**0.2840** ↑1.4%	**0.0578** ↓3.5%
Office Scenario								
Ego-only	0.8770	0.4918	0.4501	0.0436	0.8901	0.5743	0.4511	0.0632
Ego + wL	0.8990 ↑2.5%	0.5249 ↑6.7%	0.4889 ↑8.6%	0.0401 ↓8.0%	0.9035 ↑1.5%	0.5718 ↓0.4%	0.4673 ↑3.6%	0.0615 ↓2.7%
Ego + wR	0.9001 ↑2.6%	0.5250 ↑6.8%	0.4891 ↑8.7%	0.0400 ↓8.3%	0.9035 ↑1.5%	0.5718 ↓0.4%	0.4672 ↑3.6%	0.0615 ↓2.7%
Ego + wL + wR	**0.9003** ↑2.7%	**0.5256** ↑6.9%	**0.4895** ↑8.8%	**0.0399** ↓8.5%	**0.9035** ↑1.5%	**0.5718** ↓0.4%	**0.4672** ↑3.6%	**0.0615** ↓2.7%
Retail Scenario								
Ego-only	0.7907	0.4454	0.4112	0.0411	0.5841	0.3935	0.3636	0.0313
Ego + wL	0.8334 ↑5.4%	0.4901 ↑10.0%	0.4414 ↑7.3%	0.0392 ↓4.6%	0.6591 ↑12.8%	0.4011 ↑1.9%	0.3811 ↑4.8%	0.0310 ↓1.0%
Ego + wR	0.8241 ↑4.2%	0.4851 ↑8.9%	0.4390 ↑6.8%	0.0395 ↓3.9%	0.6670 ↑14.2%	0.4008 ↑1.9%	0.3812 ↑4.8%	0.0310 ↓1.0%
Ego + wL + wR	**0.8316** ↑5.2%	**0.4900** ↑10.0%	**0.4411** ↑7.3%	**0.0391** ↓4.9%	**0.6591** ↑12.8%	**0.4007** ↑1.8%	**0.3810** ↑4.8%	**0.0309** ↓1.3%
Outdoor Scenario								
Ego-only	0.7905	0.5339	0.4769	0.0456	0.9745	0.4879	0.3750	0.1071
Ego + wL	0.8060 ↑2.0%	0.5533 ↑3.6%	0.5002 ↑4.9%	0.0439 ↓3.7%	0.9628 ↓1.2%	0.4978 ↑2.0%	0.3850 ↑2.7%	0.1043 ↓2.6%
Ego + wR	0.8062 ↑2.0%	0.5530 ↑3.6%	0.5001 ↑4.9%	0.0439 ↓3.7%	0.9628 ↓1.2%	0.4986 ↑2.2%	0.3854 ↑2.8%	0.1043 ↓2.6%
Ego + wL + wR	**0.8061** ↑2.0%	**0.5534** ↑3.7%	**0.5002** ↑4.9%	**0.0439** ↓3.7%	**0.9628** ↓1.2%	**0.4981** ↑2.1%	**0.3851** ↑2.7%	**0.1043** ↓2.6%
Overall (All Scenarios)								
Ego-only	0.8393	0.4792	0.4311	0.0456	0.8271	0.4396	0.3743	0.0615
Ego + wL	0.8567 ↑2.1%	0.5030 ↑5.0%	0.4575 ↑6.1%	0.0437 ↓4.2%	0.8355 ↑1.0%	0.4499 ↑2.3%	0.3856 ↑3.0%	0.0601 ↓2.3%
Ego + wR	0.8561 ↑2.0%	0.5024 ↑4.8%	0.4572 ↑6.1%	0.0437 ↓4.2%	0.8355 ↑1.0%	0.4497 ↑2.3%	0.3854 ↑3.0%	0.0602 ↓2.1%
Ego + wL + wR	**0.8566** ↑2.1%	**0.5030** ↑5.0%	**0.4575** ↑6.1%	**0.0436** ↓4.4%	**0.8347** ↑0.9%	**0.4496** ↑2.3%	**0.3852** ↑2.9%	**0.0601** ↓2.3%

Temporal Accuracy changes more modestly and varies across scenarios, suggesting that wrist views mainly help resolve where and how strongly contact occurs rather than simply whether contact occurs. We also find that a single wrist view already captures much of the complementary evidence, likely because the fisheye cameras cover a broad interaction region and can sometimes observe details of the opposite hand. Thus, the main benefit comes from adding at least one contact-aware viewpoint, while the second wrist view provides additional gains mainly under stronger bimanual occlusion.

4.3 Ablation Studies

View dropout is important. As shown in Table 3, training without view dropout leads to a substantial performance drop when wrist views are unavailable at inference time, indicating that the model relies heavily on the full multi-view setting. In contrast, incorporating view dropout significantly improves robustness: the relative performance drop (Δ V) between all-view and ego-only inference improves from -27.20% to -5.78% on the seen-object split.

Importantly, view dropout enables strong generalization to partial-view inputs while maintaining competitive full multi-view performance. This suggests that exposure to varying view combinations during training encourages the model to learn complementary representations across views, rather than relying on a fixed

Table 3 View-dropout ablation on the seen-object split. View dropout (p=0.3) improves robustness to missing wrist views; Ego + wL/wR reports the average performance over the two single-wrist configurations; ΔV denotes the relative V.IoU drop from all-view inference under the same training strategy.

Training	Ego only				Ego + wL/wR				All views			
	C.IoU	V.IoU	MAE	ΔV	C.IoU	V.IoU	MAE	ΔV	C.IoU	V.IoU	MAE	ΔV
No dropout	0.3623	0.3233	0.0826	-27.20%	0.4497	0.4073	0.0527	-8.29%	0.4883	0.4441	0.0453	–
w/ dropout	**0.4792**	**0.4311**	**0.0456**	**-5.78%**	**0.5027**	**0.4573**	**0.0437**	**-0.04%**	**0.5030**	**0.4575**	**0.0436**	**--**

Figure 6 Data scaling ablation study. Performance improves consistently with more training data across all metrics, demonstrating the model's ability to leverage larger datasets effectively.

camera configuration. As a result, the model can flexibly operate under different deployment conditions where some views may be missing or occluded.

Performance scales with data. We study how performance varies with the amount of training data by training the model on 25%, 50%, 75%, and 100% of the dataset. As shown in Figure 6, performance consistently improves as more data is used. In particular, both Contact IoU and Volumetric IoU exhibit steady gains, indicating that the model benefits from increased diversity in interaction patterns and contact configurations. Notably, the improvement does not saturate at higher data regimes, suggesting that the proposed task remains **data-hungry and can further benefit from larger-scale tactile datasets**. This highlights the importance of scaling data collection for learning robust vision-to-tactile mappings.

4.4 Qualitative Results

Figure 7 illustrates how occlusion in egocentric views leads to incomplete tactile predictions. When the contact interface is not directly visible, the model lacks sufficient visual evidence to infer pressure accurately, resulting in missing contact regions. By incorporating wrist-mounted views that directly observe the contact interface, the model gains access to previously occluded information and is able to recover both the location and intensity of contact. This demonstrates that multi-view observations provide critical complementary evidence for resolving occlusion-induced ambiguity in vision-to-tactile prediction.

5 Conclusion

We presented EgoTouch, a large-scale multi-view egocentric dataset with dense tactile supervision for bimanual hand-object interaction. EgoTouch provides synchronized head- and wrist-mounted RGB videos, bimanual 3D hand pose, tracker poses, and continuous tactile pressure maps across 208 tasks and 1,891 episodes. We further introduced TouchAnything, a multi-view vision-to-touch baseline with cross-view fusion and view dropout for flexible inference under different camera configurations. Experiments show that wrist-mounted views provide complementary contact evidence and improve tactile prediction, especially for contact localization and pressure estimation under egocentric occlusion. We hope EgoTouch will support future research on

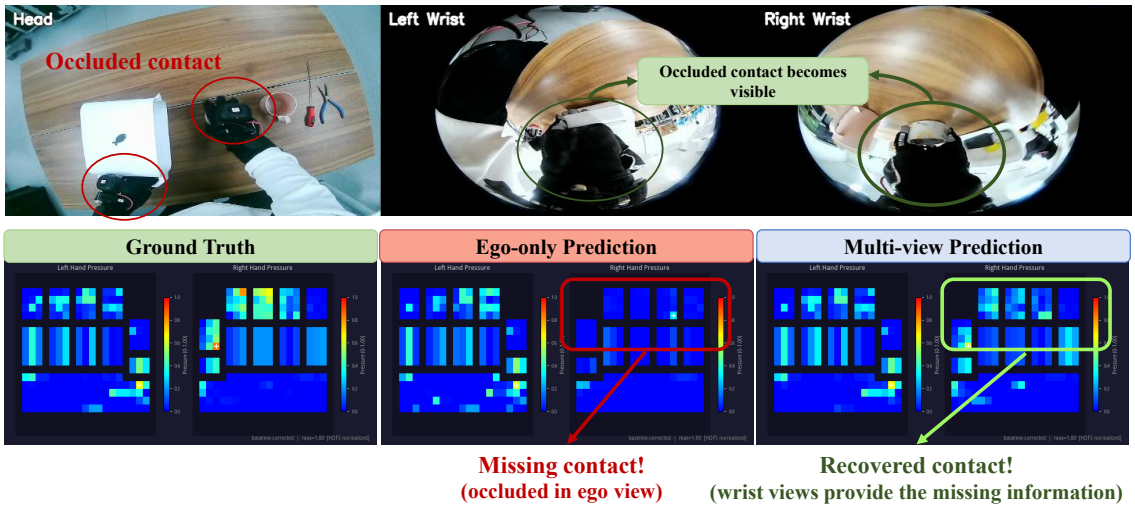


Figure 7 Multi-view wrist cameras recover occluded hand--object contact. Top: the egocentric view suffers from occlusion, while wrist-mounted views reveal the contact interface. Bottom: ego-only prediction misses contact in occluded regions, whereas multi-view prediction recovers accurate pressure distributions consistent with the ground truth.

tactile-grounded embodied perception, manipulation, and learning from egocentric human interaction data.

6 Limitations and Future Work

Our work has several limitations. First, all current training data are collected with tactile gloves, which may introduce glove-specific appearance bias and limit generalization to bare-hand tactile estimation. Future work will explore glove-to-bare-hand retargeting and augmentation to improve robustness in natural human interactions.

Second, our data-scaling analysis shows that model performance has not yet saturated. As shown in Fig. 6, Contact IoU and Volumetric IoU continue to improve with more training data, suggesting that vision-to-touch prediction remains data-hungry. We therefore plan to expand EgoTouch with more diverse objects, environments, contact patterns, and manipulation behaviors.

Finally, the current benchmark focuses primarily on tactile estimation. We hope EgoTouch can support future research on tactile-grounded embodied intelligence, including contact-aware manipulation, grasp stability prediction, affordance learning, tactile-enhanced world models, and robot policy learning from egocentric human demonstrations.

References

[1] Amrit Aryal, Santosh Giri, Sanjeeb Prasad Panday, Suman Sharma, Babu R. Dawadi, and Sushant Chalise. Efficient 3d scene reconstruction from multi-view RGB images using optimized gaussian splatting. IEEE Access, 14: 1269–1286, 2026. doi: 10.1109/ACCESS.2025.3648171. URL https://doi.org/10.1109/ACCESS.2025.3648171.

[2] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard A. Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. HOT3D: hand and object tracking in 3d from egocentric multi-view videos. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025, pages 7061–7071. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00662. URL https://openaccess.thecvf.com/content/CVPR2025/html/Banerjee_HOT3D_Hand_and_Object_Tracking_in_3D_from_Egocentric_Multi-View_CVPR_2025_paper.html.

[3] Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging, 2019. URL https://arxiv.org/abs/1904.06830.

[4] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. Dexycb: A benchmark for capturing hand grasping of objects. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[5] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, et al. Scaling egocentric vision: The epic-kitchens dataset. In European Conference on Computer Vision (ECCV), 2018.

[6] Joseph Del Preore and Daniela Rus. Actionsense: A multimodal dataset and recording framework for human activities using wearable sensors in a kitchen environment. In Advances in Neural Information Processing Systems (NeurIPS), 2022.

[7] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[8] Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, Qianli Ma, Seungjun Nah, Loic Magne, Jiannan Xiang, Yuqi Xie, Ruijie Zheng, Dantong Niu, You Liang Tan, K. R. Zentner, George Kurian, Suneel Indupuru, Pooya Jannaty, Jinwei Gu, Jun Zhang, Jitendra Malik, Pieter Abbeel, Ming-Yu Liu, Yuke Zhu, Joel Jang, and Linxi "Jim" Fan. Dreamdojo: A generalist robot world model from large-scale human videos. CoRR, abs/2602.06949, 2026. doi: 10.48550/ARXIV.2602.06949. URL https://doi.org/10.48550/arXiv.2602.06949.

[9] Patrick Grady et al. Egopressure: A dataset for hand pressure and pose estimation in egocentric views. arXiv preprint, 2024.

[10] Kristen Grauman, Andrew Westbury, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[11] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zachary Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, María Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Dutt Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J. Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina González, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran K. Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbeláez, Gedas Bertasius, David Crandall, Dima Damen, Jakob Julian Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard A. Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shout, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. Int. J. Comput. Vis., 133(12): 8356–8435, 2025. doi: 10.1007/S11263-025-02557-6. URL https://doi.org/10.1007/s11263-025-02557-6.

[12] Kristen Grauman et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[13] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 11090–11100, June 2022.

[14] Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video, 2026. URL https://arxiv.org/abs/2505.11709.

[15] Umar Iqbal, Pavlo Molchanov, Thomas Breuel, Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. arXiv preprint arXiv:1804.09534, 2018.

[16] Mike Lambeta, Po-Wei Chou, Stephen Tian, et al. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. In IEEE Robotics and Automation Letters, 2020.

[17] Jae Yong Lee, Daniel Scharstein, Akash Bapat, Hao Hu, Andrew Fu, Haoru Zhao, Paul Sammut, Xiang Li, Stephen Jeapes, Anik Gupta, Lior David, Saketh Madhuvarasu, Jay Girish Joshi, and Jason Wither. Ego-1k - A large-scale multiview video dataset for egocentric vision. CoRR, abs/2603.13741, 2026. doi: 10.48550/ARXIV.2603.13741. URL https://doi.org/10.48550/arXiv.2603.13741.

[18] Yunzhu Li, Jun-Yan Li, Antonio Torralba, et al. Connecting touch and vision via cross-modal prediction. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[19] Yijiong Lin, Mauro Comi, Alex Church, Dandan Zhang, and Nathan F. Lepora. Attention for robot touch: Tactile saliency prediction for robust sim-to-real tactile control. In IROS, pages 10806–10812, 2023. doi: 10.1109/IROS55552.2023.10341888. URL https://doi.org/10.1109/IROS55552.2023.10341888.

[20] Yunze Liu, Yun Liu, Che Jiang, et al. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[21] Hao Luo, Ye Wang, Wanpeng Zhang, Sipeng Zheng, Ziheng Xi, Chaoyi Xu, Haiweng Xu, Haoqi Yuan, Chi Zhang, Yiqing Wang, Yicheng Feng, and Zongqing Lu. Being-h0.5: Scaling human-centric robot learning for cross-embodiment generalization. CoRR, abs/2601.12993, 2026. doi: 10.48550/ARXIV.2601.12993. URL https://doi.org/10.48550/arXiv.2601.12993.

[22] Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. Dinov2: Learning robust visual features without supervision. Transactions on Machine Learning Research, 2023.

[23] Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, Jacob Chalk, Zhifan Zhu, Rhodri Guerrier, Fahd Abdelazim, Bin Zhu, Davide Moltisanti, Michael Wray, Hazel Doughty, and Dima Damen. HD-EPIC: A highly-detailed egocentric video dataset. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025, pages 23901–23913. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02226. URL https://openaccess.thecvf.com/content/CVPR2025/html/Perrett_HD-EPIC_A_Highly-Detailed_Egocentric_Video_Dataset_CVPR_2025_paper.html.

[24] Jiyoung Seo, Dong In Lee, Pilhyeon Lee, Jiwoo Lee, Youn-Hee Gil, Karthik Ramani, and Sangpil Kim. Egocentric hand activity video dataset and bidirectional motion-priors for hand action recognition. IEEE Access, 14:8128–8148, 2026. doi: 10.1109/ACCESS.2026.3652803. URL https://doi.org/10.1109/ACCESS.2026.3652803.

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

[26] Haisheng Su, Feixiang Song, Cong Ma, Wei Wu, and Junchi Yan. Robosense: Large-scale dataset and benchmark for egocentric robot perception and navigation in crowded and unstructured environments. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025, pages 27446–27455. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02556. URL https://openaccess.thecvf.com/content/CVPR2025/html/Su_RoboSense_Large-scale_Dataset_and_Benchmark_for_Egocentric_Robot_Perception_and_CVPR_2025_paper.html.

[27] Markku Suomalainen, Yiannis Karayiannidis, and Ville Kyrki. A survey of robot manipulation in contact. Robotics Auton. Syst., 156:104224, 2022. doi: 10.1016/J.ROBOT.2022.104224. URL https://doi.org/10.1016/j.robot.2022.104224.

[28] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In European Conference on Computer Vision (ECCV), 2020.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS), 2017.

[30] Akihiko Yamaguchi and Christopher G. Atkeson. Recent progress in tactile sensing and sensors for robotic manipulation: can we turn tactile sensing into vision? Adv. Robotics, 33(14):661–673, 2019. doi: 10.1080/01691864.2019.1632222. URL https://doi.org/10.1080/01691864.2019.1632222.

[31] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/354892587fe39b17c2b727af02abff4a-Abstract-Datasets_and_Benchmarks.html.

[32] Linhan Yang, Bidan Huang, Qingbiao Li, Ya-Yen Tsai, Wang Wei Lee, Chaoyang Song, and Jia Pan. Tacgnn: Learning tactile-based in-hand manipulation with a blind robot using hierarchical graph neural network. IEEE Robotics Autom. Lett., 8(6):3605–3612, 2023. doi: 10.1109/LRA.2023.3264759. URL https://doi.org/10.1109/LRA.2023.3264759.

[33] Lixin Yang, Kailin Li, Xinyu Zhan, et al. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[34] Patrick Grady Yang, Christian Haase-Schütz, Marcel Leonardi, et al. Pressurevision: Estimating hand pressure from a single rgb image. In European Conference on Computer Vision (ECCV), 2022.

[35] Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, Hongxu Yin, Sifei Liu, Song Han, Yao Lu, and Xiaolong Wang. Egovla: Learning vision-language-action models from egocentric human videos. CoRR, abs/2507.12440, 2025. doi: 10.48550/ARXIV.2507.12440. URL https://doi.org/10.48550/arXiv.2507.12440.

[36] Taegyoon Yoon, Yegyu Han, Seojin Ji, Jaewoo Park, Sojeong Kim, Taein Kwon, and Hyung-Sin Kim. Egoxtreme: A dataset for robust object pose estimation in egocentric views under extreme conditions. CoRR, abs/2603.25135, 2026. doi: 10.48550/ARXIV.2603.25135. URL https://doi.org/10.48550/arXiv.2603.25135.

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

[38] Gu Zhang, Qicheng Xu, Haozhe Zhang, Jianhan Ma, Long He, Yiming Bao, Zeyu Ping, Zhecheng Yuan, Chenhao Lu, Chengbo Yuan, Tianhai Liang, Xiaoyu Tian, Maanping Shao, Feihong Zhang, Mingyu Ding, Yang Gao, Hao Zhao, Hang Zhao, and Huazhe Xu. Unidex: A robot foundation suite for universal dexterous hand control from egocentric human videos. CoRR, abs/2603.22264, 2026. doi: 10.48550/ARXIV.2603.22264. URL https://doi.org/10.48550/arXiv.2603.22264.

[39] Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, Trevor Darrell, Furong Huang, Yuke Zhu, Danfei Xu, and Linxi Fan. Egoscale: Scaling dexterous manipulation with diverse egocentric human data, 2026. URL https://arxiv.org/abs/2602.16710.

[40] Shaohong Zhong et al. Touching a nerf: Leveraging neural radiance fields for tactile sensory data generation. arXiv preprint arXiv:2304.12828, 2023.

[41] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan C. Russell, Max J. Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single RGB images. In 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pages 813–822. IEEE, 2019. doi: 10.1109/ICCV.2019.00090. URL https://doi.org/10.1109/ICCV.2019.00090.

Appendix

A Related Work

A.1 Egocentric Hand-Object Interaction Datasets

Recent egocentric datasets [17, 23, 24, 26, 36] have substantially advanced the study of hand-object interaction from a first-person perspective. Ego4D [10] and EPIC-KITCHENS [5] provide large-scale egocentric video for activity understanding, but do not offer the paired 3D hand pose and dense tactile annotations needed for tactile reasoning. EgoDex [14] scales up egocentric manipulation data with 3D hand and finger tracking across 194 tasks, but relies on a single egocentric view and does not provide tactile annotations. EgoPressure [9] pairs egocentric video with real pressure supervision, but focuses on single-hand hand-surface interactions collected in a controlled indoor setup, lacking diverse hand-object manipulation scenarios. OpenTouch [25] introduces in-the-wild full-hand tactile sensing with synchronized video-touch-pose data, but remains limited to single-hand interactions and a single first-person viewpoint.

A key limitation shared by these datasets is the lack of viewpoints that can directly observe hand-object contact regions, leading to severe occlusion of critical contact areas, especially the palmar surfaces where pressure is applied. While some datasets introduce additional views, they do not provide complementary perspectives that explicitly capture the contact interface. More importantly, no existing dataset jointly provides synchronized multi-view video, bimanual hand pose, and dense real tactile sensing. EgoTouch addresses this gap by combining a head-mounted egocentric camera with dual wrist-mounted cameras that directly observe contact regions, together with dense continuous pressure maps, enabling tactile prediction under realistic occlusion and viewpoint variation.

A.2 Vision-to-Touch Prediction

While hardware-based tactile sensors such as GelSight [37] and DIGIT [16] provide high-resolution contact signals, they are difficult to deploy at scale. Vision-to-touch prediction has therefore emerged as an alternative, aiming to infer tactile feedback from visual observations. PressureVision [34] predicts hand pressure maps from a single RGB image using a convolutional network. VisGel [18] learns cross-modal representations between vision and touch using paired GelSight data. Touching a NeRF [40] leverages neural radiance fields to synthesize tactile signals from 3D geometry. EgoPressure [9] further explores pressure estimation in egocentric settings by predicting surface pressure from RGB observations, but focuses on relatively simple hand-surface interactions rather than diverse hand-object manipulation scenarios.

However, existing vision-to-touch approaches remain limited for realistic egocentric manipulation. In such settings, critical hand-object contact regions are frequently occluded, especially on the palmar surfaces. As a result, visual inputs often lack direct observations of the contact interface, regardless of model capacity. This limitation cannot be resolved by better models alone; it requires complementary viewpoints that explicitly capture the contact region. Building on EgoTouch, we propose TouchAnything, a multi-view tactile prediction model that leverages wrist-mounted cameras to directly observe contact regions, thereby enabling robust tactile estimation under occlusion and viewpoint variation.

A.3 Multi-View Learning for Hand Interaction

Multi-view learning [1, 2] mitigates occlusion by leveraging complementary observations from different viewpoints. In hand understanding, it enables accurate 3D hand pose and mesh reconstruction by recovering geometry that is not visible from a single view [13, 15, 41]. In egocentric perception, Ego-Exo4D [12] combines egocentric and exocentric views to improve activity understanding, highlighting the benefits of cross-view complementarity.

However, existing multi-view approaches primarily target geometric reconstruction or high-level action understanding, rather than modeling physical interaction signals such as tactile feedback. Moreover, while some datasets and methods incorporate multiple views, these viewpoints are typically external or global and do not directly observe the hand-object contact interface. As a result, critical contact regions—especially

on the palmar surfaces—remain occluded or only indirectly inferred. In this work, we introduce wrist-mounted cameras that provide contact-aware viewpoints, directly capturing hand-object interactions from complementary perspectives. This design enhances tactile estimation by supplying visual evidence of contact regions that are otherwise difficult to infer from standard viewpoints.

B Additional Dataset Details

B.1 Data Collection Setup

We collect a multimodal dataset of bimanual interactions using a wearable acquisition system that records synchronized RGB videos, tracker poses, bimanual hand kinematics, and dense tactile measurements at a target rate of 30Hz.

Hardware Configuration. The visual subsystem contains three RGB cameras: a chest/head-mounted egocentric camera and two wrist-mounted cameras observing the left and right hands. During acquisition, the GUI stores these views directly as three 30 FPS videos, `chest.mp4`, `left.mp4`, and `right.mp4`. The wrist cameras provide close-range observations of hand-object contacts that are often occluded in the egocentric view.

For global spatial localization, we use three HTC Vive Trackers assigned to the roles `chest`, `left_wrist`, and `right_wrist`. The trackers are read through OpenXR using the `XR_HTCX_vive_tracker_interaction` extension. For each collection tick, the latest valid tracker state is stored as a 3D translation and quaternion rotation in `vive_poses.json`.

Bimanual hand pose is captured using Rokoko motion-capture gloves. The Rokoko stream is received through UDP, parsed into 21 3D joints per hand, and cached for low-latency access by the GUI. When calibration matrices are available, the system transforms the left and right Rokoko hand joints into the Vive coordinate frame using the corresponding wrist tracker pose and records this with an `aligned_to_vive` flag. The resulting per-frame hand joints are saved in `rokoko_hands.json`.

Dense tactile feedback is captured by custom pressure-sensing gloves. Each hand sends a 256-channel 8-bit pressure vector through a serial connection at 921600 baud, together with an IMU packet decoded into a quaternion. The latest pressure vectors and IMU quaternions for both hands are saved in `jq_pressure.json`.

Data Acquisition Pipeline. The current acquisition software is implemented in the PyQt GUI. Each recording episode is stored under a timestamped directory with the following structure:

```
<root>/<category>/<date>/<task>/<episode_timestamp>/
    chest.mp4
    left.mp4
    right.mp4
    jq_pressure.json
    rokoko_hands.json
    vive_poses.json
    camera_matrix.txt
```

The three `.json` files are JSON Lines files: each line stores one frame-level record with a common timestamp `ts` and integer `frame_index`. The saved records contain:

- `jq_pressure.json`: `sensor_left` and `sensor_right`, each a 256-value pressure vector, plus `quat_left` and `quat_right`;
- `rokoko_hands.json`: `left_pos` and `right_pos`, each a 21×3 joint array, plus `aligned_to_vive`;
- `vive_poses.json`: a dictionary of tracker poses for `chest`, `left_wrist`, and `right_wrist`, each containing `trans` and `rot`.

If camera intrinsics are available, they are written once as `camera_matrix.txt`. This episode-level storage format avoids creating thousands of small image and text files and keeps the RGB streams temporally aligned with compact frame-wise sensor metadata.

Synchronization Strategy. The system uses software synchronization. Sensor readers for cameras, Rokoko, Vive Trackers, and pressure gloves run asynchronously and keep their latest valid measurements in memory. A 30Hz timer submits frame indices to a background saving worker. For each frame index, the GUI builds a snapshot containing the latest available RGB frames, hand joints, tracker poses, pressure vectors, and glove quaternions. The worker then appends one row to each JSON Lines file and writes the corresponding RGB frames to the video streams. When recording stops, the GUI waits for the queued frames to be written before closing the video and JSON files, so the number of saved JSON rows matches the number of submitted collection frames whenever possible.

B.2 Tactile Grid Mapping and Preprocessing

The raw tactile glove stream contains 256 sensor values per hand. Although these values can be interpreted as a compact 16×16 array, the physical sensor layout on the glove is not a regular image grid: different sensors correspond to different fingers, palm regions, and bending/contact locations. To preserve the spatial structure of the hand, we remap the raw 256-dimensional tactile vector into a 21×21 hand-shaped pressure grid before training. The mapping is defined by hand-specific JSON files, where each key specifies a target grid coordinate (r, c) and each value specifies the corresponding raw sensor index. This produces a sparse hand-shaped tactile map whose valid locations follow the physical arrangement of the glove sensors.

For each frame, we initialize a 21×21 grid with invalid locations marked as NaN and fill only the mapped sensor locations. The left hand is placed directly according to the mapping, while the right hand is horizontally mirrored so that left and right tactile maps share a consistent canonical hand coordinate system. This canonical representation makes the model output directly interpretable as a hand-shaped pressure distribution rather than an arbitrary sensor vector, providing stronger spatial priors for learning contact location and pressure magnitude.

We further apply several preprocessing steps to improve data quality. First, we optionally subtract the first-frame baseline pressure when the first frame is judged to be contact-free, either according to manual contact annotations or a low-pressure threshold fallback. This removes static sensor bias while avoiding over-correction when the sequence starts with an active contact. Second, known broken columns in the right-hand tactile grid are repaired by interpolation from neighboring valid columns. Third, tactile sensors and bending-related sensors are normalized separately, preventing high bending-sensor values from compressing the dynamic range of true contact-pressure sensors. The processed pressure grids, baseline-correction flags, grid size, and normalization metadata are saved into `pressure_grids.npz` for downstream conversion and training.

B.3 HDF5 Dataset Conversion

After cleaning and tactile-grid preprocessing, we convert each trajectory into a compact HDF5 file for efficient training and inference. The conversion script expects each trajectory to contain synchronized videos from the three cameras, processed pressure grids, and hand-pose annotations. During conversion, RGB frames are decoded from `ego.mp4`, `left.mp4`, and `right.mp4`; when available, GPU-accelerated FFmpeg decoding is used, with OpenCV as a fallback. The final frame count is set to the minimum valid length across the three camera streams to keep all views temporally aligned.

Each HDF5 file stores a stable hierarchy of modalities. Metadata include the trajectory id, task name, number of frames, FPS, image resolution, and quality flags such as duplicated wrist-camera frames. The `images` group stores the three RGB streams. The `poses` group stores 7D Vive Tracker poses for the head/chest and both wrists. The `hands` group stores Rokoko hand joints when available, as well as WiLoR-based left/right 21-joint hand poses and validity masks. The `pressure` group stores the processed left and right 21×21 pressure grids together with preprocessing metadata. Optional glove masks are stored under a separate `masks` group when available.

For large-scale conversion, we use the batch conversion script with `64` trajectory workers, gzip compression level `4`, and `-skip_existing` to avoid recomputing valid HDF5 files. The same script also supports regenerating trajectories listed in a bad-file list, filtered by failure reason, which is useful for repairing read errors or quality-control failures without reprocessing the entire dataset. This HDF5 format substantially reduces data-loading overhead and ensures that all experiments use the same cleaned, temporally aligned, and spatially remapped tactile representation.

C Additional Implementation Details

We propose a multi-view tactile prediction model that takes as input any subset of the three camera views and bimanual hand pose, and predicts dense bilateral pressure maps. The architecture is designed to gracefully handle missing views at inference time through a view dropout training strategy.

C.1 Problem Formulation

Given a video clip of T frames from any subset $\mathcal{V} \subseteq \{V^{ego}, V^{wL}, V^{wR}\}$ of available views and the corresponding bimanual hand pose sequence $\mathbf{P} \in \mathbb{R}^{T\times 42\times 3}$, our goal is to predict the bilateral pressure maps $\hat{\mathbf{M}} \in \mathbb{R}^{T\times 2\times 21\times 21}$ for both hands at each timestep. The model must produce reasonable predictions regardless of which views are available.

C.2 Multi-View Vision Encoder

Shared backbone with view embeddings. All views are processed by a shared DINOv2-ViT-B/14 [22] backbone, which extracts $N = 256$ patch tokens of dimension $D = 768$ per frame. To enable the model to distinguish which camera a patch originates from, we add a learnable view embedding $\mathbf{e}_v \in \mathbb{R}^D$ for each view $v \in \{ego, wL, wR\}$:

$$\mathbf{F}_v = \text{DINOv2}(V_v) + \mathbf{e}_v, \quad \mathbf{F}_v \in \mathbb{R}^{T\times N\times D} \tag{3}$$

Sharing the backbone across views reduces the parameter count from $3 \times 86\text{M}$ (separate encoders) to $86\text{M} + 3 \times 768$ (shared encoder + view embeddings), improving both efficiency and generalization.

Cross-view attention. Rather than performing expensive attention over all $N \times |\mathcal{V}|$ patch tokens, we extract a summary token for each view via global average pooling and apply a lightweight cross-view transformer [29] over the $|\mathcal{V}|$ summary tokens:

$$\mathbf{s}_v = \text{MeanPool}(\mathbf{F}_v), \quad [\hat{\mathbf{s}}_1, \ldots, \hat{\mathbf{s}}_{|\mathcal{V}|}] = \text{CrossViewTransformer}([\mathbf{s}_1, \ldots, \mathbf{s}_{|\mathcal{V}|}]) \tag{4}$$

This allows each view's summary to attend to summaries from other views, enabling complementary information exchange (e.g., the wrist view can inform the egocentric view about occluded contact regions).

Gated view fusion. The fused summary tokens are passed through a gating network that learns view-dependent importance weights:

$$w_v = \text{softmax}\big(\text{MLP}(\hat{\mathbf{s}}_v)\big), \quad \mathbf{F}^{fused} = \sum_{v\in\mathcal{V}} w_v \cdot \mathbf{F}_v \tag{5}$$

The output $\mathbf{F}^{fused} \in \mathbb{R}^{T\times N\times D}$ has the same shape as single-view features, ensuring compatibility with downstream modules.

C.3 Temporal Modeling and Pose-Vision Fusion

Temporal transformer. A windowed temporal transformer [29] is applied across the time dimension to capture manipulation dynamics:

$$\mathbf{H} = \text{TemporalTransformer}(\mathbf{F}^{fused}), \quad \mathbf{H} \in \mathbb{R}^{T\times N\times D} \tag{6}$$

Pose encoder. The bimanual hand pose $\mathbf{P} \in \mathbb{R}^{T\times 42\times 3}$ is encoded by a transformer-based pose encoder that produces per-joint features $\mathbf{G} \in \mathbb{R}^{T\times 42\times D}$.

Pose-vision cross-attention fusion. Each joint token queries the visual patch tokens via cross-attention, enabling spatially grounded fusion:

$$\mathbf{Z} = \text{CrossAttn}(Q{=}\mathbf{G},\ K{=}\mathbf{H},\ V{=}\mathbf{H}), \quad \mathbf{Z} \in \mathbb{R}^{T\times 42\times D} \tag{7}$$

This allows each joint to attend to the visual patches most relevant to its spatial location and contact state.

C.4 Joint-Level Tactile Decoder

The fused joint features $\mathbf{Z}$ are decoded into bilateral pressure maps. The 42-joint features are split into left-hand (joints 1–21) and right-hand (joints 22–42) groups, each decoded independently into a 21×21 pressure map via an MLP followed by a reshape operation:

$$\hat{\mathbf{M}}_t^{left} = \sigma\big(\text{MLP}(\mathbf{Z}_t^{left})\big) \in [0,1]^{21\times 21}, \quad \hat{\mathbf{M}}_t^{right} = \sigma\big(\text{MLP}(\mathbf{Z}_t^{right})\big) \in [0,1]^{21\times 21} \tag{8}$$

where σ is the sigmoid function ensuring outputs are in $[0,1]$.

C.5 View Dropout Training Strategy

A critical design requirement is that the model must work with any available subset of views at inference time. To achieve this, we employ view dropout during training: the egocentric view is always retained, while each wrist view is independently dropped with probability $p = 0.3$, matching the configuration used in our training script. This exposes the model to four possible input configurations:

- Ego only (both wrist views dropped)
- Ego + left wrist
- Ego + right wrist
- All three views (no views dropped)

At inference time, the same model accepts whichever views are available without architectural changes or fine-tuning. This enables systematic evaluation under different camera configurations, including ego-only deployment and full multi-view inference.

C.6 Training Objective and Optimization

The training objective combines pixel-wise regression, sparsity-aware pressure fitting, and spatial regularization. Let $\hat{\mathbf{M}}$ and $\mathbf{M}$ denote the predicted and ground-truth tactile maps. We optimize

$$\mathcal{L} = \lambda_{mse}\mathcal{L}_{mse} + \lambda_{l1}\mathcal{L}_{l1} + \lambda_{tv}\mathcal{L}_{TV}, \tag{9}$$

where $\mathcal{L}_{mse}$ and $\mathcal{L}_{l1}$ measure pressure reconstruction error and $\mathcal{L}_{TV}$ encourages spatial smoothness in the predicted pressure maps. To reduce the tendency of the model to predict all-zero pressure maps under sparse contact supervision, pixels with pressure greater than 0.1 are treated as contact regions and assigned a larger loss weight. In our experiments, we use $\lambda_{mse} = 1.0$, $\lambda_{l1} = 0.5$, $\lambda_{tv} = 0.01$, and a contact-region weight of 3.0.

We train the model with AdamW using a learning rate of 5×10^{-5}, weight decay of 0.05, and betas $(0.9, 0.999)$. The learning rate follows a cosine schedule with 10 warmup epochs and a minimum learning rate of 10^{-6}. Training runs for 25 epochs. We use distributed data parallel training launched by `torchrun`; by default, the launcher uses six GPUs with per-GPU batch size 16 and gradient accumulation over 3 steps, resulting in an effective batch size of 288. The DINOv2 ViT-B/14 visual encoder is initialized from pretrained weights and kept frozen during training. We use clips of 8 frames sampled every 2 frames, RGB inputs resized to 224×224, 42 bimanual hand joints from WiLoR, glove color augmentation with probability 0.2, and 21×21 tactile maps aligned with the native pressure-sensor grid.

C.7 Inference and Evaluation Protocol

For evaluation, we load the best model checkpoint and run batched inference with the same configuration file used during training. The inference script supports both full evaluation and a lightweight mode for quick inspection. In the full setting, it evaluates up to 1000 trajectories per split; in lightweight mode, it samples one trajectory per task category from the split file. Unless otherwise specified, inference uses batch size 64, two worker processes, 30 FPS visualization, and skips saving HDF5 outputs to reduce storage usage while retaining videos and metrics.

The inference pipeline supports configurable view subsets through a `views` argument. We use this to evaluate ego-only input, single-wrist variants, and full multi-view input with the same trained model. Outputs are organized by configuration name, checkpoint name, view setting, and dataset split. For each evaluated split, the script saves visualizations and computes tactile prediction metrics including Temporal Accuracy, Contact IoU, Volumetric IoU, and MAE. This shared inference protocol ensures that all view configurations are compared under the same checkpoint, dataset split, preprocessing, and metric implementation.

D Additional Qualitative Results

We provide additional qualitative examples of tactile prediction results across diverse manipulation tasks. Each visualization shows a 2×3 grid of frames sampled uniformly from the video sequence, with the egocentric RGB input (top row), ground truth pressure maps for both hands (middle row), and predicted pressure maps (bottom row). The model accurately captures contact locations, pressure intensity, and temporal dynamics across a wide range of interactions.

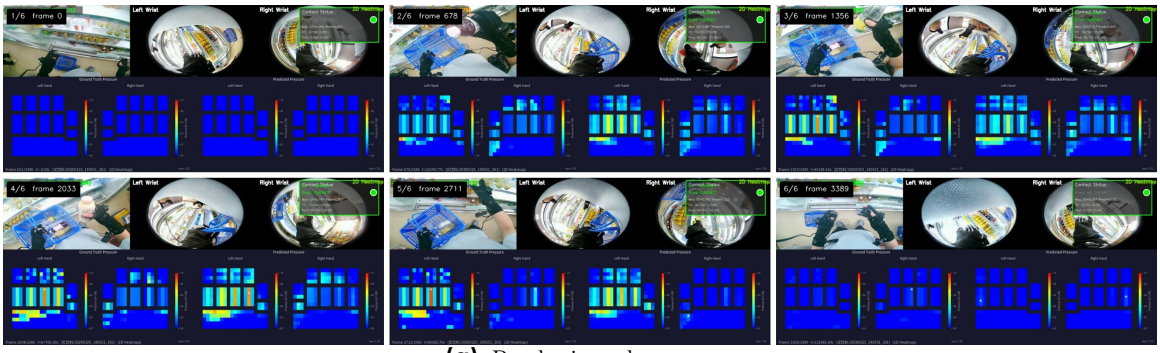
These examples demonstrate the model's ability to generalize across diverse object categories, manipulation strategies, and contact configurations. The predicted pressure maps maintain spatial consistency with the hand pose and visual observations, while capturing the temporal evolution of contact during continuous manipulation.

D.1 Failure Cases and Limitations

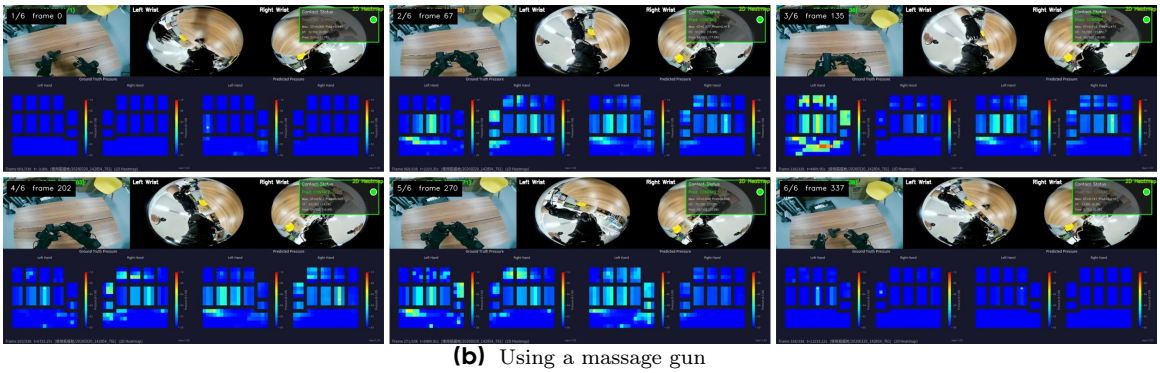
While the model achieves strong performance on most manipulation tasks, we observe failure modes in challenging visual conditions. Figure 12 shows a representative failure case where the model incorrectly predicts contact in the first frame when no contact has occurred yet.

This failure is caused by the low visual contrast between the black glove and the black shorts. From the egocentric view, the hand appears to be in close proximity to or overlapping with the dark fabric, creating an ambiguous visual signal that the model interprets as contact. This demonstrates that the model has learned to associate visual proximity and occlusion patterns with tactile contact, but can be misled when color similarity makes it difficult to distinguish hand-object boundaries.

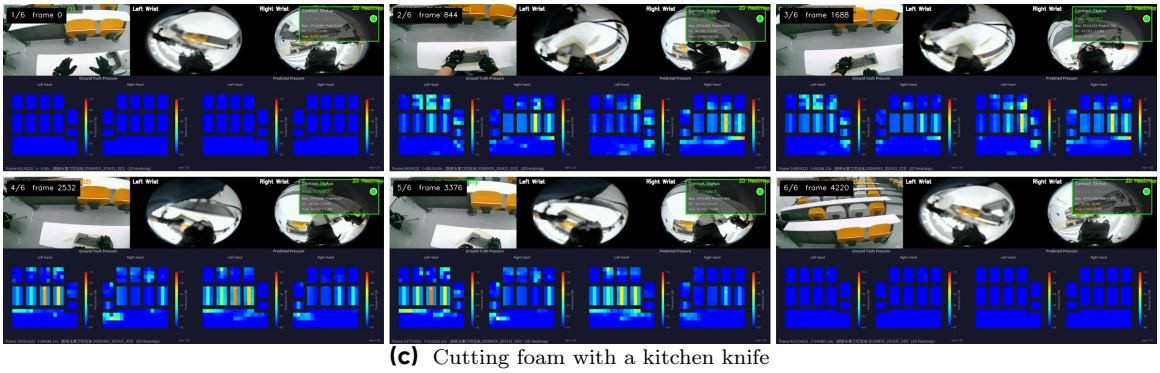
Such failures suggest several directions for improvement: (1) incorporating explicit depth or hand-object segmentation to disambiguate proximity from contact, (2) augmenting training data with more challenging color combinations, and (3) leveraging temporal consistency to suppress isolated false-positive predictions in the first frame when no prior contact history exists.



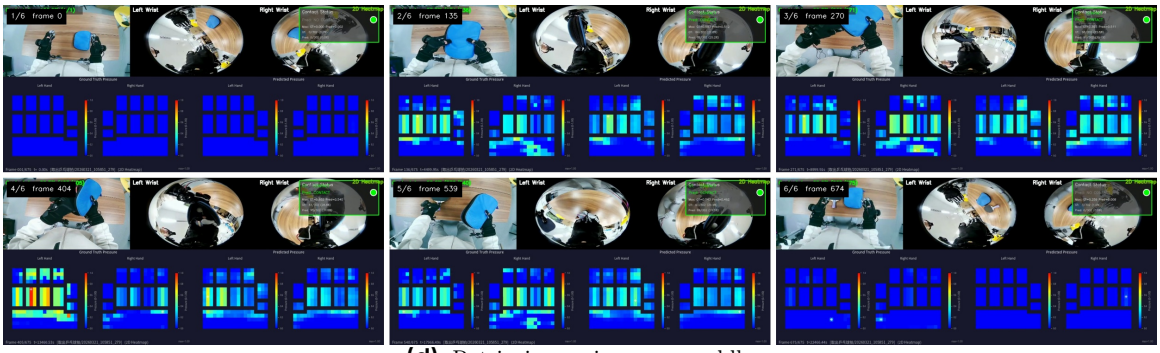
(a) Purchasing a beverage



(b) Using a massage gun

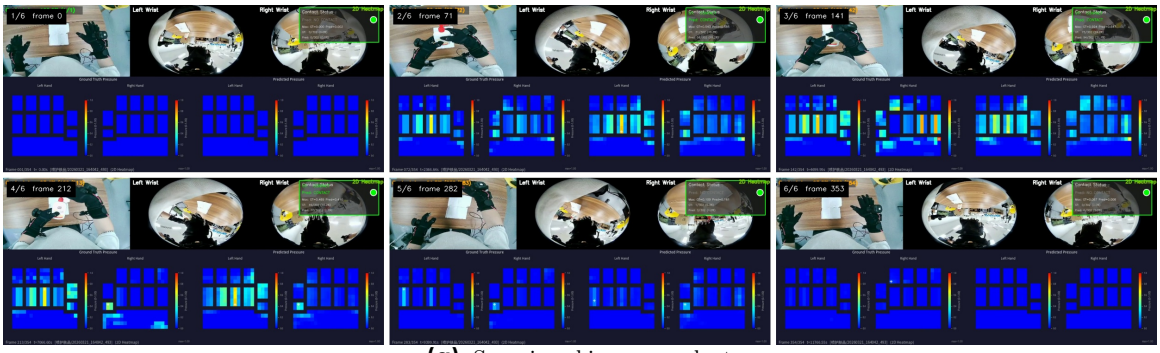


(c) Cutting foam with a kitchen knife

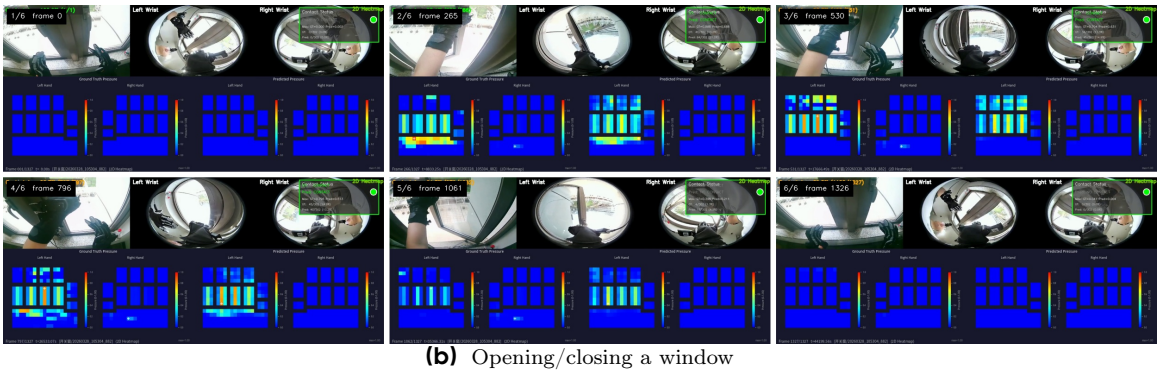


(d) Retrieving a ping-pong paddle

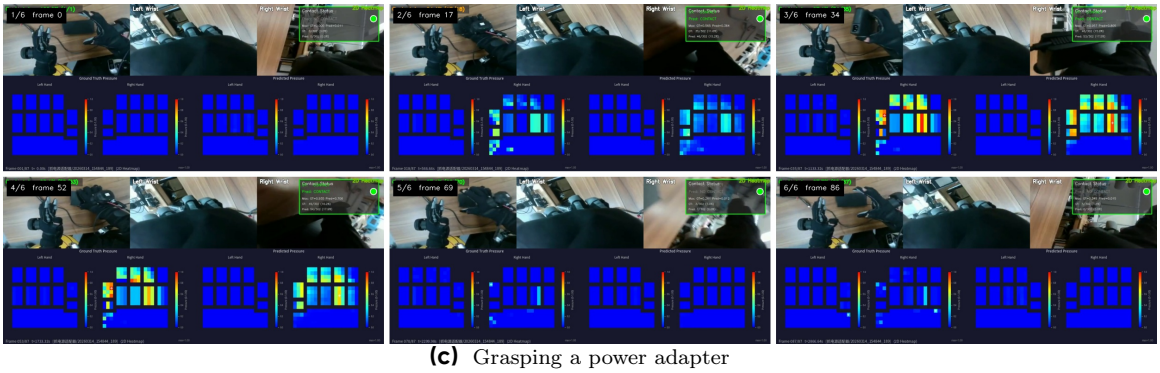
Figure 8 Tactile prediction results (1–4). Each subfigure shows a 2×3 grid of frames with predicted tactile pressure maps overlaid on the egocentric RGB input.



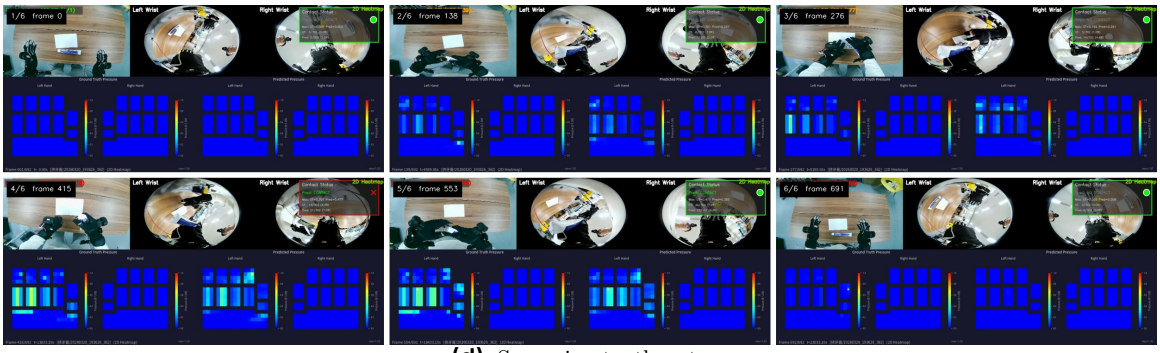
(a) Spraying skincare product



(b) Opening/closing a window

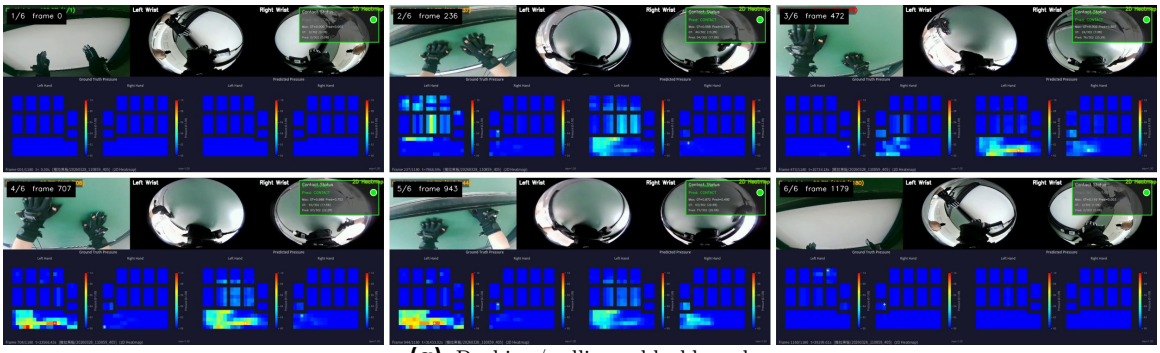


(c) Grasping a power adapter

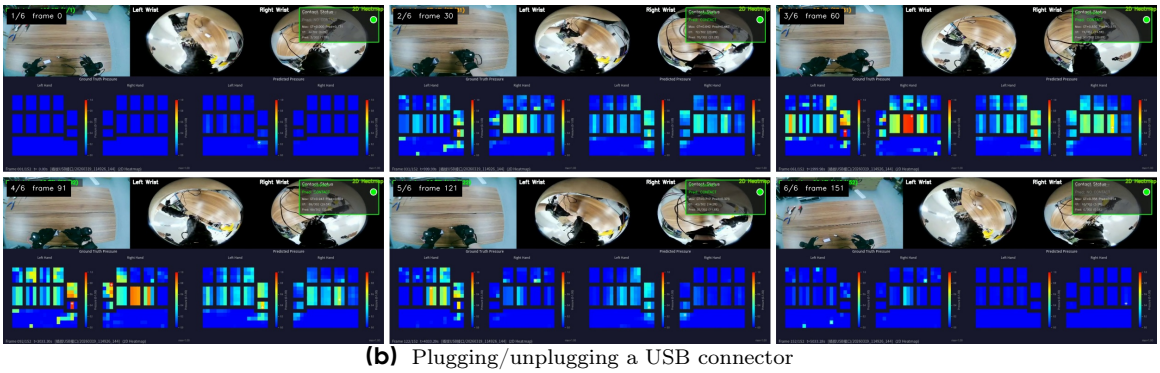


(d) Squeezing toothpaste

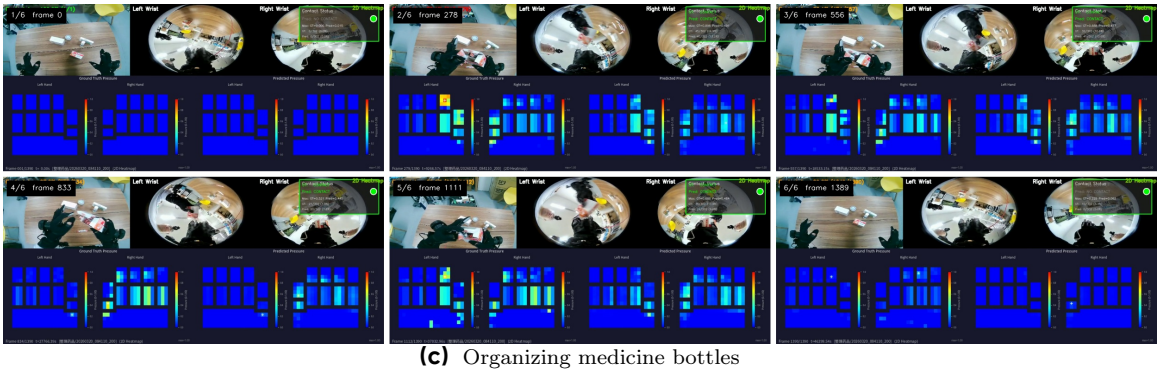
Figure 9 Tactile prediction results (5–8). The model captures fine-grained contact patterns and bimanual coordination across diverse manipulation tasks.



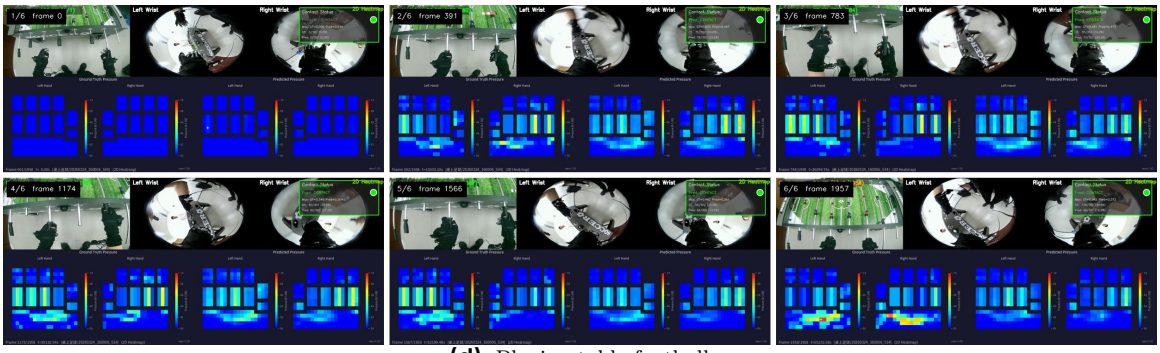
(a) Pushing/pulling a blackboard



(b) Plugging/unplugging a USB connector

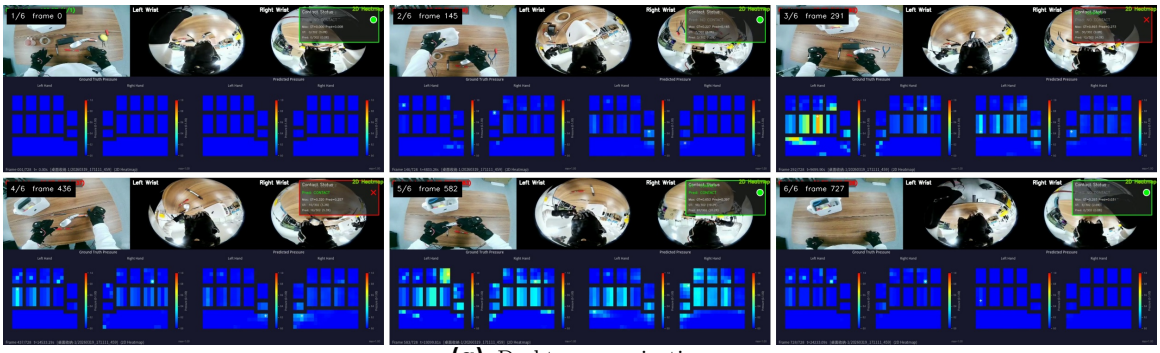


(c) Organizing medicine bottles

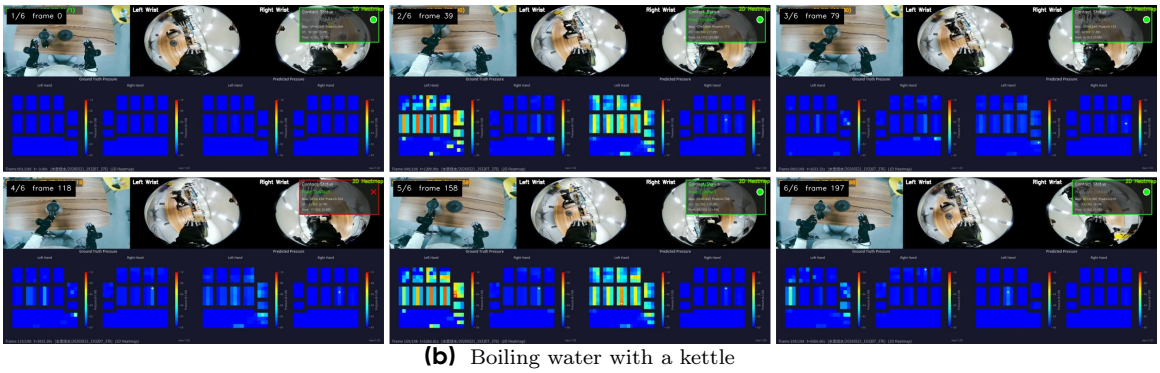


(d) Playing table football

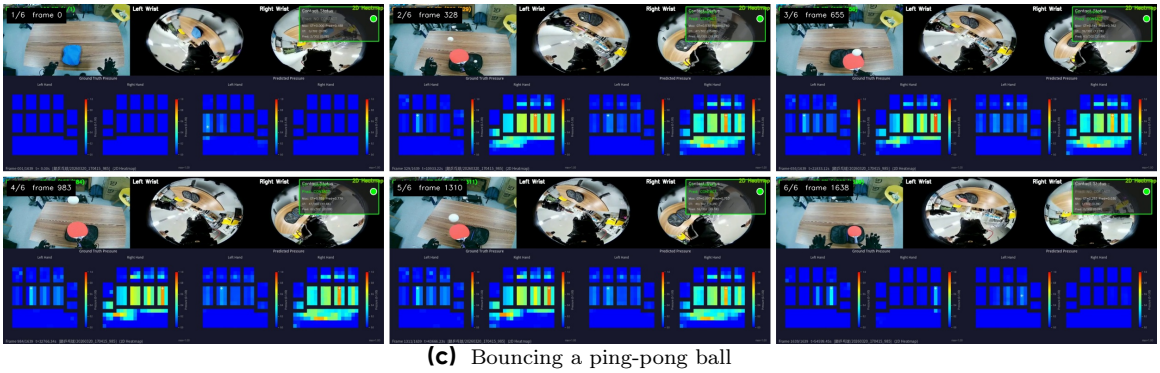
Figure 10 Tactile prediction results (9–12). The model accurately predicts contact locations during precision manipulation and dynamic gameplay.



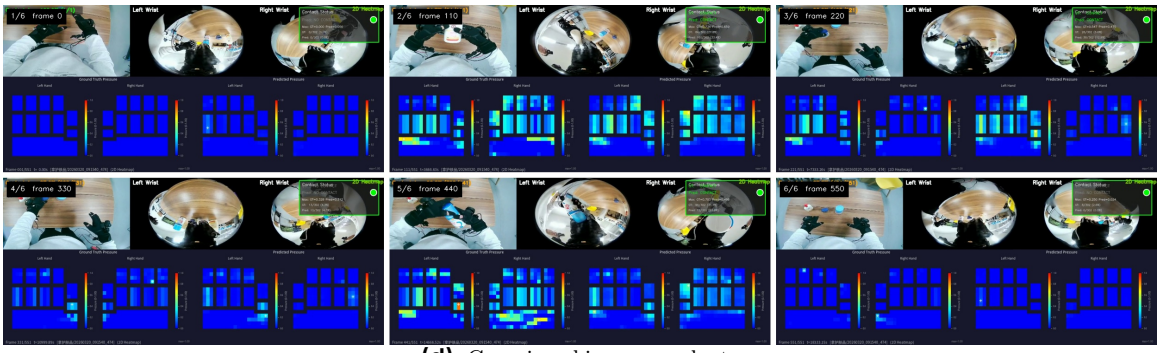
(a) Desktop organization



(b) Boiling water with a kettle



(c) Bouncing a ping-pong ball



(d) Grasping skincare products

Figure 11 Tactile prediction results (13–16). The predicted pressure maps maintain spatial consistency with hand pose and visual observations while capturing temporal evolution of contact.

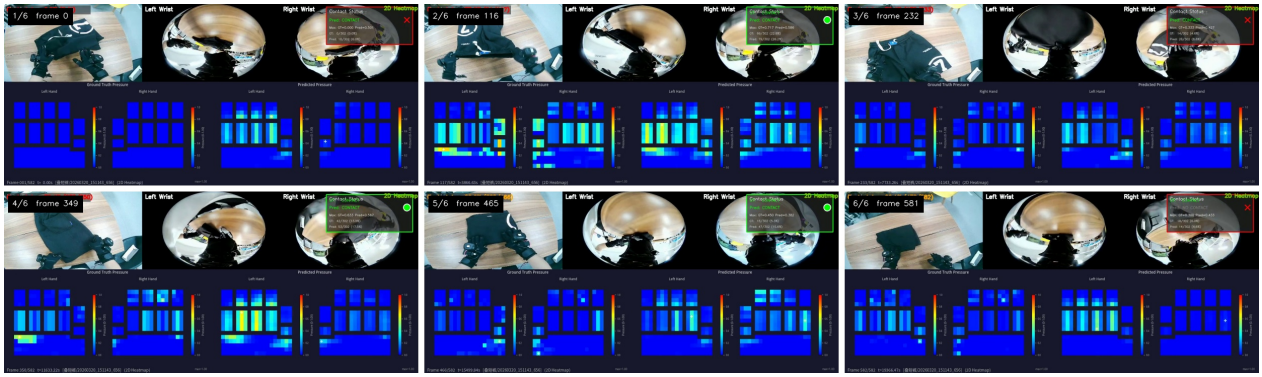


Figure 12 Failure case: folding black shorts. In the first frame, the model incorrectly predicts contact on the left hand even though no contact has occurred. The black glove and black shorts create a low-contrast visual appearance that resembles contact, causing the model to hallucinate pressure. This highlights a key limitation: the model relies heavily on visual cues and can be confused by color similarity between the hand and object.