TOUCHANNOTATOR: MULTI-SOURCE TACTILE SUPERVISION WITH PRIVILEGED-VIEW DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Egocentric video captures human–object interaction at scale but rarely includes tactile labels for hidden contact and pressure. We introduce TouchAnnotator, an ego-only bimanual tactile annotator trained with multi-source tactile supervision and privileged-view distillation. A canonical hand field unifies glove pressure, geometry-derived human contact, and robot fingertip deformation from six datasets and three embodiments, while source-aware masks preserve valid targets. During training, an EMA teacher observes synchronized ego and wrist views and distills hand-local representations and tactile predictions into a student that uses only ego video and estimated pose at inference. Compact past-clip memory supplies temporal context. We evaluate on EgoTouch, zero-shot FPHA transfer, and 860 synchronized glove recordings. The model transfers to bare-hand contact and produces qualitatively coherent tactile responses across bare-hand, gloved-hand, and robot embodiments. On measured recordings, it improves temporal accuracy by 5.38 points and Volumetric IoU by 5.02 points under a shared protocol. These results show that heterogeneous human–robot supervision and privileged training views can turn egocentric video into dense tactile annotations.

1 INTRODUCTION

Understanding manipulation requires knowing where the hands make contact and how that contact changes over time. Egocentric datasets capture diverse everyday activities and detailed hand–object motion (Damen et al., 2018; Grauman et al., 2022; Liu et al., 2022), but these visual records do not directly measure touch. A video may show a hand closing around a bottle while leaving the contact regions hidden behind the object. Annotating existing videos with contact and relative-pressure maps could add a useful source of supervision for understanding human interaction and learning from demonstrations, extending the value of recordings collected without tactile sensors.

Vision-based tactile prediction offers a route toward this goal. PressureVision and EgoPressure infer pressure from visual observations of hand–surface interaction (Grady et al., 2022; Zhao et al., 2025); OpenTouch and EgoTactile extend tactile data collection to everyday object manipulation (Song et al., 2025; Zeng et al., 2026); and EgoTac combines measured intensity with geometry-derived contact across human datasets (Zhang et al., 2026b). These sources provide complementary supervision but differ in sensor layout, spatial support, scale, embodiment, and whether they measure pressure, contact, or deformation. Moreover, an egocentric camera can miss the hand–object interface precisely when contact occurs, even when synchronized wrist views are available during data collection (Zhou et al., 2026). This raises a broader question: *how can heterogeneous tactile sources and privileged training views be combined to train a single annotator that requires only egocentric input at deployment?*

Related advances in egocentric learning motivate treating auxiliary observations as privileged training information. Ego-Exo4D provides synchronized first- and third-person views of skilled activities (Grauman et al., 2024), while Ego-Exo transfers representations learned from third-person videos to egocentric recognition (Li et al., 2021). Multimodal distillation similarly transfers complementary training signals into an RGB-only action recognizer (Radevski et al., 2023). These approaches connect to the broader principle of learning with information available only during training (Lopez-Paz et al., 2016). Tactile annotation adds a second challenge: supervision from different sensors and embodiments must share an anatomical representation without treating missing regions or incompatible targets as negatives.

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.

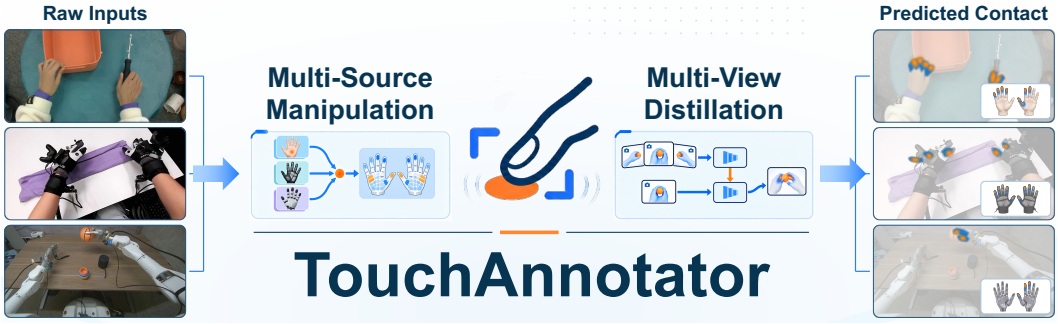

Figure 1: **Tactile annotation from egocentric video.** Human and robot supervision supports bimanual relative-pressure and contact estimates from ego RGB and pose estimated from that stream. The heatmaps are schematic and do not depict predictions or sensor readings.

We introduce TouchAnnotator, a framework built around multi-source tactile supervision and privileged-view distillation (Figure 1). First, a canonical bilateral hand field unifies glove pressure, geometry-derived human contact, and sparse robot fingertip deformation from six datasets and three embodiments. Source-aware masks preserve which hand regions and target types are valid, allowing heterogeneous signals to train a shared tactile representation without inventing supervision for unobserved cells.

Second, an exponential-moving-average teacher observes clean egocentric and synchronized wrist videos during training. An ego-only student learns from the teacher's hand-local representations and pressure/contact predictions while receiving degraded visual input. At deployment, the student uses only egocentric video and hand pose estimated from the same stream. A compact memory of preceding clips supplies temporal context, while the predicted intensity remains sensor-relative because visual observations do not uniquely determine physical pressure. We additionally collect egocentric video with synchronized glove measurements to validate predictions on everyday manipulation tasks.

Our contributions are threefold:

1. **Multi-source tactile supervision**: a canonical bimanual field with source-aware masks unifies pressure, contact, and deformation signals from six datasets and three human–robot embodiments.
2. **Privileged-view distillation**: an ego–wrist teacher transfers hand-local representations and tactile predictions to a student that requires only ego video and estimated pose at deployment.
3. **Extensive validation in real-world settings**: experiments on EgoTouch and FPHA, together with an evaluation on our synchronized glove recordings, examine tactile prediction, cross-dataset transfer, and agreement with measured touch.

Our experiments examine ego-only tactile prediction, cross-dataset transfer, and agreement with measured touch. On EgoTouch, the model predicts bimanual tactile fields from ego RGB and estimated hand pose, while additional visual evidence improves contact and intensity overlap for the same checkpoint. Without FPHA adaptation or threshold tuning, TouchAnnotator reaches F1 0.3658 on geometry-derived bare-hand contact. On 860 synchronized glove recordings, it improves temporal accuracy by 5.38 points and Volumetric IoU by 5.02 points under the shared sensor-grounded protocol. Qualitative examples further show task-relevant tactile responses across bare-hand, gloved-hand, and robot embodiments. These results support multi-source transfer and the use of privileged training views while keeping cross-protocol comparisons explicitly qualified.

2 Related Work

2.1 Egocentric Interaction and Tactile Prediction

Egocentric datasets span diverse activities (Damen et al., 2018; Grauman et al., 2022), RGB-D hand motion (Garcia-Hernando et al., 2018), and detailed hand–object geometry (Liu et al., 2022; Fan et al., 2023; Zhan et al., 2024; Liu et al., 2024). ContactDB records thermal grasp traces, while

Table 1: **Scale and diversity of the configured training corpus.** Counts cover unique segments before ratio sampling and native synchronized timesteps counted once across cameras. P., C., and Priv. denote relative pressure, contact, and non-ego teacher views; Geom. and Deform. denote geometry- and deformation-derived contact.

Source	Segments	Frames	Hours	P.	C.	Priv.
Tactile gloves: measured sensor responses						
EgoTouch (Zhou et al., 2026)	541	411,568	3.8	✓	✓	✓
HumanTouch (Spark Lab, 2026)	17,545	28,505,945	132.0	✓	✓	✓
Bare hands: geometry-derived contact						
HOI4D (Liu et al., 2022)	1,415	351,050	6.5	–	Geom.	–
TACO (Liu et al., 2024)	2,311	361,631	3.3	–	Geom.	–
OakInk2 (Zhan et al., 2024)	822	1,004,594	9.3	–	Geom.	–
Dexterous robot hands: optical deformation						
T-Rex (Niu et al., 2026)	4,371	4,363,007	40.4	–	Deform.	✓
Total	**27,005**	**34,997,795**	**195.3**			

DECO and HACO estimate dense human–scene or hand contact from RGB (Brahmbhatt et al., 2019; Tripathi et al., 2023; Jung & Lee, 2025).

HOI reconstruction uses contact as a geometric cue: Hasson et al. regularize a joint hand–object model with a contact loss, RHO fits interactions with 3D contact priors, and HOLD reconstructs both from monocular video without object templates (Hasson et al., 2019; Cao et al., 2021; Fan et al., 2024). InterField predicts dense hand–object distances on ARCTIC (Fan et al., 2023). G-HOP supplies a generative prior for reconstruction and grasps; ScoreHOI uses score-guided diffusion with contact refinement (Ye et al., 2024; Li et al., 2025). These methods estimate geometry and contact; pressure remains unmeasured.

Vision–touch learning connects cross-modal representations (Li et al., 2019; Yang et al., 2022) with visual pressure estimation (Grady et al., 2022; Zhao et al., 2025). OpenTouch, EgoTouch/TouchAnything, and EgoTactile extend tactile sensing and prediction to everyday manipulation (Song et al., 2025; Zhou et al., 2026; Zeng et al., 2026). TouchAnything supports ego-only inference through view dropout, while EgoTac combines measured intensity and geometry-derived contact with task-valid masks, dual heads, and short RGB histories (Zhou et al., 2026; Zhang et al., 2026b). We emphasize privileged multi-view distillation, bounded memory across clips, and source-masked robot deformation supervision.

2.2 Tactile policy learning and privileged supervision

Vision–language–action (VLA) models connect pretrained visual–language knowledge to robot control. RT-2 represents actions as tokens (Zitkovich et al., 2023); OpenVLA scales open-source policies across robot embodiments (Kim et al., 2025); π_0 uses flow matching for continuous actions (Black et al., 2024). These models motivate tactile feedback for physical interaction.

Tactile-VLA combines VLA priors with tactile feedback and hybrid position–force control, while OmniVTLA aligns touch with vision and language (Huang et al., 2025; Cheng et al., 2025). T-Rex separates visuomotor planning from high-rate tactile refinement (Niu et al., 2026). TTP transfers human tactile–action pretraining through shared human/robot spaces, and FTP-1 scales tactile pretraining across sensors and embodiments (Zhang et al., 2026a; Yuan et al., 2026). Being-H0.8 combines pseudo-touch and measured touch in a latent world–action model; Motus2 refines action sub-chunks and predicts future force (BeingBeyond Team, 2026; Bi et al., 2026). Our dense, hand-aligned video annotations could supply tactile supervision for such policies; downstream control remains outside our evaluation.

Privileged learning and distillation transfer training-only information to a deployable student (Vapnik & Vashist, 2009; Hinton et al., 2015), including complementary views and modalities for egocentric learning (Grauman et al., 2024; Li et al., 2021; Radevski et al., 2023). We distill masked tactile fields with confidence- and discrepancy-based gating, while segment-recurrent memory (Dai et al., 2019) compresses past clips into tokens for current tactile queries.

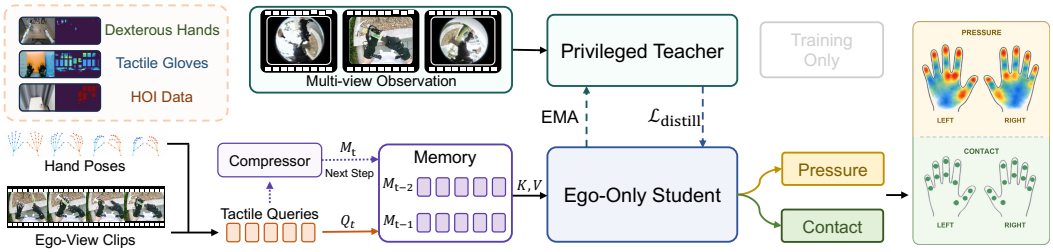

Figure 2: **Training and deployment of TouchAnnotator.** Human and robot supervision supplies tactile targets, and a multi-view EMA teacher guides the ego-only student during training. Compressed memory provides past-clip context for the student at deployment.

3 METHOD

3.1 PROBLEM FORMULATION

Index clips by t. Each contains T egocentric frames $\mathbf{X}_t^e \in \mathbb{R}^{T\times 3\times H\times W}$ and bimanual pose $\mathbf{P}_t \in \mathbb{R}^{T\times 42\times 3}$. Training samples may also provide synchronized wrist clips $\mathbf{X}_t^{w_l}, \mathbf{X}_t^{w_r}$. For each clip, the model predicts bimanual pressure and contact maps on a hand-aligned grid:

$$\hat{\mathbf{Y}}_t^p, \hat{\mathbf{Y}}_t^c \in [0,1]^{T\times 2\times G\times G}, \qquad G = 21, \tag{1}$$

where p is relative pressure and c is contact probability. A finite-cell mask $\mathbf{M}$ and sample-level availability flags $a^p, a^c \in \{0,1\}$ define task masks $\mathbf{M}^p = a^p\mathbf{M}$ and $\mathbf{M}^c = a^c\mathbf{M}$. Deployment uses

$$f_\theta : \left(\mathbf{X}^e_{t-K:t}, \mathbf{P}_{t-K:t}\right) \mapsto \left(\hat{\mathbf{Y}}_t^p, \hat{\mathbf{Y}}_t^c\right), \tag{2}$$

with pose estimated from ego RGB using WiLoR in the reported evaluations. Ego-only deployment thus estimates pose without tactile measurements or wrist images.

3.2 UNIFYING HETEROGENEOUS TACTILE SUPERVISION

Our configured training corpus spans **six datasets and three embodiment types**, totaling **27,005 video segments and 35.0M synchronized source frames** (195.3 hours; Table 1). It combines measured glove responses, geometry-derived human contact, and robot fingertip deformation. Only EgoTouch and HumanTouch supervise relative pressure; all six sources provide contact supervision.

All sources share a bilateral 21×21 anatomical layout. EgoTouch remaps each 16×16 glove array; HumanTouch projects 360 taxels onto the same 217-cell support and uses one dataset-level monotone map, fixed across clips, to match EgoTouch's active-taxel distribution. This statistical alignment does not establish common physical pressure units. Geometry-derived contact thresholds MANO vertex-object distances at 5 mm before interpolation to the grid; proximity remains a noisy contact proxy. T-Rex deformation is pooled into corresponding fingertip regions, leaving unobserved cells invalid. Hand mirroring follows a shared anatomical convention. The mixed loader uses fixed source proportions (Table C.1), avoiding repetition where possible.

3.3 HAND-LOCAL TACTILE QUERIES AND DUAL PREDICTION HEADS

Each available stream yields local tokens $\mathbf{F}^v \in \mathbb{R}^{T\times N\times D}$ through a shared Qwen3.5 video tower. A pose encoder supplies hand-specific tokens. Each hand has 64 learned tactile queries on an 8×8 latent grid, attending to ego and its corresponding wrist stream; both use ego only at deployment.

A pose-conditioned soft ROI weights visual tokens at within-clip frame τ:

$$r^{h,v}_{\tau,n} = \sigma\left(\frac{\langle W_q\bar{\mathbf{p}}^h_\tau, W_k\mathbf{F}^v_{\tau,n}\rangle}{\sqrt{D}}\right), \tag{3}$$

where $\bar{\mathbf{p}}^h_\tau$ averages hand h's pose tokens. Two cross-attention blocks fuse ROI-weighted features, pose, and tactile queries into $\mathbf{Q}_t \in \mathbb{R}^{T\times 2\times 64\times D}$. After memory augmentation (Section 3.5), a Qwen3.5 reasoner produces $\mathbf{Z}_t$ using shared forward and reversed-time passes. Independent output projections predict pressure and contact, separating their logits while retaining a shared representation.

3.4 MULTI-VIEW DISTILLATION

Multi-view distillation applies to samples with valid privileged views. A stop-gradient teacher, initialized from the student, receives clean ego and available wrist clips. We update its parameters by an exponential moving average:

$$\bar{\theta} \leftarrow m\bar{\theta} + (1-m)\theta, \qquad m = 0.996, \tag{4}$$

The student receives ego only, with stochastic downsampling and sensor noise sharing one degradation plan across current and historical clips. Single-view samples train without distillation.

For reasoned queries, a student predictor g matches normalized teacher features:

$$\mathcal{L}_q = \frac{1}{|\Omega_q|} \sum_{i \in \Omega_q} \left(2 - 2 \left\langle \frac{g(\mathbf{Z}_i^s)}{\|g(\mathbf{Z}_i^s)\|_2}, \frac{\mathbf{Z}_i^t}{\|\mathbf{Z}_i^t\|_2} \right\rangle \right). \tag{5}$$

Here Ω_q contains queries from privileged samples whose mean $1 - \cos(\mathbf{Z}^s, \mathbf{Z}^t)$, computed before g, lies in $[0.002, 0.50]$. Pressure transfer uses smooth-ℓ_1 regression under this sample gate and $\mathbf{M}^p$. Contact transfer uses Bernoulli Jensen–Shannon divergence under $\mathbf{M}^c$, an independent sample-mean JSD gate $[10^{-5}, 0.10]$, and cellwise teacher confidence of at least 0.60 in either class. These gates heuristically select moderate discrepancies; they do not guarantee teacher correctness. Empty masks yield zero loss. The combined term is

$$\mathcal{L}_{\text{distill}} = \mathcal{L}_q + 0.5\mathcal{L}_p^{\text{distill}} + 0.5\mathcal{L}_c^{\text{distill}}. \tag{6}$$

Output distillation respects task availability; query distillation has no cell-level mask.

3.5 PAST-ONLY CAUSAL CLIP MEMORY

The current queries read the $K = 2$ preceding non-overlapping clips. For each, $M = 16$ learned seeds cross-attend to queries with sinusoidal time encoding across frames, hands, and sites, producing $\mathbf{C}_{t-k} \in \mathbb{R}^{M \times D}$. Two cross-attention/MLP layers read recency-encoded caches:

$$\widetilde{\mathbf{Q}}_t = \text{MemRead}(\mathbf{Q}_t, [\mathbf{C}_{t-2} + \mathbf{e}_2; \mathbf{C}_{t-1} + \mathbf{e}_1]). \tag{7}$$

Missing history is masked. Each cache compresses its clip's unconditioned queries, excluding memory readouts and temporal-reasoner outputs, so history does not accumulate recursively. The state is bounded to two caches. Causality holds *across clips*; bidirectional reasoning requires observing the complete current eight-frame clip and is not frame-online inference.

3.6 TRAINING OBJECTIVE

In the saved Full configuration, $\mathcal{L}_{\text{press}}$ combines contact-weighted MSE and ℓ_1 regression with total variation under $\mathbf{M}^p$; $\mathcal{L}_{\text{contact}}$ is positive-weighted BCE under $\mathbf{M}^c$. The objective is

$$\mathcal{L} = \mathcal{L}_{\text{press}} + \lambda_c\mathcal{L}_{\text{contact}} + \lambda_{\text{roi}}\mathcal{L}_{\text{roi}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{distill}}\mathcal{L}_{\text{distill}}. \tag{8}$$

ROI supervision uses available glove masks. Alignment couples visual, pose, fused, and tactile embeddings; its separate tactile encoder reconstructs source signals on valid cells, including contact proxies. The distillation weight ramps as $\lambda_{\text{distill}} = 0.1 \min(1, e/3)$ at epoch $e \geq 1$. Settings appear in Appendix A.

4 EXPERIMENTS

We first evaluate on the established EgoTouch and FPHA benchmarks and compare qualitative predictions on both datasets. We then test agreement with measured touch on our collected recordings and examine whether a shared canonical field supports qualitatively coherent predictions across embodiments.

Table 2: **Tactile prediction on EgoTouch.** Local TouchAnnotator results use 184 seen and 76 unseen manifest-A episodes; all-view input adds both wrist cameras. TouchAnything rows are published references whose split equivalence to the local evaluations is unverified.

	Seen objects				Unseen objects			
Method	T.Acc.↑	C.IoU↑	V.IoU↑	MAE↓	T.Acc.↑	C.IoU↑	V.IoU↑	MAE↓
Multi-view								
TouchAnything (Zhou et al., 2026)	0.8566	0.5030	0.4575	0.0436	0.8347	0.4496	0.3852	0.0601
TouchAnnotator student (all views)	**0.8713**	**0.5139**	**0.4578**	0.0471	**0.9127**	**0.5038**	**0.4061**	0.0652
Egocentric								
TouchAnything (Zhou et al., 2026)	0.8393	0.4792	0.4311	0.0456	0.8271	0.4396	0.3743	0.0615
TouchAnnotator student (ego only)	**0.8613**	**0.5051**	**0.4478**	0.0479	**0.8687**	**0.4779**	**0.3821**	0.0638

4.1 Experimental setup

EgoTouch results use ego-estimated WiLoR pose (Potamias et al., 2025) and episode-macro metrics on 184 seen and 76 unseen manifest-A episodes. Contact IoU thresholds pressure at 0.1; the published TouchAnything scores are a reference because split equivalence is unverified. FPHA groups RGB-only baselines separately from methods given estimated hand pose. The sensor-grounded comparison uses TouchAnything on measured cells. We report overlap and absolute-error metrics together to distinguish spatial contact coverage from pressure calibration. Appendix C.1 documents the evaluation protocols and checkpoints.

4.2 Main results on EgoTouch

Table 2 juxtaposes the local Full seen/unseen evaluations with a published reference whose split equivalence is unverified; it supports no SOTA margin. On the same 76 episodes, providing all views increases Contact IoU from 0.4779 to 0.5038 and Volumetric IoU from 0.3821 to 0.4061, while MAE rises from 0.0638 to 0.0652. A similar trade-off appears against TouchAnything: TouchAnnotator achieves higher overlap but slightly higher MAE. This is consistent with sparse tactile maps: MAE over all cells favors conservative near-zero predictions, whereas our positive-pressure weighting and contact supervision encourage broader active regions. The latter improves contact coverage but can accumulate small background residuals. On seen objects, additional views instead reduce MAE from 0.0479 to 0.0471, suggesting better pressure calibration for familiar objects. Both input conditions use the same checkpoint, so this comparison isolates inference-time view access.

The all-view result is a diagnostic rather than the deployment setting. Because both rows use the same student weights and ego-estimated pose, the difference measures the value of additional image evidence without changing the learned predictor or pose source. The remaining gap indicates that wrist cameras reveal contact cues hidden from the head-mounted view, while the ego-only row represents the intended annotation protocol for videos that lack synchronized auxiliary cameras.

4.3 Transfer to bare-hand contact on FPHA

FPHA tests whether the tactile predictor transfers to contact annotations on bare hands. We evaluate the 124 sequences (10,998 frames) in its official test split with complete hand/object geometry (Garcia-Hernando et al., 2018). Every method uses the same cached labels: a MANO hand vertex is in contact when its nearest object-mesh vertex is within 10 mm. Predictions at shared 15 Hz timestamps are linearly interpolated to the 30 Hz source grid; metrics are computed over 778 right-hand vertices per timestamp and then averaged. We perform no FPHA training, threshold tuning, or label-dependent mapping fit.

Table 3 separates the RGB-only methods DECO (Tripathi et al., 2023) and HACO (Jung & Lee, 2025) from TouchAnything and TouchAnnotator, which also receive estimated 3D right-hand joints. DECO's full-body scores are restricted to the same right-hand vertex subset, HACO predicts native MANO scores, and both pressure-based models use the same fixed 217-to-778 nearest-support projection. Binary contact uses DECO probability ≥ 0.5, HACO probability > 0.52, or mapped pressure > 0.1; AUROC uses continuous scores.

Without FPHA adaptation, TouchAnnotator reaches F1 0.3658 on these bare-hand labels, close to HACO's 0.3768. Its AUROC (0.7431 versus 0.7376) and recall (0.6169 versus 0.5742) are also

Table 3: **Cross-dataset contact transfer on FPHA.** Methods share geometry-derived labels and timestamps on 124 official-test sequences. Groups distinguish ego RGB alone from ego RGB with estimated hand pose; bold marks the best result within each group.

Method	AUROC↑	Recall↑	F1↑
Ego RGB only			
DECO	0.6763	0.1550	0.0858
HACO	0.7376	0.5742	**0.3768**
Ego RGB + estimated hand pose			
TouchAnything	0.5486	0.1153	0.1209
TouchAnnotator	**0.7431**	**0.6169**	0.3658

numerically higher. Under the shared estimated-pose protocol, its F1 exceeds TouchAnything's 0.1209. This shows that a predictor trained with heterogeneous tactile supervision can remain competitive on geometry-derived bare-hand contact. Figure 3 shows representative EgoTouch pressure and FPHA contact maps. Appendix C.5 details preprocessing, timestamp handling, and mapping sensitivity.

The input groups in Table 3 should be interpreted separately: the pose-conditioned models receive information unavailable to the RGB-only baselines. Within this protocol, AUROC evaluates score ranking independently of a single operating threshold, whereas F1 reflects the fixed threshold used by each output type. TouchAnnotator's higher AUROC and recall but slightly lower F1 than HACO therefore indicate transferable contact ranking without establishing a universally better calibrated decision rule.

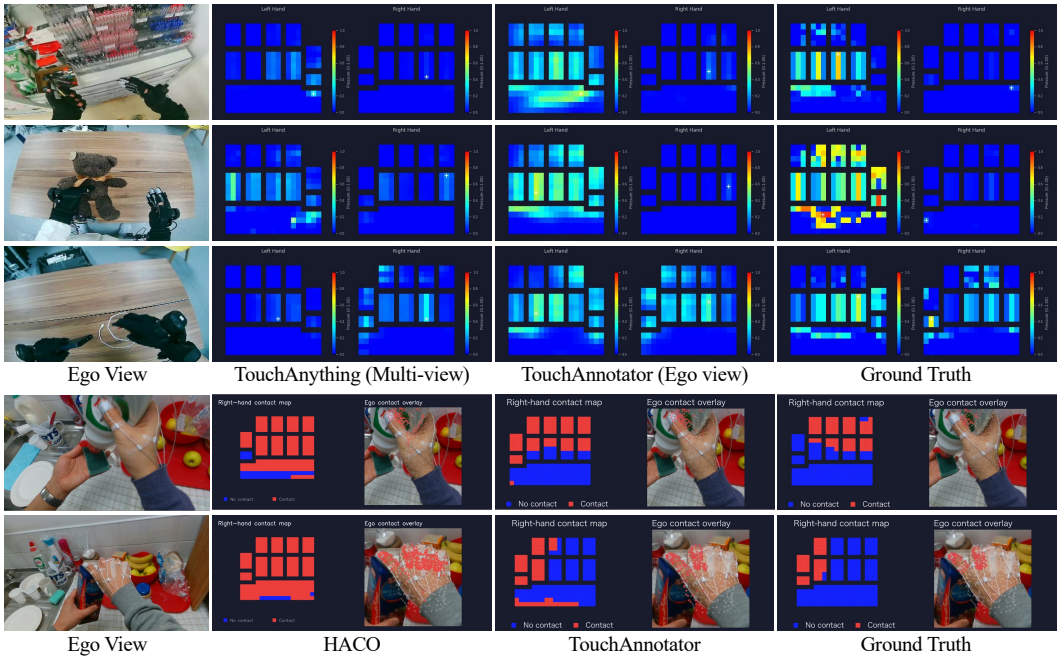

Figure 3: **Qualitative tactile predictions on EgoTouch and FPHA.** The upper examples compare EgoTouch pressure estimates from multi-view TouchAnything and ego-only TouchAnnotator with measured maps. The lower examples compare FPHA contact estimates from HACO and TouchAnnotator with geometry-derived reference labels.

4.4 Validation on collected tactile recordings

For additional validation, we collected egocentric video with synchronized tactile measurements from both hands: **897 retained sequences across 30 everyday manipulation tasks** (Figure 4). These recordings provide a measured reference for contact location and relative-intensity changes when

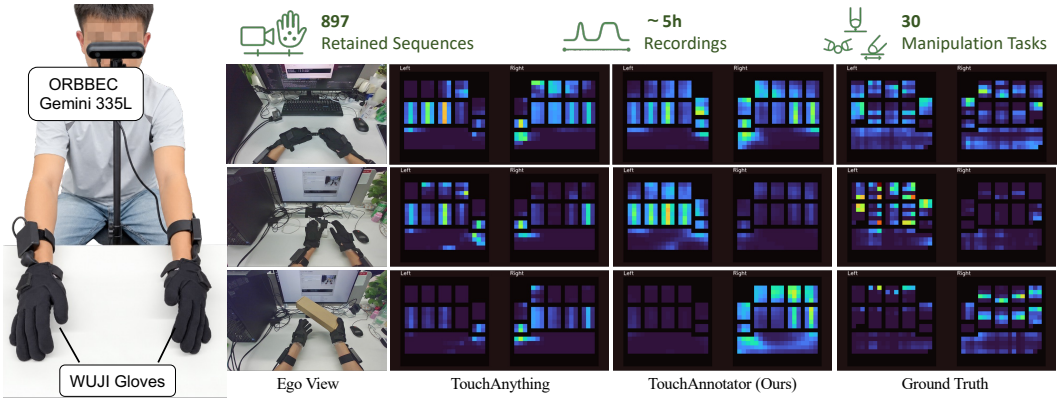

Figure 4: **Collected recordings and measured tactile examples.** An ego camera and Wuji gloves record synchronized bimanual interactions; the panels compare example predictions with measured maps. The collection contains 897 retained sequences across 30 tasks (about five hours).

Table 4: **Sensor-grounded tactile prediction by expected tactile value.** Each cell compares TouchAnything / TouchAnnotator on a cached cohort of 860 episodes from 21 tasks using measured-support 217-cell split-pressure targets. Lower to High are preliminary labels based on task descriptions; the operation-complexity breakdown appears in Table D.2.

Metric (TA / Ours)	Overall (21/860)	Lower (3/110)	Moderate (11/430)	High (7/320)
T.Acc.↑	63.29/**68.67**	65.12/**72.41**	71.59/**75.31**	48.20/**55.66**
T.F1↑	76.87/**81.08**	78.86/**83.99**	82.61/**85.47**	65.02/**71.46**
C.IoU↑	6.89/**10.73**	7.11/**10.11**	8.04/**11.76**	5.21/**9.38**
V.IoU↑	9.10/**14.12**	9.23/**13.97**	10.37/**14.99**	7.30/**12.83**
MAE↓	**0.083013**/0.087488	**0.095079**/0.106340	**0.073978**/0.081586	0.093117/**0.089496**

visible motion alone may be ambiguous. Appendix Table D.1 defines every task and its principal tactile events.

Diverse contact mechanisms. The tasks span brief presses, insertion and separation, sustained grasping, and wiping. This coverage supports evaluation of contact onset and release, changes in active hand regions, and intensity variation during continued interaction. It includes simple actions with brief tactile events as well as manipulation involving sustained or changing contact constraints.

Bimanual coordination and temporal structure. Joint support, handovers, and complementary hand roles extend evaluation to contact distribution across both hands. The synchronized streams allow predictions to be checked for coordinated contact changes and continuity during stable manipulation. These interactions motivate temporal diagnostics for brief events, delayed transitions, and spurious changes during sustained contact.

Evaluation protocol. The annotator receives ego RGB and hand pose estimated from that stream. Evaluation uses common measured hand regions and trajectory-relative sensor responses. Contact IoU assesses spatial contact, Volumetric IoU and MAE assess relative intensity, and temporal accuracy and F1 assess frame-level contact presence. Alignment, normalization, and metric definitions appear in Appendix D. Table 4 reports a cached raw 860-episode cohort; it does not cover all 897 retained sequences in the collection.

Table 4 shows that TouchAnnotator agrees more closely with measured tactile signals than TouchAnything. Overall, it improves temporal accuracy and F1 by 5.38 and 4.21 percentage points, respectively, while raising Contact IoU from 6.89 to 10.73 and Volumetric IoU from 9.10 to 14.12. The gains are consistent across tactile-value strata: TouchAnnotator improves both overlap metrics in every group and delivers the largest temporal gains in the high-value group, where temporal accuracy and F1 increase by 7.46 and 6.44 points. Although its overall MAE is slightly higher, it reduces MAE in the high-value group (0.0931 to 0.0895), indicating a clear advantage on interactions for which

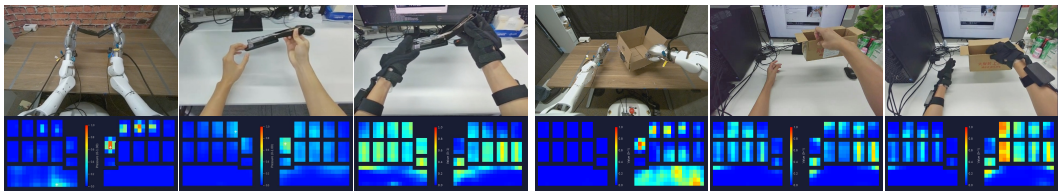

Figure 5: **Illustrative tactile patterns across embodiments.** Two examples pair camera views with estimated tactile maps from dexterous robot hands, bare hands, and tactile gloves.

tactile information is most important. Additional breakdowns and evaluation details are provided in the appendix.

This combination of overlap gains and overall MAE regression reflects the different emphasis of the metrics. Contact and Volumetric IoU focus on recovering active support and its magnitude, whereas MAE averages errors over all valid cells, including the many inactive or weakly active locations. Reporting them together prevents a sparse near-zero prediction from appearing uniformly preferable. The tactile-value strata are descriptive task groupings rather than sensor-derived classes, so they reveal where the aggregate differences occur without defining a new supervised label.

4.5 Cross-embodiment examples

Figure 5 compares matched manipulation tasks performed by dexterous robot hands, bare hands, and gloved hands. For the same operation, TouchAnnotator predicts localized tactile responses across all three embodiments despite substantial differences in hand appearance and sensor layout. These examples qualitatively show that the learned representation can recover task-relevant touch across embodiments. Because the robot and glove responses have different sensor semantics and scales, the figure demonstrates cross-embodiment prediction rather than numerical alignment.

The canonical field makes these outputs visually comparable by anatomical region, not by physical unit. Accordingly, the figure is evidence of qualitative localization across appearance and embodiment shifts; it does not claim that equal colors correspond to equal forces. Establishing such a correspondence would require paired calibration across glove taxels, bare-hand contact proxies, and robot deformation sensors.

5 Conclusion

TouchAnnotator combines masked human/robot supervision, multi-view distillation, and past-only clip memory for bimanual tactile annotation from egocentric video. EgoTouch results show improved contact overlap when the same student receives wrist views. The cached sensor-grounded cohort shows higher contact overlap than TouchAnything, and FPHA shows stronger contact transfer under shared annotated-pose input. Together, these evaluations support dense tactile annotation from ordinary ego video while distinguishing deployment results from diagnostic settings with additional inputs.

The predictions remain relative sensor responses rather than calibrated forces. Geometry-derived contact depends on hand/object reconstruction and distance thresholds, and ego-only inference remains sensitive to pose failure, blur, truncation, and prolonged occlusion. Causality is enforced across clips, but the complete current clip is processed bidirectionally, so the system is intended for offline annotation rather than frame-online feedback. Evaluating whether the resulting tactile annotations improve downstream policy or representation learning is an important next step.

REPRODUCIBILITY STATEMENT

Appendix A records configuration, conversion, and evaluation details. Current audits cannot verify the published EgoTouch split or every historical initialization.

ETHICS STATEMENT

The collected recordings contain egocentric human interaction and therefore may include privacy-sensitive visual information. Data collection should use informed consent, minimize the capture of bystanders and personally identifying content, and provide a deletion mechanism. Tactile values are normalized sensor responses unless a sensor is explicitly force-calibrated; they should not be interpreted as safety-certified force estimates.

REFERENCES

BeingBeyond Team. Being-H0.8: A latent tactile world-action model at scale. Project page, 2026.

Hongzhe Bi, Zihao Zhou, Yihang Tang, Jingrui Pang, Shuhe Huang, Haitian Liu, Runqing Wang, et al. Motus2: A self-evolving general world model for dexterous manipulation. *arXiv preprint arXiv:2608.30237*, 2026.

Kevin Black, Noah Brown, Danny Driess, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. URL `https://arxiv.org/abs/2410.24164`.

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

Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *IEEE/CVF International Conference on Computer Vision*, pp. 12417–12426, 2021.

Zhengxue Cheng, Yiqian Zhang, Anni Tang, Keyu Wang, Wenkang Zhang, Haoyu Li, Hengdi Zhang, and Li Song. OmniVTLA: Vision-tactile-language-action models with semantic-aligned tactile sensing. *arXiv preprint arXiv:2508.08706*, 2025.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Annual Meeting of the Association for Computational Linguistics*, 2019.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, et al. Scaling egocentric vision: The EPIC-KITCHENS dataset. In *European Conference on Computer Vision*, 2018.

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*, pp. 12943–12954, 2023.

Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Xu Chen, Muhammed Kocabas, Michael J. Black, and Otmar Hilliges. HOLD: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 494–504, 2024.

Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with RGB-D videos and 3d hand pose annotations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

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 *European Conference on Computer Vision*, pp. 328–345, 2022.

Kristen Grauman, Andrew Westbury, Eugene Byrne, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, et al. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Yana Hasson, G"ul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*, 2015.

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

Daniel Sungho Jung and Kyoung Mu Lee. Learning dense hand contact estimation from imbalanced data. In *Advances in Neural Information Processing Systems*, 2025.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, et al. OpenVLA: An open-source vision-language-action model. In *Conference on Robot Learning*, pp. 2679–2713, 2025. URL https://proceedings.mlr.press/v270/kim25c.html.

Ao Li, Jinpeng Liu, Yixuan Zhu, and Yansong Tang. ScoreHOI: Physically plausible reconstruction of human-object interaction via score-guided diffusion. In *IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2509.07920.

Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-Exo: Transferring visual representations from third-person to first-person videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6943–6953, 2021.

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

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. TACO: Benchmarking generalizable bimanual tool-action-object understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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*, 2022.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations*, 2016.

Dantong Niu, Zhuoyang Liu, Zekai Wang, et al. T-Rex: Tactile-reactive dexterous manipulation. *arXiv preprint arXiv:2606.17055*, 2026.

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

Gorjan Radevski, Dusan Grujicic, Matthew Blaschko, Marie-Francine Moens, and Tinne Tuytelaars. Multimodal distillation for egocentric action recognition. In *IEEE/CVF International Conference on Computer Vision*, pp. 5213–5224, 2023.

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. *arXiv preprint arXiv:2512.16842*, 2025.

Spark Lab. HumanTouch: A multimodal system for scalable human-hand tactile acquisition. Project page, 2026.

Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J. Black. DECO: Dense estimation of 3D human-scene contact in the wild. In *IEEE/CVF International Conference on Computer Vision*, pp. 8001–8013, 2023.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5–6):544–557, 2009.

Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. In *Advances in Neural Information Processing Systems*, 2022.

Yufei Ye, Abhinav Gupta, Kris Kitani, and Shubham Tulsiani. G-HOP: Generative hand-object prior for interaction reconstruction and grasp synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1911–1920, 2024.

Chengbo Yuan, Zicheng Zhang, Mingjie Zhou, Wendi Chen, Yi Wang, Zhuoyang Liu, Dantong Niu, et al. FTP-1: A generalist foundation tactile policy across tactile sensors for contact-rich manipulation. *arXiv preprint arXiv:2606.13102*, 2026.

Yuan Zeng, Yujia Shi, Tiao Tan, Xingting Li, Yaqi Qin, Zongqing Lu, Wenming Yang, Jing-Hao Xue, and Qingmin Liao. EgoTactile: Learning grasp pressure for everyday objects from egocentric video. *arXiv preprint arXiv:2606.09243*, 2026.

Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. OAKINK2: A dataset of bimanual hands-object manipulation in complex task completion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 445–456, 2024.

Chi Zhang, Penglin Cai, Ziheng Xi, Haoqi Yuan, Hao Luo, Wanpeng Zhang, Sipeng Zheng, Chaoyi Xu, and Zongqing Lu. Human-centric transferable tactile pre-training for dexterous robotic manipulation. *arXiv preprint arXiv:2607.01067*, 2026a.

Wenkang Zhang, Chengbo Yuan, Zicheng Zhang, Zhengxue Cheng, and Yang Gao. EgoTac: In-the-wild tactile prediction from egocentric vision. *arXiv preprint arXiv:2608.15060*, 2026b.

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

Jianyi Zhou, Ziteng Gao, Feiyang Hong, Zirui Liu, Guannan Zhang, Weisheng Dai, Ruichen Zhen, Chuqiao Lyu, Haotian Wu, Yinian Mao, Xushi Wang, Yuxiang Jiang, Wenbo Ding, and Shuo Yang. TouchAnything: A dataset and framework for bimanual tactile estimation from egocentric video. *arXiv preprint arXiv:2605.13083*, 2026.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183, 2023. URL `https://proceedings.mlr.press/v229/zitkovich23a.html`.