TOUCHANYTHING: A DATASET AND FRAMEWORK FOR BIMANUAL TACTILE ESTIMATION FROM EGOCENTRIC VIDEO

Anonymous authors
Paper under double-blind review

ABSTRACT

Egocentric video captures diverse human–object interactions at scale, but usually lacks tactile measurements of contact and pressure. Predicting these signals from video could extend tactile supervision to recordings acquired without tactile sensors, although occlusion and ambiguous visual evidence make this challenging. 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 software-aligned multi-view RGB (head-mounted egocentric and dual wrist-mounted cameras), bimanual 3D hand pose, and continuous tactile signals from wearable sensors. Building on EgoTouch, we introduce **TouchAnything**, a baseline multi-view vision-to-touch prediction framework conditioned on hand poses estimated from the egocentric video, with optional wrist-mounted RGB views at inference time. Multi-view training improves ego-only inference over an independently trained ego-only model, increasing seen-object Volumetric IoU by 7.2%. Within the multi-view-trained model, adding wrist RGB at inference further improves Contact IoU by 5.0% and Volumetric IoU by 6.1%. We will publicly release the dataset, code, and benchmark.

1 INTRODUCTION

Egocentric video datasets have become a key source of scalable supervision for embodied intelligence (Kareer et al., 2025), as they are relatively easy to collect, capture natural human-environment interactions, and cover diverse real-world manipulation behaviors (Grauman et al., 2022; Zheng et al., 2026; Hoque et al., 2026; Zhang et al., 2026; Luo et al., 2026; Grauman et al., 2025; Gao et al., 2026; Yang et al., 2025). However, most egocentric recordings lack tactile measurements. Appearance and motion alone do not directly reveal contact force and pressure distributions, leaving physical interaction dynamics without direct supervision (Yang et al., 2023; Lin et al., 2023; Yamaguchi & Atkeson, 2019; Suomalainen et al., 2022). Collecting tactile data requires wearable hardware that is costly and intrusive. This motivates learning to infer tactile signals from visual observations.

Vision-to-touch prediction has shown promise (Grady et al., 2022; Li et al., 2019; Yang et al., 2022a), with recent datasets extending tactile supervision to egocentric settings. EgoPressure (Zhao et al., 2025) studies hand-surface pressure, while OpenTouch (Song et al., 2025) provides in-the-wild full-hand tactile signals for single-hand interactions. EgoTouch complements these efforts by combining bilateral tactile supervision with head- and wrist-mounted RGB views, supporting a benchmark for bimanual tactile prediction under varying view availability. Further related work appears in Appendix A.1. 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** (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 tactile maps and bimanual hand pose on a common frame timeline.

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: EgoTouch combines egocentric and wrist-mounted views with 3D hand pose and dense tactile maps on a common frame timeline, providing complementary visual evidence for learning contact-aware interactions.

It contains 208 manipulation tasks, 1,891 episodes, and approximately 20 hours of interaction data across diverse indoor and outdoor environments.

Building on EgoTouch, we establish **TouchAnything**, a baseline that combines RGB observations with hand poses estimated from the egocentric video using WiLoR. During training, view dropout randomly removes wrist views while retaining the egocentric view. The same model supports four inference configurations: ego only, ego with either wrist view, and all three views. Pose inputs are held fixed across these configurations. In our benchmark, adding wrist-mounted views improves aggregate 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, software-aligned RGB videos from one head-mounted camera and two wrist-mounted cameras, bimanual 3D hand pose, and dense continuous tactile 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 approximately 20 hours of multi-view egocentric video collected by 10 participants at 30 Hz, comprising 1,891 episodes across 208 manipulation tasks. All 10 participants consented to data collection and public release of their recordings. This amounts to approximately 2.1 million frame timepoints covering over 1,000 objects in indoor and outdoor environments. Each timepoint associates three RGB views, bimanual 3D hand pose, and tactile maps for both hands through the software alignment protocol below. We compare EgoTouch with existing hand-object interaction and tactile datasets in Table 1.

Table 1: Comparison of EgoTouch with existing hand interaction and tactile datasets. EgoTouch combines head- and wrist-mounted video, bimanual hand pose, and measured tactile signals across diverse scenes.

Dataset	In-the-wild	Hand Pose	Contact	Wrist Views	Hands	Objects	Frames
GRAB (Taheri et al., 2020)	✗	MoCap	Analytical	✗	Biman.	51	1.6M
ContactDB (Brahmbhatt et al., 2019)	✗	✗	Thermal	✗	Biman.	50	375k
ARCTIC (Fan et al., 2023)	✗	MoCap	Analytical	✗	Biman.	11	2.1M
OakInk (Yang et al., 2022b)	✗	MoCap	Analytical	✗	Single	100	230k
DexYCB (Chao et al., 2021)	✗	Est.	Analytical	✗	Single	20	582k
ActionSense (DelPreto et al., 2022)	✗	Glove	Pressure	✗	Biman.	21	521k
HOI4D (Liu et al., 2022)	✗	Est.	✗	✗	Single	800	2.4M
EgoPressure (Zhao et al., 2025)	✗	Est.	Pressure	✗	Single	31	4.3M
EgoDex (Hoque et al., 2026)	✗	Est.	Analytical	✗	Biman.	500	90M
OpenTouch (Song et al., 2025)	✓	Glove	Pressure	✗	Single	800	∼ 500k
EgoTouch (Ours)	✓	**Glove + Est.**	**Pressure**	✓	**Biman.**	**∼1000**	**2.1M**

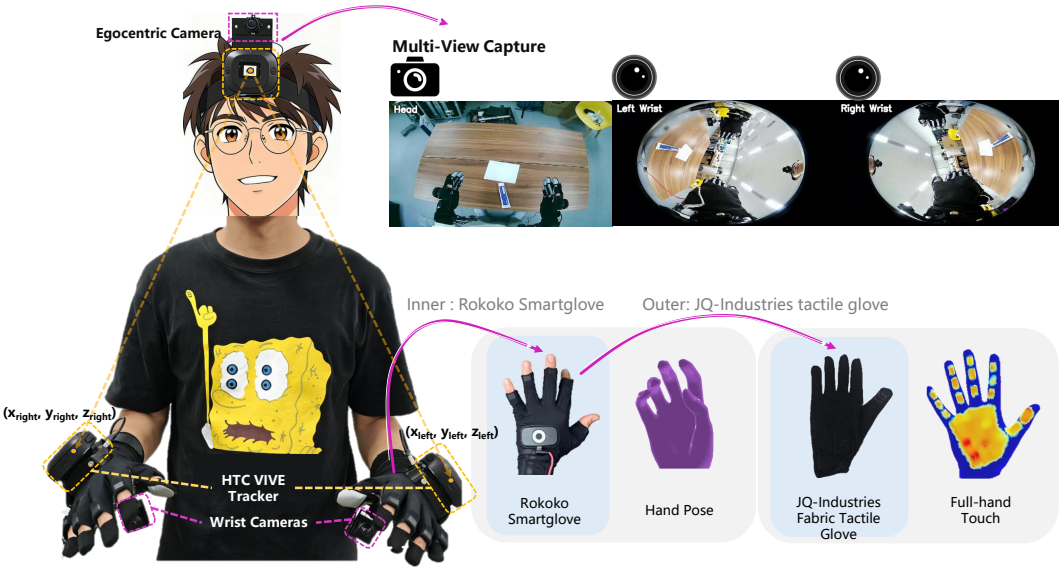


Figure 2: Data collection setup and example multi-modal data. The participant wears a head-mounted egocentric camera, two wrist-mounted cameras, and tactile gloves. Software associates the modalities on a common 30 Hz frame timeline.

2.1 Data Collection Setup

EgoTouch is collected with a wearable system that records complementary visual, kinematic, and tactile signals during bimanual manipulation (Figure 2). It includes a head-mounted RGB camera, two wrist-mounted RGB cameras, Rokoko motion-capture gloves, full-hand tactile gloves providing 256-channel readings per hand, and HTC Vive Trackers for 6-DoF head and wrist localization. The tactile readings cover the fingers and palm and are stored as 16×16 arrays before canonical remapping. We associate the streams on a common 30 Hz frame timeline using the latest available sensor measurements. This software alignment does not guarantee simultaneous acquisition; Appendix A.2.1 describes the asynchronous capture procedure and timing limitations. Figure 3 shows representative observations.

2.2 Data Modalities

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

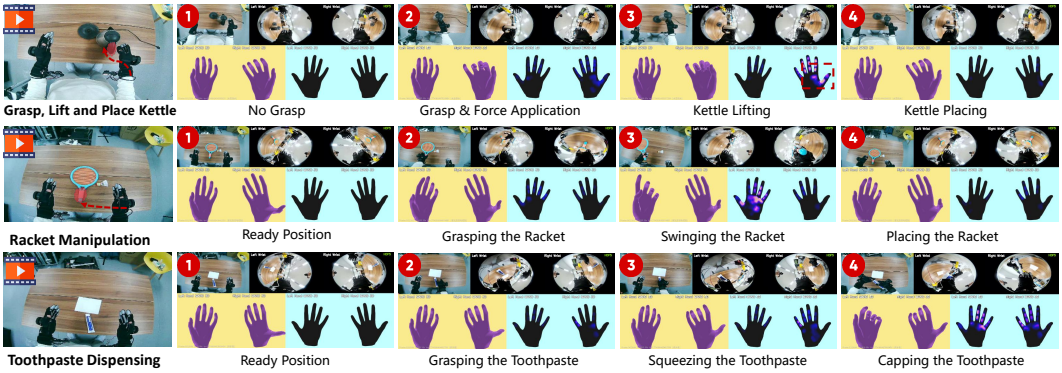


Figure 3: Example EgoTouch sequences pair RGB observations with tactile measurements and hand pose, showing contact patterns and hand motion during object manipulation.

- **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 ($V^{wL}, V^{wR} \in \mathbb{R}^{640\times480\times3}$) 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. The dataset stores captured kinematics and vision-estimated annotations; the benchmark consistently uses WiLoR estimates from the egocentric video for both training and testing.
- **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 A.2.2.
- **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.

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.
- **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.

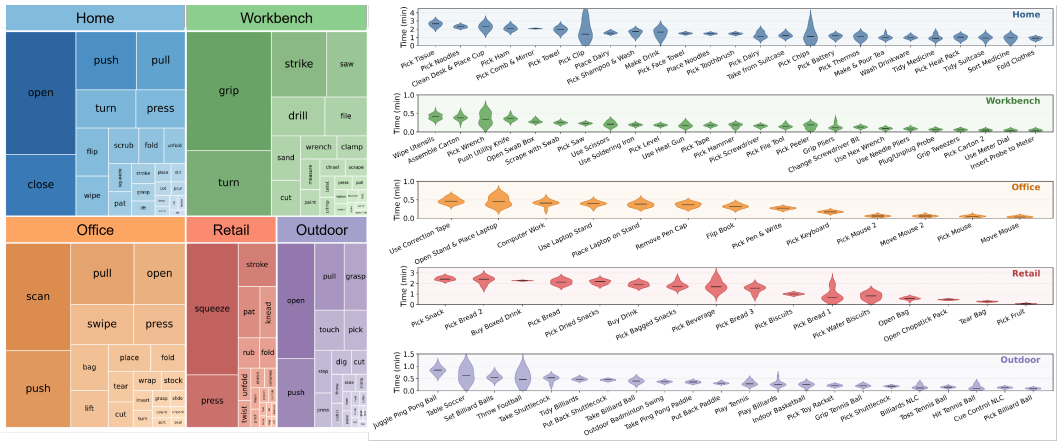


Figure 4: Dataset overview across five environment categories. Left: interaction taxonomy. Right: duration distributions in minutes for the displayed tasks, grouped by environment.

3.1 Problem Formulation

Given a video clip of T frames from views $\{V^{ego}\} \subseteq \mathcal{V} \subseteq \{V^{ego}, V^{wL}, V^{wR}\}$ and the corresponding ego-derived 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 and spatial remapping of the raw tactile sensor layout (Appendix A.2.2).

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 view is encoded by a shared, frozen DINOv2 backbone (Oquab et al., 2023). Adding a learned camera-identity embedding gives patch features $\mathbf{F}_{t,v,n} \in \mathbb{R}^D$, where t, v, and n index frame, view, and patch. We average the N patches within each frame and view to obtain summary tokens $\mathbf{s}_{t,v} = N^{-1}\sum_n \mathbf{F}_{t,v,n}$. A cross-view transformer updates these summaries, and a shared MLP maps each updated summary to a scalar logit $a_{t,v}$. Fusion normalizes over the available views:

$$w_{t,v} = \frac{\exp(a_{t,v})}{\sum_{u\in\mathcal{V}} \exp(a_{t,u})}, \qquad \mathbf{F}^{\text{fused}}_{t,n} = \text{LayerNorm}\left(\sum_{v\in\mathcal{V}} w_{t,v}\mathbf{F}_{t,v,n}\right). \tag{1}$$

The weights vary by frame and are shared across patch positions. Fusion retains N tokens regardless of view count, keeping the downstream representation fixed. With ego alone, cross-view attention is bypassed and ego features are directly layer-normalized. The weighted patch mixture operates in feature space; identical patch indices across cameras do not imply geometric correspondence.

The fused visual features first undergo temporal modeling within each input clip. A pose encoder maps the ego-derived hand joints to joint-level features, which then query the temporally enhanced visual tokens through cross-attention. This pose-conditioned attention learns associations between hand configurations and visual evidence without explicit joint-to-image projection.

The fused joint features are split into left and right hands. A shared decoder projects each hand's features onto a learned 8×8 grid, then applies convolutional refinement and upsampling to produce a 21×21 normalized tactile map. Appendix A.3.4 specifies this joint-to-grid mapping.

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 each wrist view is independently dropped with probability

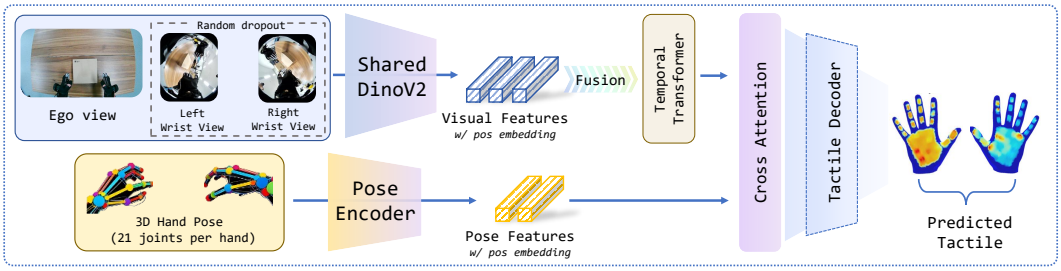


Figure 5: Architecture of the multi-view tactile prediction model. A shared backbone encodes each view with view embeddings. Cross-view attention and gated fusion produce unified visual features, which undergo temporal modeling before pose-conditioned cross-attention and joint-to-grid decoding into bilateral tactile maps.

$p = 0.3$. This exposes the model to all four input combinations and enables flexible inference at test time without architectural changes. The model uses egocentric RGB and the same ego-derived hand poses in every configuration, with optional wrist-mounted RGB inputs.

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

$$\mathcal{L} = \lambda_{mse}\mathcal{L}_{MSE} + \lambda_{l1}\mathcal{L}_{L1} + \lambda_{tv}\mathcal{L}_{TV}(\hat{\mathbf{M}}) \tag{2}$$

where $\mathcal{L}_{MSE}$ and $\mathcal{L}_{L1}$ measure reconstruction error over valid mapped locations, and $\mathcal{L}_{TV}$ regularizes adjacent predictions over the full output grid. Invalid targets are replaced with zero before loss computation and excluded from reconstruction by a sensor-validity mask. Appendix A.3.6 gives the exact reductions. To address sparse supervision, we upweight valid mapped cells whose normalized target response exceeds 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 partition episodes into training, validation, and test sets with an approximately 80% / 10% / 10% allocation. Episodes are kept intact to avoid sharing neighboring frames across splits. The test set contains two object-based subsets: *seen objects*, whose instances also appear in training episodes, and *unseen objects*, whose instances are excluded from training and validation. Evaluation covers the five environment categories. These splits assess object generalization; they do not hold out participant identities.

Evaluation metrics. The task is to predict bilateral tactile maps from visual observations; TouchAnything also encodes ego-derived hand pose. Contact maps are derived by thresholding both predictions and targets at $\tau = 0.1$ in normalized full-hand tactile units. We adapt the metrics of PressureVision (Grady et al., 2022; 2024) to the hand-shaped grids. Let Ω_e contain all valid mapped locations across both hands and evaluated frames of episode e; locations marked as NaN in the targets are excluded.

- **Temporal Accuracy** ↑: Frame-level agreement in contact presence. A frame is labeled as in contact when at least 5% of valid locations across both hands exceed τ, within a fixed contact-detection mask. This measures contact-state agreement, not onset or offset timing error.
- **Contact IoU** ↑: Intersection over union of thresholded prediction and target maps, pooled over Ω_e.
- **Volumetric IoU** ↑: Overlap of normalized tactile magnitudes, computed as

$$\mathrm{IoU}^{(e)}_{\mathrm{vol}} = \frac{\sum_{q\in\Omega_e} \min(M_q, \hat{M}_q)}{\sum_{q\in\Omega_e} \max(M_q, \hat{M}_q)}. \tag{3}$$

  Here q indexes frame, hand, and grid location. Neither Contact IoU nor Volumetric IoU is generally an upper bound on the other.

Table 2: Tactile prediction across five scenarios. PV++ is adapted PressureVision++; TA-Ego is independently trained. Other TouchAnything configurations share one multi-view-trained (MV) checkpoint. Bold marks the best reported score in each scenario or overall block. Percentages in the overall block show relative changes from Ego (MV); green indicates improvement.

	Seen Objects				*Unseen Objects*			
Model / input	**T.Acc ↑**	**C.IoU ↑**	**V.IoU ↑**	**MAE ↓**	**T.Acc ↑**	**C.IoU ↑**	**V.IoU ↑**	**MAE ↓**
Home Scenario								
PV++ (adapted)	0.369	0.250	0.178	0.060	0.354	0.210	0.149	0.064
TA-Ego	0.8418	0.4610	0.4103	0.0503	0.7881	0.3963	0.3474	0.0615
Ego (MV)	0.8532	0.4869	0.4361	0.0470	0.8175	0.4250	0.3671	0.0604
Ego + wL	**0.8659**	0.5074	**0.4598**	0.0453	**0.8278**	**0.4375**	**0.3779**	0.0591
Ego + wR	0.8658	0.5072	0.4597	0.0453	0.8276	0.4372	0.3776	0.0592
Ego + wL + wR	**0.8659**	**0.5075**	**0.4598**	**0.0452**	0.8267	0.4371	0.3775	**0.0590**
Workbench Scenario								
PV++ (adapted)	0.345	0.215	0.151	0.058	0.337	0.165	0.121	0.064
TA-Ego	0.7697	0.3776	0.3386	0.0465	0.7432	0.2537	0.2525	0.0615
Ego (MV)	0.7954	0.4131	0.3747	0.0420	**0.8096**	0.2932	0.2801	0.0599
Ego + wL	**0.8250**	**0.4422**	**0.4080**	**0.0395**	0.7525	0.3016	**0.2841**	0.0579
Ego + wR	0.8242	0.4408	0.4070	**0.0395**	0.7539	**0.3017**	0.2840	0.0579
Ego + wL + wR	0.8241	0.4414	0.4078	**0.0395**	0.7538	0.3015	0.2840	**0.0578**
Office Scenario								
PV++ (adapted)	0.379	0.254	0.182	0.059	0.391	0.250	0.169	0.066
TA-Ego	0.8562	0.4494	0.4071	0.0503	0.8362	0.5271	0.4183	0.0655
Ego (MV)	0.8770	0.4918	0.4501	0.0436	0.8901	**0.5743**	0.4511	0.0632
Ego + wL	0.8990	0.5249	0.4889	0.0401	**0.9035**	0.5718	**0.4673**	**0.0615**
Ego + wR	0.9001	0.5250	0.4891	0.0400	**0.9035**	0.5718	0.4672	**0.0615**
Ego + wL + wR	**0.9003**	**0.5256**	**0.4895**	**0.0399**	**0.9035**	0.5718	0.4672	**0.0615**
Retail Scenario								
PV++ (adapted)	0.341	0.231	0.163	0.055	0.285	0.195	0.142	0.034
TA-Ego	0.7541	0.3894	0.3786	0.0447	0.4895	0.3313	0.3387	0.0326
Ego (MV)	0.7907	0.4454	0.4112	0.0411	0.5841	0.3935	0.3636	0.0313
Ego + wL	**0.8334**	**0.4901**	**0.4414**	0.0392	0.6591	**0.4011**	0.3811	0.0310
Ego + wR	0.8241	0.4851	0.4390	0.0395	**0.6670**	0.4008	**0.3812**	0.0310
Ego + wL + wR	0.8316	0.4900	0.4411	**0.0391**	0.6591	0.4007	0.3810	**0.0309**
Outdoor Scenario								
PV++ (adapted)	0.352	0.271	0.191	0.062	0.438	0.228	0.151	0.105
TA-Ego	0.7765	0.5094	0.4515	0.0487	0.9384	0.4607	0.3556	0.1082
Ego (MV)	0.7905	0.5339	0.4769	0.0456	**0.9745**	0.4879	0.3750	0.1071
Ego + wL	0.8060	0.5533	**0.5002**	**0.0439**	0.9628	0.4978	0.3850	**0.1043**
Ego + wR	**0.8062**	0.5530	0.5001	**0.0439**	0.9628	**0.4986**	**0.3854**	**0.1043**
Ego + wL + wR	0.8061	**0.5534**	**0.5002**	**0.0439**	0.9628	0.4981	0.3851	**0.1043**
Overall (All Scenarios)								
PV++ (adapted)	0.364	0.246	0.175	0.060	0.359	0.213	0.150	0.065
TA-Ego	0.8238	0.4493	0.4023	0.0493	0.7923	0.4078	0.3526	0.0628
Ego (MV)	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%**

- **MAE ↓**: Mean absolute error over Ω_e. This global average can be dominated by zero-valued non-contact locations.

Contact IoU, Volumetric IoU, and MAE use all valid locations in the full-hand tactile maps. We average episode-level scores equally, omitting undefined scores separately for each metric. A zero union makes an episode's IoU undefined. Appendix A.3.7 specifies the masks, aggregation, and empty-contact cases.

Implementation. TouchAnything uses a frozen DINOv2-Base backbone, 224×224 RGB inputs, 8-frame clips sampled at interval 2, and 42 WiLoR hand joints. Evaluation is offline, with bidirectional attention within each clip. We train for 25 epochs with AdamW (learning rate 5×10^{-5}, weight decay 0.05) and a cosine schedule with 10 warmup epochs. Each wrist view is independently dropped with probability 0.3 while ego is retained. The reconstruction objective and detailed optimization settings are given in Appendix A.3.6.

Baselines. We separately train *TA-Ego* using only ego RGB, retaining the temporal and pose modules. We also adapt PressureVision++ (Grady et al., 2024) (*PV++*) to predict canonical tactile maps from single-frame ego hand crops, with shared weights across hands. WiLoR is used only for crop localization in PV++, whereas TouchAnything encodes its joint estimates. This external baseline

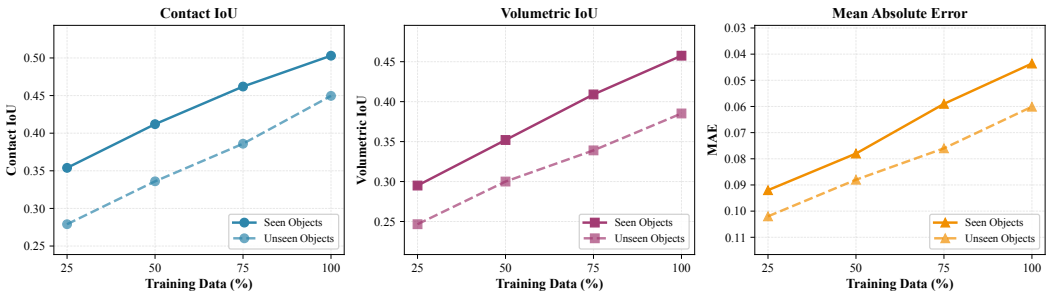


Figure 6: Data scaling ablation. The 100% endpoints correspond to the full-view Overall results in Table 2. Contact IoU, Volumetric IoU, and MAE improve over the evaluated training-data range.

comparison therefore includes differences in temporal context and pose conditioning. Appendix A.4 specifies the adaptation.

4.2 Main Results

Table 2 separates training-view and inference-view comparisons. With ego-only inference, the multi-view-trained model outperforms independently trained TA-Ego: Contact IoU increases from 0.4493 to 0.4792 on seen objects and from 0.4078 to 0.4396 on unseen objects; Volumetric IoU increases from 0.4023 to 0.4311 and from 0.3526 to 0.3743, respectively. These results support a benefit from multi-view training even when wrist RGB is unavailable at deployment.

Within the multi-view-trained checkpoint, adding wrist RGB further increases Contact IoU to 0.5030/0.4496 and Volumetric IoU to 0.4575/0.3852 on seen/unseen objects. Thus, the overall comparison distinguishes gains associated with multi-view training from those obtained by providing wrist views at inference.

PV++ (adapted) obtains overall Volumetric IoU of 0.175/0.150 on seen/unseen objects, compared with 0.4023/0.3526 for TA-Ego. This external comparison combines differences in backbone, temporal context, pose conditioning, and output formulation, so the gap cannot be attributed to multi-view fusion alone.

An object-generalization gap remains with all views: overall Volumetric IoU falls from 0.4575 on seen objects to 0.3852 on unseen objects, while MAE rises from 0.0436 to 0.0601. Additional viewpoints improve predictions within each split, but do not eliminate the challenge of estimating tactile responses for objects absent from training.

The gains vary across metrics and scenarios: unseen Workbench Temporal Accuracy decreases from 0.8096 to 0.7538 (5.58 percentage points), and unseen Outdoor Temporal Accuracy and Office Contact IoU also decrease. One wrist accounts for most of the aggregate gain: Ego+wL matches full-view IoUs on seen objects and slightly exceeds them on unseen objects. Establishing a benefit from the second wrist under severe bimanual occlusion requires stratified evaluation.

4.3 Ablation Studies

View dropout. Table 3 shows improved robustness to missing wrist views: dropout reduces the relative V.IoU decrease from all-view to ego-only inference from 27.20% to 5.78% on seen objects.

Full-view Contact IoU also rises from 0.4883 to 0.5030 and Volumetric IoU from 0.4441 to 0.4575, consistent with a regularizing effect.

Performance scales with data. Training on 25%, 50%, 75%, and 100% of the dataset improves Contact IoU, Volumetric IoU, and MAE over the evaluated range (Figure 6). From 25% to 100%, unseen-object Volumetric IoU increases from 0.2466 to 0.3852 and MAE decreases from 0.1020 to 0.0601, showing that the trend extends to held-out object instances. These results do not establish saturation or separate data diversity from optimization budget.

Table 3: View-dropout ablation on seen objects. Ego only is an inference condition, not TA-Ego. Ego + wL/wR averages the single-wrist configurations; ΔV is the relative V.IoU drop from all-view inference within each training strategy ($p = 0.3$ for dropout).

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**	**–**

4.4 Qualitative Results

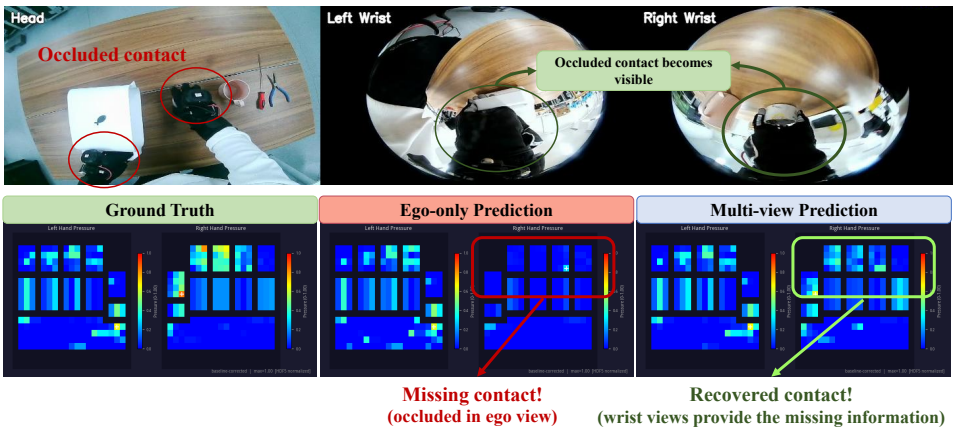


Figure 7: **Qualitative view ablation of the multi-view-trained model.** Top: wrist views reveal hand–object contact occluded in ego. Bottom: adding wrist RGB recovers contact missed with ego-only inference; ground truth is shown for reference.

Figure 7 illustrates contact missed under egocentric occlusion and recovered with wrist views. Both predictions use the same multi-view-trained checkpoint and ego-derived poses, isolating the change in available RGB evidence. The highlighted regions illustrate improved contact localization; they do not establish the frequency of such recoveries. Appendix A.7 presents a complementary failure in which visual proximity is mistaken for contact.

5 Conclusion

We introduced EgoTouch, a dataset pairing multi-view RGB, bimanual pose, and full-hand tactile signals, and TouchAnything, a baseline with flexible view availability. The independent ego-only comparison and shared-checkpoint view ablations support benefits from multi-view training and wrist-assisted inference. EgoTouch provides a basis for studying tactile-grounded perception and learning from human manipulation.

6 Limitations and Future Work

EgoTouch is collected from 10 participants wearing tactile gloves, without a held-out-participant split; transfer to bare hands and unseen participants remains unevaluated. Wearable hardware may alter dexterity and grasping. Visual appearance does not uniquely determine internal forces, and low-contrast boundaries can cause false contact predictions (Appendix A.7). Results concern normalized tactile maps, without evaluating force calibration in physical units. Bidirectional clip attention also leaves causal, real-time deployment unevaluated. Occlusion-stratified evaluation, repeated-seed studies, and broader scaling experiments would help distinguish viewpoint benefits from dataset-specific effects.

AI USE STATEMENT

Generative AI assisted with manuscript writing and review, baseline implementation, and consistency checks. The authors are responsible for verifying scientific claims, experimental descriptions, and final revisions.

REPRODUCIBILITY STATEMENT

The acquisition procedure and tactile-grid representation are described in Appendices A.2.1 and A.2.2. Architecture, losses, optimization settings, and evaluation reductions are specified in Appendix A.3; Appendix A.4 describes TA-Ego and the PressureVision++ adaptation.

REFERENCES

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.

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*, pp. 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.

Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8701–8711. IEEE, 2019.

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9040–9049. IEEE, 2021.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130(1):33–55, 2022.

Joseph DelPreto, Chao Liu, Yiyue Luo, Michael Foshey, Yunzhu Li, Antonio Torralba, Wojciech Matusik, and Daniela Rus. Actionsense: A multimodal dataset and recording framework for human activities using wearable sensors in a kitchen environment. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 13800–13813. Curran Associates, Inc., 2022. doi: 10.52202/068431-1003. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/5985e81d65605827ac35401999aea22a-Paper-Datasets_and_Benchmarks.pdf.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12943–12954. IEEE, 2023.

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.

Patrick Grady, Chengcheng Tang, Samarth Brahmbhatt, Christopher D. Twigg, Chengde Wan, James Hays, and Charles C. Kemp. Pressurevision: Estimating hand pressure from a single RGB image. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VI*, volume 13666 of *Lecture Notes in Computer Science*, pp. 328–345. Springer, 2022. doi: 10.1007/978-3-031-20068-7_19. URL https://doi.org/10.1007/978-3-031-20068-7_19.

Patrick Grady, Jeremy A Collins, Chengcheng Tang, Christopher D Twigg, Kunal Aneja, James Hays, and Charles C Kemp. Pressurevision++: Estimating fingertip pressure from diverse rgb images. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8698–8708. IEEE, 2024.

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

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.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11080–11090. IEEE, 2022.

Ryan Hoque, Peide Huang, David Yoon, Jian Zhang, et al. Egodex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations*, volume 2026, pp. 4218–4237, 2026.

Umar Iqbal, Pavlo Molchanov, Thomas Breuel, Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5d heatmap regression, 2018. URL https://arxiv.org/abs/1804.09534.

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13226–13233. IEEE, 2025.

Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, et al. 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.

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.

Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting touch and vision via cross-modal prediction, 2019. URL https://arxiv.org/abs/1906.06322.

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*, pp. 10806–10812, 2023. doi: 10.1109/IROS55552.2023.10341888. URL https://doi.org/10.1109/IROS55552.2023.10341888.

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20981–20990. IEEE, 2022.

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.

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.

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*, pp. 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.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12242–12254. IEEE, 2025.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6):245:1–245:17, 2017. doi: 10.1145/3130800.3130883. URL https://mano.is.tue.mpg.de/.

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.

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. *CoRR*, abs/2512.16842, 2025. doi: 10.48550/ARXIV.2512.16842. URL https://doi.org/10.48550/arXiv.2512.16842.

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*, pp. 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.

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.

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*, pp. 581–600. Springer, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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.

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 (eds.), *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*, 2022a. URL http://papers.nips.cc/paper_files/paper/2022/hash/354892587fe39b17c2b727af02abff4a-Abstract-Datasets_and_Benchmarks.html.

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.

Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20921–20930. IEEE, 2022b.

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.

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.

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

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.

Yiming Zhao, Taein Kwon, Paul Streli, Marc Pollefeys, and Christian Holz. Egopressure: A dataset for hand pressure and pose estimation in egocentric vision. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27727–27738. IEEE, 2025.

Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, et al. Egoscale: Scaling dexterous manipulation with diverse egocentric human data. *arXiv preprint arXiv:2602.16710*, 2026.

Shaohong Zhong, Alessandro Albini, Oiwi Parker Jones, Perla Maiolino, and Ingmar Posner. Touching a nerf: Leveraging neural radiance fields for tactile sensory data generation. In *Conference on Robot Learning*, pp. 1618–1628. PMLR, 2023.

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*, pp. 813–822. IEEE, 2019. doi: 10.1109/ICCV.2019.00090. URL https://doi.org/10.1109/ICCV.2019.00090.

A Appendix

A.1 Related Work

A.1.1 Egocentric Hand-Object Interaction Datasets

Recent egocentric datasets (Seo et al., 2026; Lee et al., 2026; Yoon et al., 2026; Perrett et al., 2025; Su et al., 2025) have substantially advanced the study of hand-object interaction from a first-person perspective. Ego4D (Grauman et al., 2022) and EPIC-KITCHENS (Damen et al., 2022) 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 (Hoque et al., 2026) 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 (Zhao et al., 2025) 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 (Song et al., 2025) 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.

These datasets provide limited support for studying bilateral tactile prediction with complementary head- and wrist-mounted views. Palmar contact regions are often occluded in the head-mounted image. EgoTouch pairs these viewpoints with bimanual hand pose and full-hand tactile signals, allowing the contribution of additional visual evidence to be evaluated under a shared prediction protocol. Wrist views can expose hand–object interaction cues hidden in ego, although the physical contact interface may remain occluded.

A.1.2 Vision-to-Touch Prediction

While hardware-based tactile sensors such as GelSight (Yuan et al., 2017) and DIGIT (Lambeta et al., 2020) 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 (Grady et al., 2022) predicts hand pressure maps from a single RGB image using a convolutional network. VisGel (Li et al., 2019) learns cross-modal representations between vision and touch using paired GelSight data. Touching a NeRF (Zhong et al., 2023) leverages neural radiance fields to synthesize tactile signals from 3D geometry. EgoPressure (Zhao et al., 2025) 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.

Egocentric tactile prediction remains challenging when contact regions are occluded, especially on the palmar surfaces. Complementary viewpoints can expose additional contact-related cues, while physical contact and force magnitude remain partly ambiguous. TouchAnything uses wrist-mounted RGB as optional evidence alongside ego RGB and ego-derived hand pose. Our quantitative comparisons evaluate aggregate changes across view configurations; the qualitative examples illustrate cases in which wrist views recover contact missed in ego-only inference.

A.1.3 MULTI-VIEW LEARNING FOR HAND INTERACTION

Multi-view learning (Aryal et al., 2026; Banerjee et al., 2025) 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 (Hampali et al., 2022; Zimmermann et al., 2019; Iqbal et al., 2018). In egocentric perception, Ego-Exo4D (Grauman et al., 2025) combines egocentric and exocentric views to improve activity understanding, highlighting the benefits of cross-view complementarity.

Many multi-view approaches target geometric reconstruction or activity understanding. EgoTouch instead supports prediction of full-hand tactile signals from head- and wrist-mounted observations. The wrist views provide close-range evidence of hand–object interactions, and the benchmark tests how predictions change when these views are available. Quantifying whether their benefit increases with occlusion severity requires a separate visibility-stratified analysis.

A.2 ADDITIONAL DATASET DETAILS

A.2.1 DATA COLLECTION SETUP

We collect a multimodal dataset of bimanual interactions using a wearable system that associates RGB, hand kinematics, and tactile feedback on a common 30 Hz frame timeline.

Hardware Configuration. Our collection platform combines three RGB viewpoints, bimanual hand kinematics, full-hand tactile sensing, and global tracker poses. The cameras are configured at 640×480 and 30 FPS: one head-mounted camera records the egocentric scene, while two wrist-mounted cameras face the left and right hand–object interaction regions. The software alignment procedure is described below.

For global spatial localization, we attach three HTC Vive Trackers to the head/ego rig and both wrists. The trackers are accessed through OpenXR with the `XR_HTCX_vive_tracker_interaction` extension and are bound to the roles `ego`, `left_wrist`, and `right_wrist`. For each valid tracker update, the system records a 6-DoF pose consisting of a 3D translation and a quaternion rotation in a shared OpenXR reference frame. These poses provide the global motion of the head-mounted and wrist-mounted sensing platforms.

Bimanual hand kinematics are recorded using Rokoko motion-capture gloves. The stream is received through UDP, parsed into 21 joints per hand, and cached in Redis for acquisition. We also provide a separate sequence of WiLoR (Potamias et al., 2025) estimates computed from the egocentric RGB video. The benchmark uses these vision-estimated poses throughout training and evaluation, including ego-only and wrist-assisted configurations. Captured Rokoko annotations are an additional dataset modality, distinct from the model inputs in the reported benchmark.

Dense tactile feedback is captured by custom pressure-sensing gloves. Each hand transmits a 256-channel 8-bit tactile frame through a serial connection at 921600 baud, together with a 16-byte IMU packet that is decoded into a quaternion. The tactile packet is parsed from a header-based binary protocol: two 128-byte sensor packets are combined into a complete 16×16 pressure array for each hand. The raw tactile array and the glove IMU quaternion are saved separately for the left and right hands.

Data Acquisition Pipeline. The collection software runs all high-rate sensor interfaces in asynchronous threads and constructs a frame-level snapshot at a target rate of 30Hz. The main loop creates a directory for each captured frame and writes all available modalities into the same folder. For each frame, the system performs the following operations:

- queries the latest OpenXR poses for the head/ego, left wrist, and right wrist trackers, and saves them as `ego_pose.txt`, `left_pose.txt`, and `right_pose.txt`;
- captures RGB frames from the three cameras and saves them as `ego_color.png`, `left_color.png`, and `right_color.png`;
- reads the latest Rokoko hand joints from Redis and saves the left and right 21×3 joint arrays;

- reads the latest complete tactile packets from the serial-reader state and saves the left/right raw pressure arrays and IMU quaternions.

This per-frame directory structure makes all modalities directly associated with a common frame index, simplifying downstream preprocessing, visualization, and model training. The collection code also stores camera intrinsics for the RGB stream, enabling geometry-aware post-processing when needed.

Synchronization Strategy. We use a software synchronization strategy designed for robust in-the-wild data collection. Camera frames are captured at 30Hz, while Rokoko hand pose, tactile packets, and Vive Tracker poses update asynchronously at their native rates. Each reader continuously maintains its most recent valid measurement in memory. At every 30Hz collection tick, the main process samples the latest value from each sensor stream and writes an atomic multi-modal snapshot to disk. This avoids blocking the capture loop on slower or temporarily unavailable devices and associates each snapshot with a common 30Hz frame index. The age of each cached measurement depends on its acquisition and transport latency; a shared frame index alone does not establish a bound on inter-sensor timing error. Frames with missing RGB views are dropped, whereas temporarily missing auxiliary streams can be recovered from the nearest valid cached measurement or supplemented during post-processing.

A.2.2 Tactile Grid Mapping and Preprocessing

The full-hand tactile glove records 256 values per hand. We normalize the full-hand tactile signals to $[0, 1]$ and remap them to canonical 21×21 hand-shaped grids using a fixed mapping for each hand. The mapping associates glove sensor readings with grid locations corresponding to the fingers and palm. The right-hand grid is mirrored to follow the canonical layout.

Locations outside the valid tactile layout are marked as NaN in the stored grids. During training, these values are replaced with zero and excluded from the reconstruction loss using a binary validity mask. Evaluation uses the same target validity information to exclude locations without tactile readings. The loss and metric reductions are specified in Appendices A.3.6 and A.3.7.

A.2.3 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 uses the three videos, processed tactile grids, and hand-pose annotations associated by the acquisition pipeline. 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. We truncate the camera streams to their minimum valid length for compatible indexing. This truncation preserves the acquisition indexing but does not correct inter-sensor timing offsets.

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 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 script also supports regenerating trajectories listed in a bad-file list, filtered by failure reason. The HDF5 format packages the cleaned RGB, hand-pose annotations, and canonical tactile grids for data loading.

A.3 Additional Implementation Details

The model predicts bilateral tactile maps from egocentric RGB, ego-derived hand poses, and zero, one, or two wrist-mounted RGB views. View dropout supports these four configurations using the same trained model.

A.3.1 Problem Formulation

Given a video clip of T frames from views $\{V^{ego}\} \subseteq \mathcal{V} \subseteq \{V^{ego}, V^{wL}, V^{wR}\}$ and the corresponding bimanual hand pose sequence $\mathbf{P} \in \mathbb{R}^{T\times 42\times 3}$, our goal is to predict the bilateral pressure maps $\hat{\mathbf{M}} \in \mathbb{R}^{T\times 2\times 21\times 21}$ for both hands at each timestep. The pose input is estimated from the egocentric stream and remains the same across the four RGB configurations.

A.3.2 Multi-View Vision Encoder

Shared backbone with view embeddings. All views are processed by a shared DINOv2-ViT-B/14 (Oquab et al., 2023) 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 $v \in \{ego, wL, wR\}$:

$$\mathbf{F}_v = \text{DINOv2}(V_v) + \mathbf{e}_v, \quad \mathbf{F}_v \in \mathbb{R}^{T\times N\times D} \tag{4}$$

Sharing the backbone avoids storing a separate visual encoder for each view. The view embeddings add 3×768 parameters; each available view still requires its own backbone forward pass.

Cross-view attention. For each frame t, we average over the N spatial tokens of each available view and apply a transformer to the resulting view summaries:

$$\mathbf{s}_{t,v} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{F}_{t,v,n}, \qquad [\hat{\mathbf{s}}_{t,v}]_{v\in\mathcal{V}} = \text{CrossViewTransformer}([\mathbf{s}_{t,v}]_{v\in\mathcal{V}}). \tag{5}$$

Pooling does not reduce the temporal dimension. Cross-view attention operates independently at each frame, producing a $T \times |\mathcal{V}| \times D$ tensor of updated summaries.

Gated view fusion. A shared MLP maps each updated summary to a scalar $a_{t,v}$. We normalize the logits over the available views and mix their patch features:

$$w_{t,v} = \frac{\exp(a_{t,v})}{\sum_{u\in\mathcal{V}} \exp(a_{t,u})}, \qquad \mathbf{F}^{fused}_{t,n} = \text{LayerNorm}\left(\sum_{v\in\mathcal{V}} w_{t,v}\mathbf{F}_{t,v,n}\right). \tag{6}$$

Each frame's view weights are shared across patch positions. The output has shape $T \times N \times D$. With only the ego view, cross-view attention is bypassed and its patch features are layer-normalized directly. This compact fusion retains a fixed number of tokens but does not establish geometric correspondences between different cameras' patch grids.

A.3.3 Temporal Modeling and Pose-Vision Fusion

Temporal transformer. A transformer (Vaswani et al., 2017) processes the fused visual features within each $T = 8$ input clip. Each layer applies temporal attention across frames at each patch index, spatial attention across patches within each frame, and a feed-forward block. Temporal sinusoidal position encodings are added before these layers:

$$\mathbf{H} = \text{TemporalTransformer}(\mathbf{F}^{fused}), \quad \mathbf{H} \in \mathbb{R}^{T\times N\times D} \tag{7}$$

Attention is bidirectional within the clip; the benchmark is an offline clip-based estimation setting. The temporal positional-encoding capacity is 16, while the evaluated clips contain 8 frames.

Pose encoder. The input $\mathbf{P} \in \mathbb{R}^{T\times 42\times 3}$ concatenates 21 WiLoR joints for each hand in the egocentric camera coordinate frame. Camera translation is retained, so the two hands share a camera-relative frame. Before encoding, coordinates are clipped to $[-10, 10]$ for x, y and $[-10, 100]$ for z in the stored coordinate scale, and transformed as $(x/5, y/5, (z - 50)/50)$. A per-joint MLP embeds these coordinates into D dimensions; learned joint-identity embeddings and a transformer produce $\mathbf{G} \in \mathbb{R}^{T\times 42\times D}$. These fixed numerical transforms normalize the estimator outputs and do not constitute a physical calibration of monocular pose scale.

Pose-vision cross-attention fusion. Each joint token queries the temporally enhanced visual tokens through pose-conditioned cross-attention:

$$\mathbf{Z} = \text{CrossAttn}(Q{=}\mathbf{G},\ K{=}\mathbf{H},\ V{=}\mathbf{H}), \quad \mathbf{Z} \in \mathbb{R}^{T\times 42\times D} \tag{8}$$

The fusion block applies cross-attention followed by joint self-attention and a feed-forward block at each frame, with residual connections and normalization. It learns feature associations conditioned on the estimated pose; camera extrinsics and explicit joint-to-image projection are not inputs to this block.

A.3.4 Joint-Level Tactile Decoder

The fused features are split into two groups of 21 joints. A shared linear projection with GELU and dropout maps each joint feature from $D = 768$ to $C = 128$ channels. For each hand h, a learned assignment matrix $\mathbf{A} \in \mathbb{R}^{21\times 64}$ scatters these features onto an 8×8 grid:

$$\alpha_{j,g} = \frac{\exp(A_{j,g})}{\sum_{g'=1}^{64}\exp(A_{j,g'})}, \qquad \mathbf{U}^h_{t,g} = \sum_{j=1}^{21} \alpha_{j,g}\,\text{Proj}(\mathbf{Z}^h_{t,j}).$$

Reshaping $\mathbf{U}^h_t$ yields $128\times 8\times 8$ features. Two 3×3 convolutional blocks refine them, with channels $128 \to 128 \to 64$. A stride-2 transposed convolution produces $32 \times 16 \times 16$ features, which are bilinearly resized to $32 \times 21 \times 21$. These blocks use batch normalization and GELU. A final 3×3 convolution and sigmoid produce the tactile map:

$$\hat{\mathbf{M}}^h_t = \sigma\big(\text{Conv}_{3\times 3}\big(\text{Upsample}(\text{Refine}(\mathbf{U}^h_t))\big)\big) \in [0,1]^{21\times 21}, \quad h \in \{left, right\}. \tag{9}$$

The projection, assignment matrix, refinement, and output layers are shared across the two hand branches. Their learned grid assignments are not prescribed anatomical or camera-projection correspondences.

A.3.5 View Dropout Training Strategy

The model is trained for four inference configurations, all of which retain the egocentric view. 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 any of these four configurations without fine-tuning. Each uses the same ego-derived poses, so adding or removing a wrist view changes only the RGB inputs.

Architecture configuration. The cross-view, pose, and pose–vision fusion transformers each use 2 layers; the temporal module uses 3 layers. All use feature dimension 768, 8 attention heads, feed-forward expansion ratio 4, and dropout 0.1. The visual backbone is shared and frozen. The decoder uses 128 projected channels and dropout 0.1. These settings refer to the tactile-map prediction model described above.

A.3.6 Training Objective and Optimization

Let $S_q \in \{0, 1\}$ mark valid mapped sensor locations in a training batch, and let q index batch item, time, hand, and grid location. Invalid target values are replaced with zero before arithmetic and assigned $S_q = 0$. Let $N_S = \max(1, \sum_q S_q)$, and set $w_q = 3$ when $M_q > 0.1$ and $w_q = 1$ otherwise. The weighted reconstruction terms are

$$\mathcal{L}_{MSE} = \frac{\sum_q S_q w_q (\hat{M}_q - M_q)^2}{N_S}, \qquad \mathcal{L}_{L1} = \frac{\sum_q S_q w_q |\hat{M}_q - M_q|}{N_S}.$$

The denominator counts valid locations rather than the sum of contact weights. Total variation is the sum of the mean absolute vertical and horizontal differences of the predicted maps:

$$\mathcal{L}_{TV} = \text{mean}\,|\hat{M}_{\ldots,r+1,c} - \hat{M}_{\ldots,r,c}| + \text{mean}\,|\hat{M}_{\ldots,r,c+1} - \hat{M}_{\ldots,r,c}|.$$

Each mean ranges over all adjacent pairs in its direction and all batch items, timesteps, and hands. Unlike reconstruction, this regularizer is applied to the full finite prediction grid, including edges next to invalid target locations; it does not take differences of NaN-valued targets. The objective is

$$\mathcal{L} = \lambda_{mse}\mathcal{L}_{MSE} + \lambda_{l1}\mathcal{L}_{L1} + \lambda_{tv}\mathcal{L}_{TV}, \tag{10}$$

with $\lambda_{mse} = 1.0$, $\lambda_{l1} = 0.5$, and $\lambda_{tv} = 0.01$. Thus, locations with target response above 0.1 receive reconstruction weight 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 canonical 21×21 hand-shaped tactile grids.

A.3.7 Inference and Evaluation Protocol

We perform batched inference on the evaluation episodes using the same preprocessing and model configuration as during training. Each frame has corresponding bimanual hand-pose input. We evaluate all frames covered by the model's clip predictions. The four RGB configurations of the multi-view-trained TouchAnything model—ego only, ego with either wrist view, and all three views—use the same trained checkpoint and evaluation episodes. TA-Ego uses its own independently trained checkpoint. These comparisons distinguish changes in training-view availability from changes in inference-time inputs.

Metric inputs and valid regions. Contact predictions are derived from the normalized full-hand tactile maps using $\tau = 0.1$ for both predictions and targets. Predictions from overlapping clips are averaged at each covered video frame before evaluation. Locations marked as NaN in the targets are masked out in both predictions and targets. Contact IoU, Volumetric IoU, and MAE use all valid locations in the full-hand tactile maps.

Contact-state classification. Temporal Accuracy uses a fixed contact-detection mask on the canonical grids, applied identically to predictions and targets. A frame is labeled as in contact when at least 5% of valid locations within this mask, pooled across both hands, have values strictly greater than 0.1. A frame with no eligible locations is assigned no contact. Frames with entirely invalid targets are omitted. The episode score is the fraction of evaluated frames with matching predicted and target contact states, including correctly predicted no-contact frames.

Episode aggregation and empty unions. For each episode, Contact IoU pools binary intersection and union over all evaluated frames, both hands, and valid mapped locations. Volumetric IoU similarly pools the minimum and maximum responses, and MAE averages absolute error over the same locations. Thus, the IoUs are ratios of episode-level sums, not means of per-frame IoUs. A frame with zero union contributes zero to both pooled sums and receives no separate IoU score. If the entire evaluated episode has zero union, the corresponding IoU is undefined and returned as NaN. MAE is undefined if no valid target locations remain. Scene and overall scores are unweighted means over the corresponding episode scores, excluding NaNs independently for each metric. Correctly empty episodes still contribute to Temporal Accuracy and to MAE when valid target locations exist, but not to an IoU whose union is zero. The overall score is therefore not generally the unweighted mean of the five scene scores.

A.4 SINGLE-VIEW BASELINES AND PRESSUREVISION++ ADAPTATION

Independently trained ego-only baseline. TA-Ego is trained and evaluated with only egocentric RGB. It retains TouchAnything's DINOv2 encoder, 8-frame temporal modeling, and WiLoR pose conditioning, but receives no wrist RGB during training. This is distinct from Ego (MV) in Table 2, which uses the multi-view-trained checkpoint with both wrist inputs removed at inference. The Ego only columns of Table 3 refer to the latter inference condition for each multi-view training strategy. Table 2 reports scenario-level and overall TA-Ego results on both seen and unseen objects.

PressureVision++ formulation and inputs. We adapt PressureVision++ (Grady et al., 2024) from image-space pressure prediction to bilateral canonical tactile grids. The adaptation processes each frame independently, using two 448×448 crops from the ego image and shared parameters for the left and right hands. WiLoR's ego-derived joint estimates provide crop localization only; no joint features, wrist images, or neighboring video frames enter the network. Crop boxes enclose the projected hand joints and are expanded to square boxes by a factor of 1.5 before clipping to image bounds. Unavailable or unprojectable estimates use the full ego frame as a fallback, preserving the evaluation frames. The stored canonical targets are used without additional mirroring; the current configuration also leaves RGB handedness unchanged.

Canonical prediction head. We retain the ImageNet-initialized SE-ResNeXt-50 encoder and FPN decoder, training both on EgoTouch. The FPN produces 128-channel features. A new head applies adaptive average pooling to 7×7, flattens the features, and uses a 512-unit fully connected layer with ReLU and dropout 0.2, followed by a linear projection to $9 \times 21 \times 21$ logits. This learned spatial mapping replaces the original image-space output head. Processing both hands yields logits of shape $2 \times 9 \times 21 \times 21$ per frame.

Pressure bins and supervision. We retain nine-class pressure prediction, adapting the bins to normalized tactile units. The upper boundaries and representative values are

$$\mathbf{e} = (0.005, 0.015, 0.035, 0.07, 0.1, 0.2, 0.4, 0.7, 1.0),$$
$$\mathbf{c} = (0, 0.01, 0.025, 0.0525, 0.085, 0.15, 0.3, 0.55, 0.85).$$

The first bin is $[0, e_0]$ and subsequent bins are $(e_{k-1}, e_k]$. This partition preserves several magnitude levels below the contact threshold rather than collapsing all values below 0.1 into a single category. At inference, we map the highest-logit class to its representative value, yielding scalar-valued tactile maps for the shared evaluation protocol.

For valid target locations $\mathcal{Q}$, the dense objective is the original structure-aware classification loss with $\omega = 1$:

$$\mathcal{L}_{\mathrm{PV++}} = -\frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \sum_{b=0}^{8} \exp(-|b - k_q|/\omega) \log p_q(b), \tag{11}$$

where k_q is the target bin. The exponential class weights are not normalized across classes, matching the original implementation. Invalid targets are excluded from the loss. The current adaptation uses dense tactile supervision only and omits the auxiliary contact-label and domain-adversarial branches used in the original fully/weakly supervised setup; it is therefore denoted PV++ (adapted).

Training and evaluation configuration. The current PV++ run is configured for 30,000 optimizer updates using Adam with an initial learning rate of 10^{-3}, reduced to 10^{-4} after 10,000 updates, and a batch of 14 frames (28 hand crops) on one GPU. Gradient norms are clipped at 5.0. We apply color jitter to crops without spatially transforming the anatomical tactile targets. Training uses the frame indices from the nonoverlapping 8-frame, interval-2 TouchAnything clips, flattened into single-frame examples; minibatch blocks are shuffled. Validation is scheduled every 1,000 updates, with checkpoint selection by episode-mean Volumetric IoU. The update budget is specific to this adaptation and does not claim to reproduce the original implementation's sample-count-based iteration counter.

The PV++ evaluation entry point uses episode splits and the frame indices covered by the TouchAnything sliding-window protocol. It applies the same $\tau = 0.1$ threshold, validity masks, temporal contact rule, and episode-level aggregation defined in Appendix A.3.7. Table 2 reports scenario-level

and overall PV++ test results on seen and unseen objects. Since PV++ is single-frame and does not encode pose features, its comparison with TouchAnything is an external method comparison rather than a controlled ablation of the architecture alone.

A.5 Existing Assets and Licenses

We credit the creators and original owners of all third-party assets used in the project through the citations in the main paper and the asset summary in Table 4. The listed assets are used for data acquisition, preprocessing, model training, or evaluation. We do not redistribute third-party datasets or proprietary hardware/software components with EgoTouch; the released code and dataset documentation will include the exact package versions, download sources, and any additional license notices required by the corresponding owners.

Prior datasets discussed in the related-work and dataset-comparison sections are not repackaged, redistributed, or used as training data unless explicitly stated otherwise. Implementations used in our experiments, including the PressureVision++ adaptation, are credited in Table 4; their original licenses and terms remain with their respective owners.

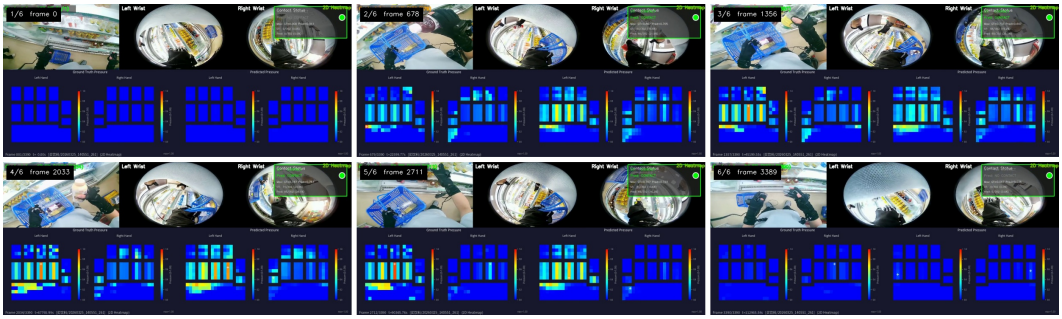
A.6 Additional Qualitative Results

Each task example contains six sampled timepoints arranged in a 2×3 panel grid. Within each timepoint, the top row shows the ego, left-wrist, and right-wrist RGB views. The lower row shows ground-truth left/right tactile maps on the left and predicted left/right maps on the right, with normalized response color scales. These examples illustrate contact patterns and their temporal evolution.

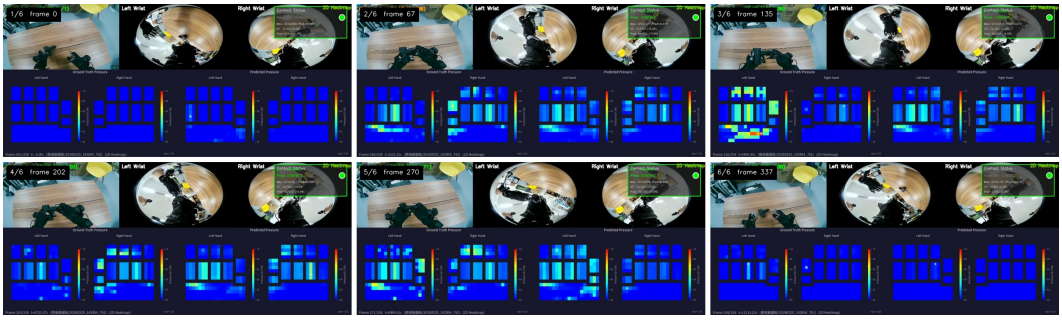
The examples cover varied object interactions and contact configurations. Quantitative generalization claims rely on the held-out evaluation in the main text; these selected visualizations provide complementary qualitative evidence.

Table 4: Third-party assets, licenses, and terms relevant to EgoTouch and TouchAnything.

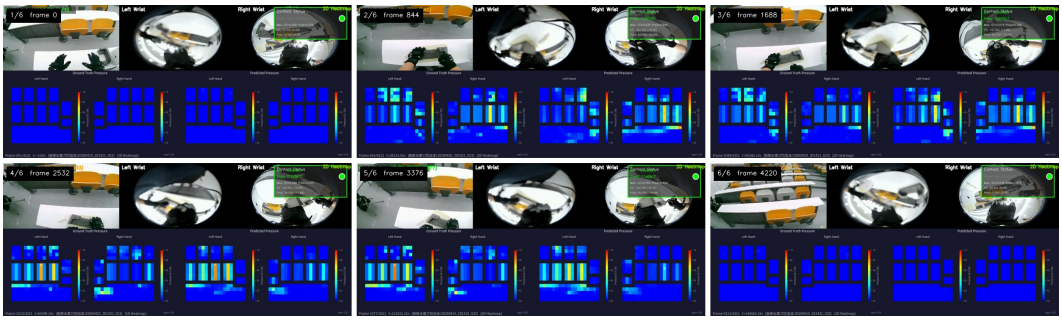
Asset	Use in this work	Source	License / terms	Release treatment
DINOv2 ViT-B/14 (Oquab et al., 2023)	Frozen visual backbone initialized from pretrained weights	https://github.com/facebookresearch/dinov2	Apache License 2.0 for the DINOv2 code and standard model weights	Credited by citation and URL; checkpoints are referenced according to the upstream license
PressureVision++ (Grady et al., 2024)	Adapted single-frame tactile baseline	https://github.com/pgrady3/pressurevision2	MIT License for the upstream code	Adaptation documented; upstream attribution and license retained
WiLoR (Potamias et al., 2025)	Ego-derived 3D hand-pose inputs for benchmark training and evaluation	https://github.com/rolpotamias/WiLoR	CC-BY-NC-ND for WiLoR models; dependent components, including Ultralytics and MANO, are governed by their own licenses	Used for pose annotation/preprocessing; users are directed to the upstream repository and terms
MANO hand model (Romero et al., 2017)	Dependency of WiLoR-style hand reconstruction pipelines	https://mano.is.tue.mpg.de/	PS:License 1.0 / research-use terms from MPI-IS	Not redistributed; users must obtain MANO from the official source if required
Ultralytics library	Dependency declared by WiLoR	https://www.ultralytics.com/license	AGPL-3.0 or Ultralytics commercial/enterprise license, depending on use case	Not redistributed as part of the dataset; downstream users must comply with the selected upstream license
PyTorch	Model implementation, training, and inference	https://pytorch.org/	BSD-3-Clause	Listed in the released environment specification with version information
OpenCV	RGB frame decoding and fallback video processing	https://opencv.org/license/	Apache License 2.0 for OpenCV 4.5.0 and later; BSD-3-Clause for earlier versions	Listed in the released environment specification with version information
FFmpeg	Video decoding and optional GPU-accelerated preprocessing	https://www.ffmpeg.org/legal.html	LGPL-2.1-or-later by default; GPL applies when GPL components are enabled	Not redistributed; users should install a build compatible with their intended license obligations
HDF5	Efficient storage format for preprocessed trajectories	https://www.hdfgroup.org/solutions/hdf5/	BSD-style license	Used as an output data format; library versions are documented in the release
OpenXR SDK	Access to Vive Tracker poses during data collection	https://github.com/KhronosGroup/OpenXR-SDK	Apache License 2.0 for the SDK	Used during acquisition; not bundled with the dataset release
Redis / Redis client	Low-latency cache for hand-pose streams during collection	https://redis.io/legal/licenses/	License depends on Redis version; Redis 7.2 and earlier are BSD-3-Clause, while newer Redis versions use the licenses stated by Redis Ltd.; Redis client libraries have their own licenses	Used only as local infrastructure during data collection; not redistributed with the dataset
Rokoko motion-capture gloves and software	Bimanual hand-pose acquisition	https://www.rokoko.com/terms	Proprietary hardware/software terms from Rokoko	Hardware/software not redistributed; only collected pose annotations are released subject to our dataset terms
HTC Vive Trackers and VIVE software	6-DoF head and wrist tracker poses	https://www.htc.com/us/terms/vive/	HTC VIVE end-user license agreements and terms of use	Hardware/software not redistributed; only recorded tracker poses are released subject to our dataset terms



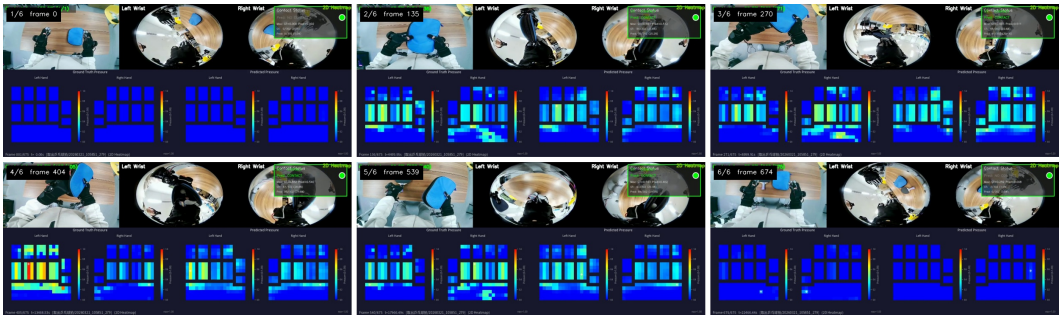
(a) Purchasing a beverage



(b) Using a massage gun

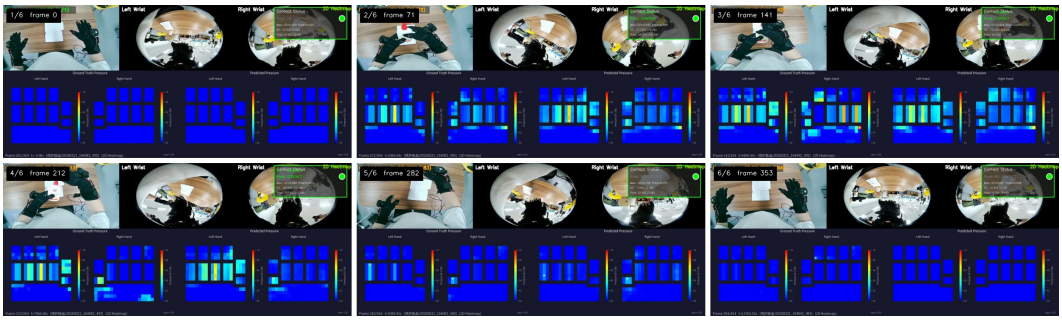


(c) Cutting foam with a kitchen knife

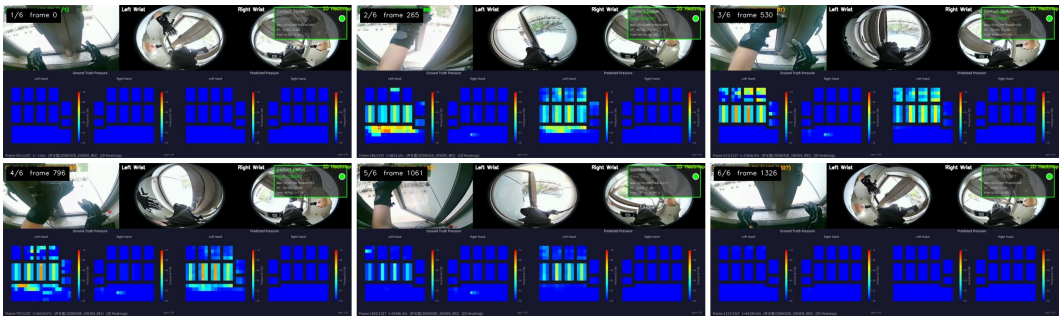


(d) Retrieving a ping-pong paddle

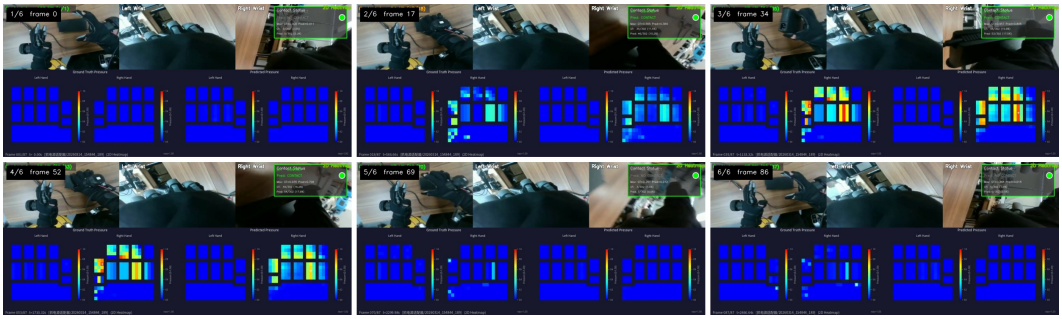
Figure 8: Tactile prediction examples (1–4). Six timepoints per task show the three RGB views above the bilateral ground-truth and predicted tactile maps. The maps are displayed side by side, not overlaid on RGB.



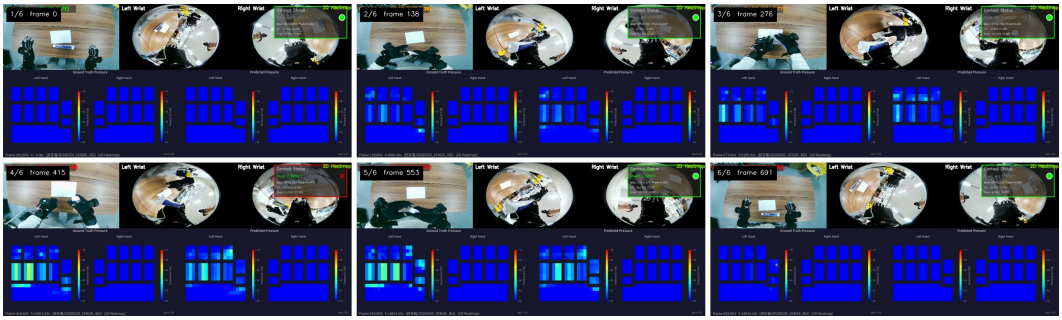
(a) Spraying skincare product



(b) Opening/closing a window

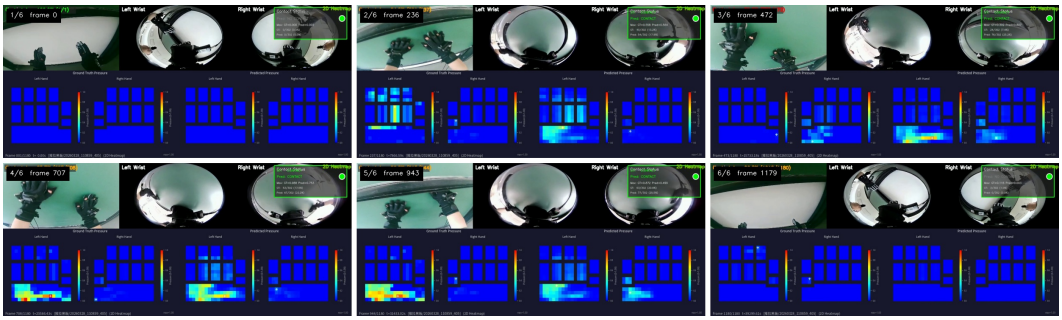


(c) Grasping a power adapter

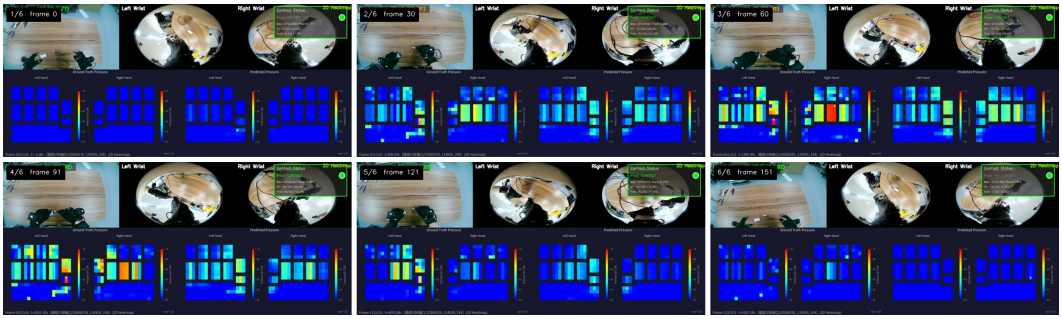


(d) Squeezing toothpaste

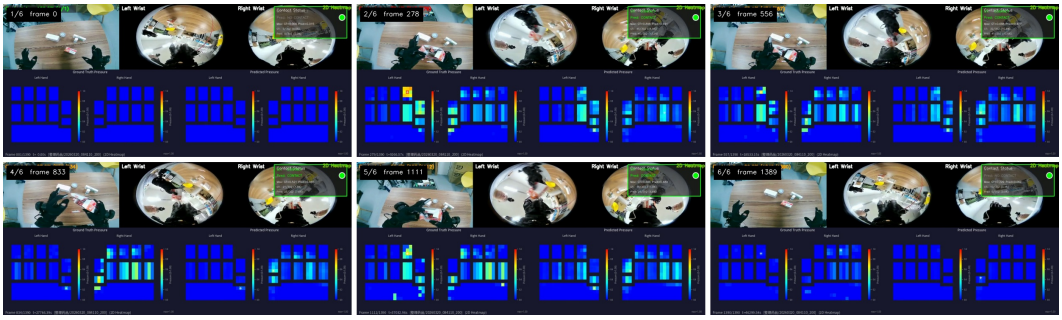
Figure 9: Tactile prediction examples (5–8), using the same RGB and tactile-map layout as Figure 8.



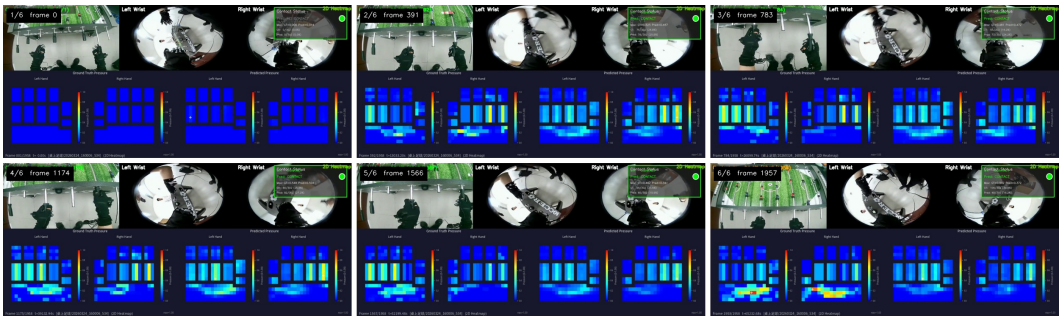
(a) Pushing/pulling a blackboard



(b) Plugging/unplugging a USB connector

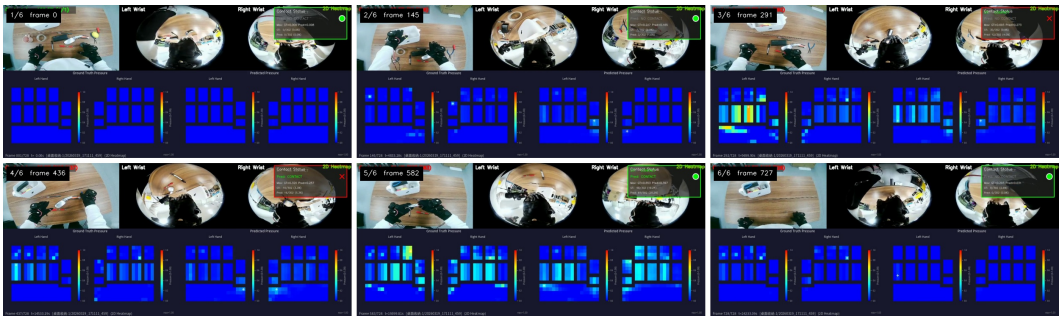


(c) Organizing medicine bottles

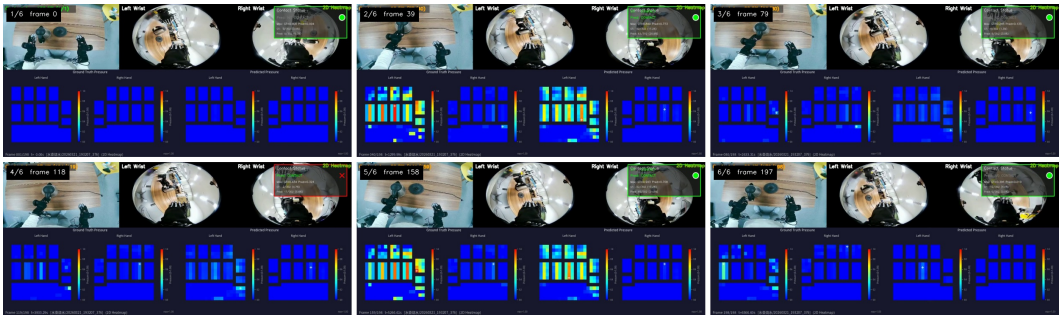


(d) Playing table football

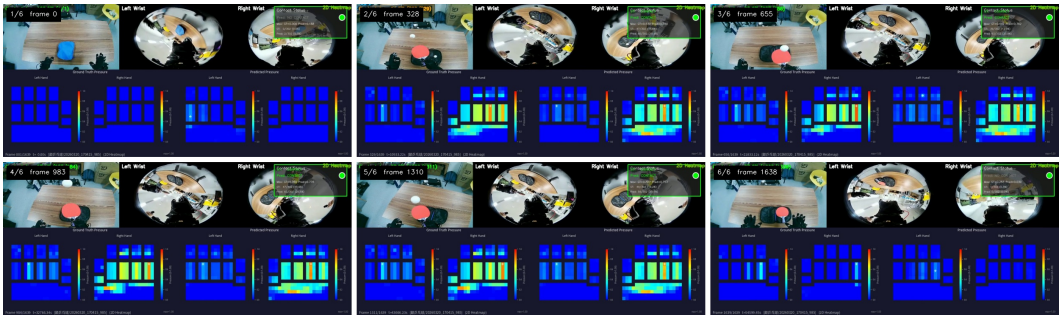
Figure 10: Tactile prediction examples (9–12), including precision manipulation and dynamic gameplay. Layout follows Figure 8.



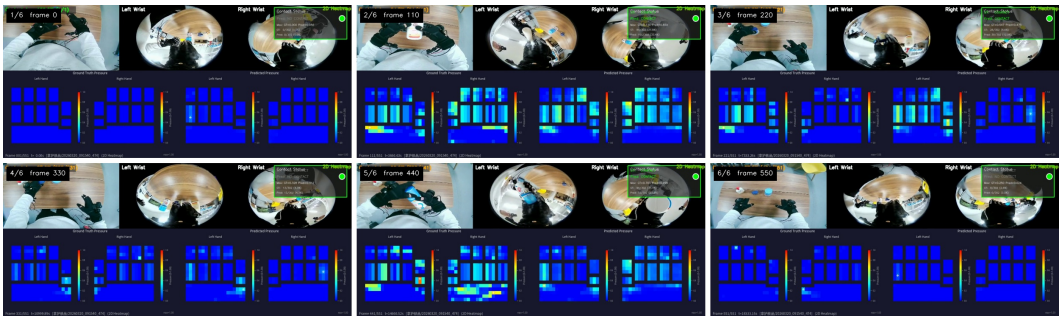
(a) Desktop organization



(b) Boiling water with a kettle



(c) Bouncing a ping-pong ball



(d) Grasping skincare products

Figure 11: Tactile prediction examples (13–16), showing RGB observations and bilateral target/prediction maps at six timepoints. Layout follows Figure 8.

A.7 Failure Cases and Limitations

Figure 12 shows a failure case in which the model predicts contact at the first displayed timepoint while the tactile ground truth indicates no contact.

In this example, the black glove and black shorts have low visual contrast, and their boundaries are difficult to distinguish in the egocentric view. The false contact prediction is consistent with confusion between visual proximity and physical contact. A single example does not isolate the cause, but illustrates an ambiguity that additional viewpoint information may not fully resolve.

Such failures motivate investigating explicit depth or hand–object segmentation and training examples with low-contrast boundaries. The current model already uses bidirectional temporal attention; additional temporal-consistency constraints would need to be evaluated for both false-positive suppression and preservation of brief real contacts.

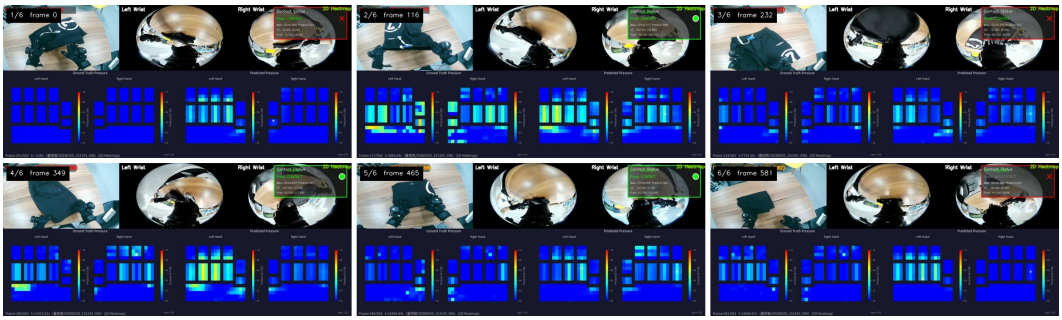


Figure 12: Failure case: folding black shorts. The model incorrectly predicts left-hand contact at the first displayed timepoint. The black glove and dark fabric have low-contrast boundaries, illustrating a setting in which visual proximity can be confused with contact.