TACTOHAND: LEARNING GEOMETRY-BASED DENSE HAND-SURFACE INTERACTION FIELDS FROM VIDEO

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding how humans interact with the world requires a representation of physical interactions at the hand surface. Tactile sensing captures these interactions, but collecting tactile data at scale requires specialized hardware. We introduce **TactoHand**, a model that predicts three dense hand-surface interaction fields—contact, proximity, and 3D force—from monocular video. We derive geometry-based contact and proximity labels and motion-derived force labels from eligible training data within a curated collection spanning approximately 50 million source frames. TactoHand uses a temporal transformer to aggregate visual evidence across frames, while three learned queries decode the corresponding hand-surface fields. A differentiable physics layer further constrains force predictions using observed object motion. Across in-domain and out-of-domain benchmarks, TactoHand improves contact and proximity estimation over prior methods. In addition, physics supervision reduces position and rotation errors in motion rollouts, and the predicted contact and proximity fields improve downstream hand–object reconstruction. Together, these results show a scalable, low-cost path to learning hand-surface interaction fields without collecting tactile measurements.

1 INTRODUCTION

Hands mediate human interactions with objects, the surrounding environment, and other hands. Although the interacting entity and its geometry vary, the hand surface provides a common reference for describing where and how these interactions occur. Contact, proximity, and force provide complementary geometric and physical cues that could support interaction reconstruction, action understanding, and embodied learning. We call these vertex-wise quantities hand-surface interaction fields: contact describes where interaction occurs, proximity captures separation from the counterpart surface, and 3D force characterizes the hand's mechanical effect.

Prior work has modeled subsets of these hand-surface quantities. HACO predicts dense contact across diverse interaction types (Jung & Lee, 2025); InterField estimates hand–object distance (Fan et al., 2023), HOPE jointly predicts contact and normal pressure from video (Jeon et al., 2026), and EgoPHI estimates contact and normal forces from egocentric images given object geometry (Ilic et al., 2026). These works motivate a shared hand-surface representation that brings contact, proximity, and 3D vertex forces together. We therefore represent all three as dense, aligned fields on the canonical MANO hand surface.

Learning these fields from monocular video, however, requires supervision that covers diverse interaction types and physical quantities. We harmonize 24 datasets comprising around 50 million frames, spanning hand–object, hand–hand, and hand–scene interactions. We derive contact and proximity labels from the available geometry. For eligible hand–object sequences, we additionally derive vertex-force labels from observed object trajectories. Object motion constrains the aggregate force and torque, but does not uniquely determine how forces are distributed over the hand surface. We therefore use an SDF-aware formulation with geometric, frictional, and temporal constraints to obtain regularized vertex-force estimates.

Inferring these fields remains challenging because contact regions are often occluded and individual frames provide limited evidence about interaction dynamics. We introduce **TactoHand**, which uses

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

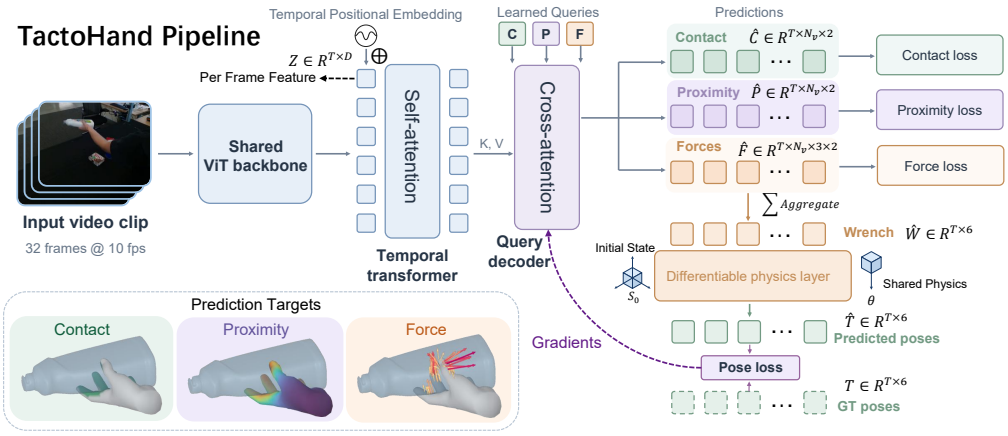

Figure 1: Overview of TactoHand. A shared video encoder extracts frame-level features, aggregated by a temporal transformer. Three learned queries decode the contact, proximity, and 3D force interaction fields at corresponding MANO vertices. Predicted vertex forces are aggregated into an object-level wrench for differentiable rigid-body supervision. Contact, proximity, force, and pose losses are applied according to the supervision available for each sample.

a temporal transformer to aggregate visual evidence across frames and three learned queries to predict contact, proximity, and 3D force fields on corresponding MANO vertices, as shown in Figure 1. To connect vertex-level force predictions with object motion, we transform the predicted vertex forces into a common coordinate frame and analytically aggregate them into object-level wrenches. For eligible rigid sequences, a differentiable dynamics layer uses these wrenches to propagate object motion and constrains force estimates through observed trajectories. This motion-level supervision complements the vertex-level targets by encouraging predicted forces to explain the resulting object motion.

We evaluate TactoHand on both interaction field prediction and downstream hand–object reconstruction. TactoHand improves contact and proximity estimation across multiple benchmarks. On HOCap, which is excluded from training, TactoHand achieves a contact Macro-F1 of 50.5, compared with 18.0 for HACO. Physics supervision further reduces position and rotation errors in reference-assisted rigid-body rollouts. In downstream hand–object reconstruction, the predicted contact and proximity fields provide geometric constraints that improve hand–object alignment and reconstruction accuracy.

Our contributions are:

- A unified representation of hand-surface interaction that jointly models dense contact, proximity, and 3D vertex forces at corresponding vertices of the MANO mesh.
- TactoHand, a model that estimates these fields from monocular video, with differentiable rigid-body supervision to constrain vertex forces.
- A heterogeneous collection of 24 hand interaction datasets, with geometry-derived contact and proximity labels and motion-derived force estimates where applicable, supporting direct field evaluation and downstream hand–object reconstruction.

2 Related Work

Hand-surface Contact and Proximity. Grasp datasets and image-based reconstruction pipelines represent hand–object contact (Brahmbhatt et al., 2020; Tse et al., 2022; Hu et al., 2024). HACO predicts dense MANO contact across interaction types, and ContactPrompt predicts contact without training (Jung & Lee, 2025; 2026). InterField instead predicts continuous hand–object distance, while contact potential fields encode vertex relationships for pose refinement (Fan et al., 2023; Yang et al., 2021). Related full-body work models human–scene contact or implicit human–object geometry (Hassan et al., 2021; Huang et al., 2022; Tripathi et al., 2023; Xie et al., 2022). TactoHand

aligns contact and proximity on MANO vertices across hand–object, hand–hand, and hand–scene data.

Visual Prediction of Pressure and Force. Visual motion has been used to estimate interaction forces (Li et al., 2022). PressureVision and EgoPressure learn from instrumented planar surfaces, whereas EgoTactile and TouchSight use glove measurements (Grady et al., 2022; Zhao et al., 2025; Zeng et al., 2026; Zhou et al., 2026). HOPE and EgoTac combine measured pressure or force with geometric contact on MANO vertices (Jeon et al., 2026; Zhang et al., 2026). EgoPHI predicts contact and normal-direction force given an egocentric image and object geometry (Ilic et al., 2026). TactoHand derives three-component vertex-force targets from object motion for sequences with suitable object trajectories.

Motion and Physical Consistency. Prior work matches simulated contact forces to observed motion or couples them with hand–object reconstruction (Ehsani et al., 2020; Hu et al., 2022); ViTaM-D adds distributed tactile sensing (Yu et al., 2025). Differentiable rigid-body models allow motion errors to supervise predicted forces (Qiao et al., 2020; Werling et al., 2021). Since motion constrains only the aggregate wrench, TactoHand regularizes its vertex-force targets and applies a reference-assisted motion loss to eligible rigid sequences.

Interaction Cues for Downstream Tasks. Contact and distance constraints aid hand–object reconstruction (Hasson et al., 2019; Grady et al., 2021; Hasson et al., 2021; Qi et al., 2024). Hand affordances and measured glove forces have also supported robot learning and physical action understanding (Srirama et al., 2024; Dessalene et al., 2026). We test downstream utility by adding predicted contact and proximity as fixed constraints to existing reconstruction pipelines.

3 Problem Formulation

Let $I_{1:T}$ denote a monocular RGB clip of T frames sampled at 10 Hz. Each hand is represented by the fixed MANO topology with $V = 778$ vertices (Romero et al., 2017). We index each vertex by $i = (h, v)$, where $h \in \{L, R\}$ denotes the hand side and $v \in \{1, \dots, V\}$ denotes the vertex index. TactoHand estimates three aligned hand-surface interaction fields on these vertices:

$$\Phi_\theta(I_{1:T}) = \left\{(\widehat{c}_{t,i}, \widehat{r}_{t,i}, \widehat{\boldsymbol{\ell}}_{t,i})\right\}_{t,i}, \quad \widehat{c}_{t,i}, \widehat{r}_{t,i} \in [0,1], \quad \widehat{\boldsymbol{\ell}}_{t,i} \in \mathbb{R}^3. \tag{1}$$

Here, $\widehat{c}_{t,i}$ denotes the contact probability, while $\widehat{r}_{t,i}$ measures proximity to the counterpart surface, with larger values indicating greater closeness. $\widehat{\boldsymbol{\ell}}_{t,i}$ contains raw vertex-local force components predicted from RGB. Converting the gated components into camera-frame forces $\widehat{\mathbf{f}}_{t,i}$ additionally requires ground-truth posed MANO geometry (Equation 7). Force supervision is applied only to eligible hand–object samples. For hand–hand and hand–scene samples, force losses are masked out, while contact and proximity supervision is retained wherever valid annotations are available.

4 Dataset Construction

4.1 Data Sources

We combine 24 sources spanning diverse hand interaction scenarios, comprising 282,874 curated sequences and 49.02 million captured frames across camera views, approximately $5.2\times$ HACO's frame count. For all datasets, we derive geometry-based contact and proximity labels on a shared set of MANO vertices. We additionally derive motion-based vertex-force labels for 15 hand–object datasets, including 13 with rigid objects and two with articulated objects. These are collection counts; training excludes GigaHands force targets and all HOCap data. Appendix Table 4 reports the sequence and frame counts, together with the available supervision, for each source.

4.2 Contact and Proximity Labeling

For a hand vertex $\mathbf{p}$, we compute its unsigned distance u to the original object surface in the object's coordinate frame. We then obtain the signed distance s using a multiray inside–outside test, described in Appendix A.2, retaining the distance magnitude and assigning a negative sign to vertices

inside the object and a nonnegative sign otherwise. For hand–hand and hand–scene data, we query the other posed hand or the available scene surface, respectively. When the counterpart surface is open and a reliable inside–outside classification is unavailable, we use the unsigned nearest-surface distance and disable inside-mesh inclusion. The proximity target is defined as:

$$r = \exp(-\tau \min\{\max(s, 0), d_{\max}\}) . \tag{2}$$

This target equals one on and inside the surface, decreases with separation outside the surface, and remains constant beyond $d_{\max}$. We use $\tau = 20\,\mathrm{m}^{-1}$ and $d_{\max} = 0.25\,\mathrm{m}$. We assign contact labels using the following rule.

Contact Decision. Let δ_D be the dataset-specific distance threshold, $\mathbf{q}$ the closest point on the counterpart surface, and $\mathbf{n}$ the outward hand-surface normal. For an outside vertex with $u > 0$, let $\kappa = \mathbf{n}^\top(\mathbf{q} - \mathbf{p})/u$ measure normal alignment. For a closed object mesh, we assign contact by

$$c(\mathbf{p}) = \begin{cases} 1, & \mathbf{p} \text{ is inside the object,} \\ 1, & \text{outside, } u \le \delta_D,\ u < 3\,\mathrm{mm}, \\ 1, & \text{outside, } 3\,\mathrm{mm} \le u \le \delta_D,\ \kappa \ge 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

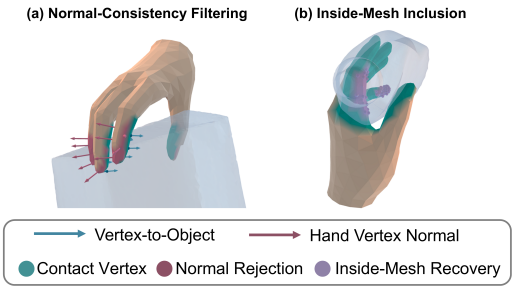

Figure 2: Contact-label construction: normal filtering (left) and inside-mesh inclusion (right).

Here, "outside" includes on-surface vertices ($u = 0$), which are labeled as contact without evaluating κ; the 3 mm band likewise bypasses the normal test where the closest-point direction is unstable. Inside-mesh inclusion retains potential contacts displaced into the object by pose error. For hand–hand and hand–scene data, we apply the same distance and normal tests, using inside-mesh inclusion only when a reliable sign is available. Appendix A.2 provides the distance thresholds and a visual comparison of the labeling rules. HACO assigns contacts by thresholding unsigned nearest-surface distance (Jung & Lee, 2025). Our geometric corrections provide more faithful contact supervision, as illustrated in Appendix Figure 6.

4.3 Motion-Derived Force Labels

The observed object motion determines the net wrench acting on the object, but does not uniquely determine the distribution of forces over the hand surface. We therefore construct force supervision in two stages: we first recover an object-level target wrench from motion and then distribute it over plausible hand vertices under geometric, frictional, and temporal constraints.

For rigid objects, we smooth the observed pose trajectories at 30 Hz and estimate linear and angular accelerations. Newton–Euler inverse dynamics then gives the non-gravitational external wrench that must be explained by the hands and any support from the environment:

$$\mathbf{w}_t^\star = \begin{bmatrix} m(\mathbf{a}_t - \mathbf{g}) \\ \mathbf{R}_t(\mathbf{I}\boldsymbol{\alpha}_t + \boldsymbol{\omega}_t \times \mathbf{I}\boldsymbol{\omega}_t) \end{bmatrix}, \qquad \mathbf{w}_t^{\mathrm{hand}} + \mathbf{w}_t^{\mathrm{env}} = \mathbf{w}_t^\star. \tag{4}$$

Here, $\mathbf{R}_t$ is the object orientation, $m = 1\,\mathrm{kg}$ is the reference object mass and $\mathbf{I}$ is the body-frame inertia scaled to this mass. The linear acceleration $\mathbf{a}_t$ and gravity $\mathbf{g}$ are expressed in the solver world frame with $\|\mathbf{g}\| = 9.81\,\mathrm{m/s}^2$, while the angular velocity $\boldsymbol{\omega}_t$ and acceleration $\boldsymbol{\alpha}_t$ are expressed in the body frame. We use dataset-provided gravity directions when available; otherwise, we estimate them from UprightNet RGB predictions (Xian et al., 2019) or scene geometry.

The target wrench alone does not uniquely determine the force at each hand vertex. We resolve this ambiguity using the local hand–object geometry illustrated in Figure 3. Vertices with signed distance $s_{t,i} < 20\,\mathrm{mm}$ are treated as force candidates, and each candidate force is constrained by a local friction cone defined by the hand-surface normal and tangent directions. We further bias the allocation toward vertices that are closer to, or penetrating, the object through

$$q_{t,i} = \frac{2}{1 + \exp(s_{t,i}/\ell)}, \qquad \ell = 4\,\mathrm{mm}. \tag{5}$$

Because the force penalty is weighted by $(q_{t,i}^2 + \epsilon)^{-1}$, vertices closer to or penetrating the object incur a lower cost.

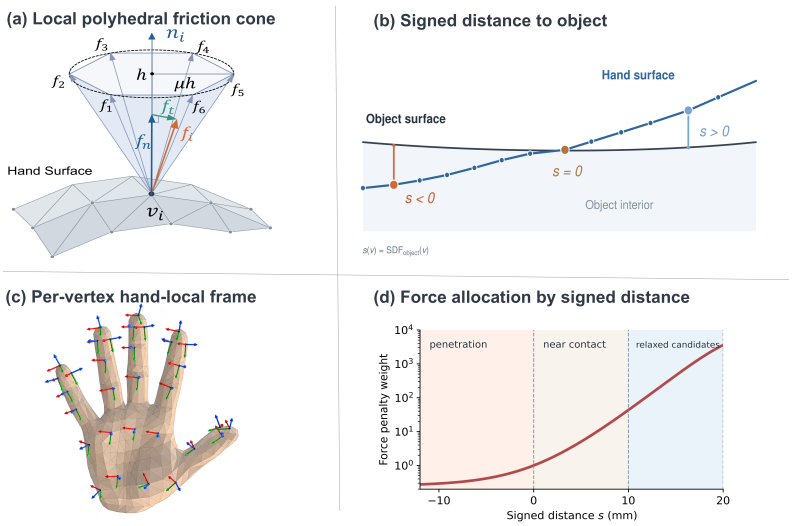

Figure 3: Force allocation geometry and weighting. Upper left: a local polyhedral friction cone constrains each active vertex force. Upper right: signed distance s distinguishes penetrated ($s < 0$), contacting ($s = 0$), and separated ($s > 0$) vertices. Lower left: vertex-local force directions on the hand surface. Lower right: the penalty weight $(q(s)^2 + \epsilon)^{-1}$ rises toward the relaxed 20 mm candidate cut-off.

We then solve for the hand forces jointly across the sequence. To account for interactions with supporting surfaces, we additionally introduce an upward environment force with magnitude $e_t \geq 0$, wrench map $\mathbf{G}_t^{\text{env}}$, and regularizer $\mathcal{R}_{\text{env}}(e_{1:T})$. The resulting sequence-level convex program is

$$\begin{aligned} \min_{\{\mathbf{f}_{t,i}\},\{e_t\}} \quad & \lambda_f \sum_{t,i} \frac{\|\mathbf{f}_{t,i}\|_2^2}{q_{t,i}^2 + \epsilon} + \lambda_{\Delta f} \sum_{t,i} \|\mathbf{f}_{t+1,i} - \mathbf{f}_{t,i}\|_2^2 + \mathcal{R}_{\text{env}}(e_{1:T}) \\ \text{s.t.} \quad & \sum_i \mathbf{G}_{t,i}\mathbf{f}_{t,i} + \mathbf{G}_t^{\text{env}} e_t = \mathbf{w}_t^\star, \\ & \mathbf{f}_{t,i} \in \mathcal{K}(\mathbf{n}_{t,i}, \mu), \quad \mathbf{f}_{t,i} = \mathbf{0} \text{ for inactive } i, \quad e_t \geq 0, \end{aligned} \tag{6}$$

where $\mathbf{G}_{t,i}$ maps a vertex force to its force-and-torque contribution about the object center of mass, and $\mathcal{K}$ is a six-facet polyhedral approximation of the unilateral Coulomb friction cone. The equality constraint ensures that the allocated hand and environment forces reproduce the motion-derived target wrench, while the magnitude and temporal terms select a spatially and temporally regularized solution among the feasible allocations.

For articulated objects, we also account for joint forces and torques, mapping each hand contact to the object part it acts on (Appendix A.3.3). Articulated sequences provide vertex-force supervision but are excluded from the rigid rollout loss. Appendix A.3 provides further details on trajectory processing, rigid-force allocation and validation, and the articulated extension.

5 TACTOHAND

5.1 VIDEO ENCODING AND INTERACTION-FIELD DECODING

Each frame is encoded by a pretrained ViT-L DINOv3 backbone (Siméoni et al., 2025). A temporal transformer aggregates clip-level visual information, and three task-specific queries decode the hand-surface interaction fields for both hands at each 10 Hz image timestamp. Architecture and input settings are provided in Appendix C.1.

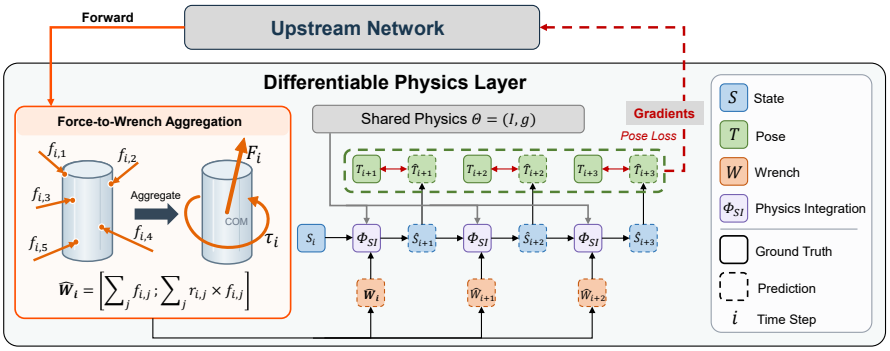

Figure 4: Differentiable rigid-body supervision. Indices $i, i+1, i+2$ are 10 Hz image timestamps; each transition contains three 30 Hz force intervals. Predicted vertex forces determine the object wrench at the image-aligned interval, while the two intermediate intervals use detached motion-derived reference forces. Pose errors supervise the propagated state.

5.2 Dense Contact and Proximity

Let $\widehat{a}_{t,h}$ denote the probability that hand h makes any contact in frame t (engagement), and $\widehat{b}_{t,h,v}$ the vertex contact probability conditioned on that event. We factor the final contact probability as $\widehat{c}_{t,h,v} = \widehat{a}_{t,h}\widehat{b}_{t,h,v}$, with engagement target $a^{\star}_{t,h} = \mathbb{1}[\sum_v c^{\star}_{t,h,v} > 0]$ on valid hand frames. Binary cross-entropy supervises engagement, while HACO's vertex-level class-balanced (VCB) loss (Jung & Lee, 2025) supervises conditional contact on engaged frames. A separate proximity head is supervised with masked mean squared error against the targets defined in Equation 2.

The same engagement scalar $\widehat{a}_{t,h}$ also gates all vertex forces of hand h. This soft gate allows force and motion losses to propagate into both the engagement and force branches while encouraging zero force on no-contact hand frames.

5.3 Vertex Forces and Analytic Wrench Aggregation

Omitting the time index, let $\widehat{\ell}_{i,1}, \widehat{\ell}_{i,2}, \widehat{\ell}_{i,n}$ denote raw local force components along two tangents and the outward normal. Ground-truth posed MANO geometry defines this basis in camera coordinates. With the engagement gate for hand $h(i)$, the hand-on-object force is

$$\widehat{\mathbf{f}}_i = \widehat{a}_{h(i)} \left(\widehat{\ell}_{i,1}\mathbf{t}_{i,1} + \widehat{\ell}_{i,2}\mathbf{t}_{i,2} + \widehat{\ell}_{i,n}\mathbf{n}_i \right). \tag{7}$$

The basis vectors and $\widehat{\mathbf{f}}_i$ are therefore expressed in the camera frame. Predicting force components relative to the hand surface reduces sensitivity to camera orientation (Figure 3, lower left; Appendix C.2).

At each image timestamp, the point-force loss $\mathcal{L}_{\text{point}}$ supervises the reconstructed camera-frame forces while balancing active and near-zero targets (Appendix C.2). Given the object center of mass $\mathbf{x}$ and vertex positions $\mathbf{p}_i$ in the same frame, the corresponding hand wrench is obtained analytically as

$$\mathbf{w}^{\text{hand}} = \begin{bmatrix} \sum_i \widehat{\mathbf{f}}_i \\ \sum_i (\mathbf{p}_i - \mathbf{x}) \times \widehat{\mathbf{f}}_i \end{bmatrix}. \tag{8}$$

5.4 Motion-Consistent Rollout

We further constrain the predicted forces through short-horizon rigid-body motion. Each 10 Hz visual interval contains three 30 Hz force intervals indexed by $s \in \{0, 1, 2\}$. At $s = 0$, we use the image-aligned network prediction $\widehat{\mathbf{f}}_{t,i}$; at $s \in \{1, 2\}$, we use the detached motion-derived reference $\mathrm{stopgrad}(\mathbf{f}^{\star}_{t,s,i})$. This keeps the image-aligned prediction differentiable while avoiding unconstrained force extrapolation between visual observations.

Starting from the observed object state, we aggregate the vertex forces into a hand wrench and transform it from camera coordinates into the solver world frame. We then combine it with gravity and the estimated environment-support terms and integrate the resulting rigid-body dynamics. Because this integration is differentiable, motion-level errors propagate back through wrench aggregation to the predicted vertex forces.

Let k index object states at the boundaries of successive 30 Hz force intervals, with $k = 0$ denoting the initial state. Let $\mathcal{H}$ denote the indices included in the short-horizon objective. We compare the propagated and observed object position and orientation:

$$\mathcal{L}_{\text{roll}} = \frac{1}{|\mathcal{H}|} \sum_{k \in \mathcal{H}} \left(\lambda_x \|\widehat{\mathbf{x}}_k - \mathbf{x}_k^\star\|_2^2 + \frac{\lambda_R}{2} \|\widehat{\mathbf{R}}_k - \mathbf{R}_k^\star\|_F^2 \right). \tag{9}$$

The differentiable state update is detailed in Appendix C.2 (Equation 17). The rollout loss is applied only to eligible rigid sequences for which the aggregated hand force can be attributed to the target object; articulated and otherwise ineligible sequences retain their available field-level supervision.

5.5 Training Objective

Each training sample activates only losses supported by its annotations. We combine the weighted engagement and conditional vertex losses into $\mathcal{L}_{\text{contact}}$ and define the overall training objective as

$$\mathcal{L} = \lambda_c \mathcal{L}_{\text{contact}} + \lambda_d \mathcal{L}_{\text{prox}} + \lambda_p \mathcal{L}_{\text{point}} + \lambda_r \mathcal{L}_{\text{roll}}. \tag{10}$$

Geometry-only sequences use contact and proximity supervision. Sequences with valid force labels additionally use $\mathcal{L}_{\text{point}}$, while eligible rigid sequences also activate $\mathcal{L}_{\text{roll}}$. Missing force supervision is masked. Repeated frames used to pad clips are excluded from force and rollout losses (Appendix C.2).

6 Experiments

6.1 Dense Proximity and Contact Comparison

Table 1 compares TactoHand with prior proximity and contact predictors. We re-score released baseline checkpoints against the same annotations. For proximity, we compare against InterField-SF and InterField-LSTM (Fan et al., 2023). For contact, we compare against HACO (Jung & Lee, 2025), EgoPHI (Ilic et al., 2026), and ContactPrompt (Jung & Lee, 2026). EgoPHI uses single RGB frames together with solver-derived MANO vertices and target-object masks and meshes. ContactPrompt is evaluated quantitatively only on MOW because full evaluation requires a separate VLM prompt for each frame; elsewhere, we include it qualitatively in Figure 5. Input settings, dataset splits, and the frames and vertices included in each metric are detailed in Appendix D.

We evaluate cross-dataset transfer on HOCap, which is excluded from training.

For proximity, InterField-SF predicts distances per frame, whereas InterField-LSTM aggregates temporal features; both are trained on ARCTIC (Fan et al., 2023). InterField covers 100 mm and TactoHand is supervised up to 250 mm, so we report full-range ADE and ADE-100, which clips both predictions and targets at 100 mm.

TactoHand lowers full-range ADE across all four benchmarks: the temporal variant performs best on ARCTIC, DexYCB, and HOCap, and the single-frame variant on MOW. The latter also leads on HOCap ADE-100. ADE-100 separates accuracy within the shared distance range from the methods' different full ranges.

Turning to contact, the temporal model consistently improves Macro-F1 over HACO on DexYCB, HOCap, and ARCTIC, with the largest gain on HOCap, from 18.0 to 50.5. On MOW, TactoHand matches HACO in Macro-F1 (34.5) while improving Micro-F1 from 41.5 to 44.3. Figure 5 illustrates representative missed and spurious contact regions across methods.

Table 1: Proximity (left) and contact (right). ADE is in mm, Accel. in m/s^2, and P/R/F1 in percent. ADE-100 clips predictions and targets at 100 mm; Macro-F1 averages F1 over frame–hand pairs with ground-truth or predicted contact. Bold marks the best per dataset. ‡ HOCap is excluded from training. See Appendix D for metrics.

	Proximity				Contact				
Dataset	Method	ADE ↓	ADE-100 ↓	Accel. ↓	Method	P ↑	R ↑	Macro-F1 ↑	Micro-F1 ↑
HOCap‡	InterField-SF	132.51	38.15	1.06	HACO	11.1	**77.7**	18.0	19.5
	InterField-LSTM	134.69	40.10	**0.47**	EgoPHI	10.3	37.3	11.4	16.2
	Ours (SF)	46.09	**6.35**	1.35	Ours (SF)	**68.9**	57.7	50.1	62.8
	Ours (T)	**42.51**	6.40	1.07	Ours (T)	68.8	60.6	**50.5**	**64.5**
ARCTIC (s05, ego)	InterField-SF	29.65	9.81	0.82	HACO	42.2	**82.2**	42.2	55.8
	InterField-LSTM	29.68	9.77	**0.76**	EgoPHI	33.5	45.2	31.8	38.5
	Ours (SF)	20.02	10.58	1.09	Ours (SF)	59.2	67.3	50.9	63.0
	Ours (T)	**16.23**	**9.32**	1.10	Ours (T)	**64.0**	63.7	**51.0**	**63.8**
DexYCB	InterField-SF	128.99	32.83	1.76	HACO	27.4	**85.5**	34.3	41.5
	InterField-LSTM	127.43	31.05	**1.18**	EgoPHI	4.4	4.2	2.7	4.3
	Ours (SF)	26.14	3.92	1.74	Ours (SF)	63.6	71.4	58.5	67.3
	Ours (T)	**23.29**	**3.42**	1.46	Ours (T)	**65.7**	69.7	**59.5**	**67.6**
MOW	InterField-SF	108.13	33.15	–	HACO	29.2	**71.5**	**34.5**	41.5
	InterField-LSTM	110.08	35.00	–	ContactPrompt	25.2	66.2	29.7	36.5
	Ours (SF)	**46.86**	**12.21**	–	EgoPHI	22.5	39.4	12.4	28.7
					Ours (T)	**44.2**	44.3	**34.5**	**44.3**

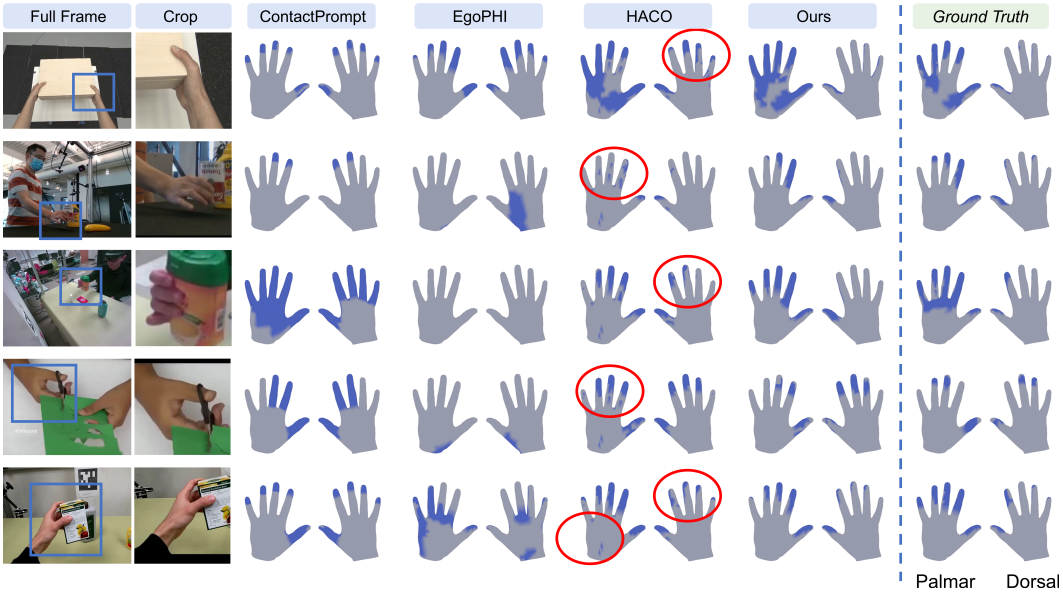

Figure 5: Contact comparison showing the input scene, hand crop, predicted contact maps, and geometry-derived reference.

6.2 Force and Rollout Ablation

We next evaluate the effect of motion-level physics supervision. We train single-frame and temporal variants with and without the rollout loss while holding all other training settings fixed. Table 2 reports both vertex-force errors and a reference-assisted rigid-body rollout diagnostic. Appendix D.3 details the evaluation protocol and window-exclusion criteria.

Table 2: Force and rollout ablation on DexYCB (DINOv3 ViT-L). SF/T: single-frame/temporal; MAE: mN; rollout ADE: cm; angles: degrees. Active targets exceed $0.05\,\mathrm{N}$. The $0.4\,\mathrm{s}$ rollout uses reference-assisted intervals and excludes divergent windows (Appendix D.3); bold indicates column minima.

Model	Physics	Vertex MAE ↓	Active MAE ↓	Force angle ↓	Rollout ADE ↓	Rollout angle ↓
SF	Off	**12.6**	124.9	28.2	5.3	45.5
SF	On	12.9	126.7	29.8	3.8	**27.4**
T	Off	13.7	126.4	31.0	5.7	45.2
T	On	12.8	**122.7**	**27.5**	**3.3**	29.4

Physics supervision primarily improves motion consistency: for the temporal model, rollout ADE falls from 5.7 to 3.3 cm and active-force MAE from 126.4 to 122.7 mN. The single-frame model also lowers rollout error, despite slightly worse active-force MAE and force-angle error.

6.3 Downstream Hand–Object Reconstruction Refinement

We replace the original contact objective in EasyHOI (Liu et al., 2025) and ArtHOI (Wang et al., 2026) with terms derived from fixed TactoHand contact and proximity predictions, keeping all other settings unchanged. This controlled setting uses metric object geometry; EasyHOI also uses ground-truth object orientation (Appendix D).

On DexYCB, HOCap, and MOW, TactoHand improves PA-VPE in all six method–dataset pairs and gives the lowest hand–object Chamfer distance in each (Table 3). Full metrics and visuals appear in Appendix Table 5 and Figures 7–8.

Table 3: Controlled reconstruction refinement. PA-VPE: mm; $\mathrm{CD}_{\mathrm{ho}}$: bidirectional mean squared hand–object Chamfer in $10^{-3}\,\mathrm{m}^2$. Bold marks each method–dataset best.

		PA-VPE ↓			$\mathrm{CD}_{\mathrm{ho}}$ ↓		
Dataset	Method	Base	+ HACO	+ Ours	Base	+ HACO	+ Ours
DexYCB	EasyHOI	9.937	9.516	**9.077**	24.140	12.552	**2.004**
	ArtHOI	7.317	6.027	**4.626**	11.587	13.627	**1.431**
HOCap	EasyHOI	9.179	9.051	**7.592**	59.416	68.943	**11.461**
	ArtHOI	6.984	6.684	**6.073**	5.234	5.199	**1.104**
MOW	EasyHOI	14.832	14.276	**13.939**	69.080	72.638	**15.266**
	ArtHOI	13.572	13.261	**13.100**	73.993	73.147	**32.413**

7 Limitations

Geometry-derived contact is a proxy sensitive to source geometry and alignment. Motion-derived vertex forces are non-unique estimates, not measurements, and depend on inertia, friction, and environmental support assumptions. Motion-level evaluation is limited to short-horizon, reference-assisted rigid-body rollouts. Future work should incorporate measured forces and articulated rollout supervision while reducing reliance on auxiliary physical inputs.

8 Conclusion

TactoHand predicts contact, proximity, and 3D force as dense interaction fields on the hand surface. Geometry-derived labels across 24 sources, supplemented with regularized motion-derived force targets where available, support heterogeneous training. In our evaluations, TactoHand improves contact and proximity estimation, and its contact and proximity fields improve controlled hand–object reconstruction. Physics supervision reduces trajectory error in the reference-assisted rigid-body evaluation. Together, these results show how a shared hand-surface representation connects visual prediction, object motion, and hand–object reconstruction.

Reproducibility Statement

The main text defines the hand-surface fields, label construction, model, and evaluation. Appendices A–D document data sources and labels, additional results, model and training settings, and evaluation protocols. We will release the code and derived annotations to support reproduction of the labels and trained model.

AI Use Statement

Generative AI tools assisted with retrieval and discovery of related work, drafting sections of the paper, and polishing its writing. The authors take responsibility for the final content, including its claims, citations, figures, and results.

Ethics Statement

This work uses only previously released datasets and collects no new human-subject data. The sources were assembled by their original authors under their own consent and licensing arrangements, and we neither re-identify participants nor add identifying information. We record each dataset's source and preprocessing steps so that derived annotations can be traced to the original data. Users must comply with the terms of each source dataset.

The labels we derive are estimates, not measurements. Vertex forces are obtained from smoothed object motion through inverse dynamics and a constrained allocation, and the construction admits many local force fields for a single aggregate wrench; the dataset records which frames satisfy the force-allocation constraints and never substitutes a zero-force target for an infeasible one. Monocular force estimation is also dual-use, in that an accurate estimate of applied force could inform physical manipulation or contact-based control. We therefore report force labels with their validity flags and residuals, and we caution against treating them, or any force prediction trained on them, as ground truth in safety-critical settings.

References

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. In *CVPR*, pp. 7061–7071, 2025. 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, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. ContactPose: A dataset of grasps with object contact and hand pose. In *ECCV*, pp. 361–378. Springer, 2020. doi: 10.1007/978-3-030-58601-0_22.

Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *ICCV*, pp. 12417–12426, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Cao_Reconstructing_Hand-Object_Interactions_in_the_Wild_ICCV_2021_paper.html.

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *CVPR*, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chao_DexYCB_A_Benchmark_for_Capturing_Hand_Grasping_of_Objects_CVPR_2021_paper.html.

Woojin Cho, Jihyun Lee, Minjae Yi, Minje Kim, Taeyun Woo, Donghwan Kim, Taewook Ha, Hyokeun Lee, Je-Hwan Ryu, Woontack Woo, and Tae-Kyun Kim. Dense hand-object (HO) graspnet with full grasping taxonomy and dynamics. In *ECCV*, volume 15140 of *Lecture*

Notes in Computer Science, pp. 284–303. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-73007-8_17. URL https://hograspnet2024.github.io/.

Eadom Dessalene, Botao He, Michael Maynord, Yonatan Tussa, Pavan Mantripragada, Yianni Karabatis, Nirupam Roy, and Yiannis Aloimonos. FEEL (force-enhanced egocentric learning): A dataset for physical action understanding. In *ECCV*, 2026. URL https://arxiv.org/abs/2603.15847. Preprint: arXiv:2603.15847.

Kiana Ehsani, Shubham Tulsiani, Saurabh Gupta, Ali Farhadi, and Abhinav Gupta. Use the force, Luke! learning to predict physical forces by simulating effects. In *CVPR*, 2020. doi: 10.1109/CVPR42600.2020.00030. URL https://arxiv.org/abs/2003.12045.

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 *CVPR*, pp. 12943–12954, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Fan_ARCTIC_A_Dataset_for_Dexterous_Bimanual_Hand-Object_Manipulation_CVPR_2023_paper.html.

Rao Fu, Dingxi Zhang, Alex Jiang, Wanjia Fu, Austin Funk, Daniel Ritchie, and Srinath Sridhar. GigaHands: A massive annotated dataset of bimanual hand activities. In *CVPR*, pp. 17461–17474, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_GigaHands_A_Massive_Annotated_Dataset_of_Bimanual_Hand_Activities_CVPR_2025_paper.html.

Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. ContactOpt: Optimizing contact to improve grasps. In *CVPR*, 2021. doi: 10.1109/CVPR46437.2021.00152. URL https://arxiv.org/abs/2104.07267.

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 *ECCV*, volume 13666 of *Lecture Notes in Computer Science*, pp. 328–345. Springer, Cham, 2022. doi: 10.1007/978-3-031-20068-7_19. URL https://arxiv.org/abs/2203.10385.

Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A method for 3D annotation of hand and object poses. In *CVPR*, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Hampali_HOnnotate_A_Method_for_3D_Annotation_of_Hand_and_Object_CVPR_2020_paper.html.

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 *CVPR*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/papers/Hampali_Keypoint_Transformer_Solving_Joint_Identification_in_Challenging_Hands_and_Object_CVPR_2022_paper.pdf.

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *ICCV*, pp. 2282–2292, 2019. doi: 10.1109/ICCV.2019.00237. URL https://prox.is.tue.mpg.de/.

Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3D scenes by learning human-scene interaction. In *CVPR*, 2021. doi: 10.1109/CVPR46437.2021.01447. URL https://arxiv.org/abs/2012.11581.

Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. URL https://hassony2.github.io/obman.

Yana Hasson, Gül Varol, Cordelia Schmid, and Ivan Laptev. Towards unconstrained joint hand-object reconstruction from RGB videos. In *3DV*, pp. 659–668, 2021. doi: 10.1109/3DV53792.2021.00075. URL https://arxiv.org/abs/2108.07044.

Haoyu Hu, Xinyu Yi, Hao Zhang, Jun-Hai Yong, and Feng Xu. Physical interaction: Reconstructing hand-object interactions with physics. In *SIGGRAPH Asia*, 2022. doi: 10.1145/3550469.3555421. URL https://arxiv.org/abs/2209.10833.

Junxing Hu, Hongwen Zhang, Zerui Chen, Mengcheng Li, Yunlong Wang, Yebin Liu, and Zhenan Sun. Learning explicit contact for implicit reconstruction of hand-held objects from monocular images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(3):2220–2228, 2024. doi: 10.1609/aaai.v38i3.27995. URL https://ojs.aaai.org/index.php/AAAI/article/view/27995.

Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *CVPR*, 2022. doi: 10.1109/CVPR52688.2022.01292. URL https://arxiv.org/abs/2206.09553.

Andela Ilic, Rachel Schuchert, Yijing Jiang, and Christian Holz. EgoPHI: Estimating 3D hand-object contact and force from egocentric vision. *arXiv preprint*, 2026. doi: 10.48550/arxiv.2608.13014. URL https://arxiv.org/abs/2608.13014.

Subin Jeon, Byungjun Kim, and Hanbyul Joo. HOPE: Hand-object pressure estimation from monocular videos. *arXiv preprint*, 2026. URL https://arxiv.org/abs/2608.06192.

Daniel Sungho Jung and Kyoung Mu Lee. Learning dense hand contact estimation from imbalanced data. In *NeurIPS*, volume 38, 2025. doi: 10.52202/085713-4017. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/ae484111f7a8f02aac14f25d5e7c498a-Abstract-Conference.html.

Daniel Sungho Jung and Kyoung Mu Lee. Training-free dense hand contact estimation with multimodal large language models. *arXiv preprint arXiv:2605.05886*, 2026.

Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2O: Two hands manipulating objects for first person interaction recognition. In *ICCV*, 2021. doi: 10.1109/ICCV48922.2021.00998. URL https://taeinkwon.com/projects/h2o/.

Zongmian Li, Jiri Sedlar, Justin Carpentier, Ivan Laptev, Nicolas Mansard, and Josef Sivic. Estimating 3D motion and forces of human-object interactions from internet videos. *International Journal of Computer Vision*, 130(2):363–383, 2022. doi: 10.1007/s11263-021-01540-1. URL https://arxiv.org/abs/2111.01591.

Yumeng Liu, Xiaoxiao Long, Zemin Yang, Yuan Liu, Marc Habermann, Christian Theobalt, Yuexin Ma, and Wenping Wang. EasyHOI: Unleashing the power of large models for reconstructing hand-object interactions in the wild. In *CVPR*, pp. 7037–7047, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liu_EasyHOI_Unleashing_the_Power_of_Large_Models_for_Reconstructing_Hand-Object_CVPR_2025_paper.html.

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 *CVPR*, 2024. URL https://taco2024.github.io/.

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 *CVPR*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Liu_HOI4D_A_4D_Egocentric_Dataset_for_Category-Level_Human-Object_Interaction_CVPR_2022_paper.html.

Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In *ECCV*, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/3540_ECCV_2020_paper.php.

Gyeongsik Moon, Shunsuke Saito, Weipeng Xu, Rohan Joshi, Julia Buffalini, Harley Bellan, Nicholas Rosen, Jesse Richardson, Mallorie Mize, Philippe de Bree, Tomas Simon, Bo Peng, Shubham Garg, Kevyn McPhail, and Takaaki Shiratori. A dataset of relighted 3D interacting hands. In *NeurIPS*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/396beafa6feba781a7114780e6837253-Abstract-Datasets_and_Benchmarks.html.

Haozhe Qi, Chen Zhao, Mathieu Salzmann, and Alexander Mathis. HOISDF: Constraining 3D hand-object pose estimation with global signed distance fields. In *CVPR*, pp. 10392–10402, 2024. doi: 10.1109/CVPR52733.2024.00989. URL https://arxiv.org/abs/2402.17062.

Yi-Ling Qiao, Junbang Liang, Vladlen Koltun, and Ming C. Lin. Scalable differentiable physics for learning and control. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7847–7856. PMLR, 2020. URL https://arxiv.org/abs/2007.02168.

Patrick Rim, Kevin Harris, Braden Copple, et al. SHOW3D: Capturing scenes of 3D hands and objects in the wild. In *CVPR*, 2026. URL https://show3d-dataset.github.io/.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6), 2017. doi: 10.1145/3130800.3130883. URL https://doi.org/10.1145/3130800.3130883.

Soshi Shimada, Vladislav Golyanik, Patrick Pérez, and Christian Theobalt. Decaf: Monocular deformation capture for face and hand interactions. *ACM Transactions on Graphics*, 42(11), 2023. doi: 10.1145/3618329. URL https://vcai.mpi-inf.mpg.de/projects/Decaf/.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Mohan Kumar Srirama, Sudeep Dasari, Shikhar Bahl, and Abhinav Gupta. HRP: Human affordances for robotic pre-training. In *RSS*, 2024. doi: 10.15607/RSS.2024.XX.068. URL https://www.roboticsproceedings.org/rss20/p068.html.

Bartolomeo Stellato, Goran Banjac, Paul Goulart, Alberto Bemporad, and Stephen Boyd. OSQP: An operator splitting solver for quadratic programs. *Mathematical Programming Computation*, 12(4):637–672, 2020. doi: 10.1007/s12532-020-00179-2. URL https://stanford.edu/~boyd/papers/osqp.html.

The Newton Contributors. Newton: GPU-accelerated physics simulation for robotics and simulation research. Software, 2025. URL https://github.com/newton-physics/newton.

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 *ICCV*, 2023. doi: 10.1109/ICCV51070.2023.00735. URL https://arxiv.org/abs/2309.15273.

Tze Ho Elden Tse, Zhongqun Zhang, Kwang In Kim, Ales Leonardis, Feng Zheng, and Hyung Jin Chang. S^2Contact: Graph-based network for 3D hand-object contact estimation with semi-supervised learning. In *ECCV*, volume 13661 of *Lecture Notes in Computer Science*, pp. 568–584. Springer, Cham, 2022. doi: 10.1007/978-3-031-19769-7_33. URL https://arxiv.org/abs/2208.00874.

Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision*, 2016. URL https://hic.is.tuebingen.mpg.de/.

Jikai Wang, Qifan Zhang, Yu-Wei Chao, Bowen Wen, Xiaohu Guo, and Yu Xiang. HO-Cap: A capture system and dataset for 3D reconstruction and pose tracking of hand-object interaction. In *NeurIPS*, 2025. URL https://arxiv.org/abs/2406.06843.

Zikai Wang, Zhilu Zhang, Yiqing Wang, Hui Li, and Wangmeng Zuo. ArtHOI: Taming foundation models for monocular 4D reconstruction of hand-articulated-object interactions. In *CVPR*, pp. 15998–16009, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_ArtHOI_Taming_Foundation_Models_for_Monocular_4D_Reconstruction_of_Hand-Articulated-Object_CVPR_2026_paper.html.

Keenon Werling, Dalton Omens, Jeongseok Lee, Ioannis Exarchos, and C. Karen Liu. Fast and feature-complete differentiable physics engine for articulated rigid bodies with contact constraints. In *RSS*, 2021. doi: 10.15607/RSS.2021.XVII.034. URL https://arxiv.org/abs/2103.16021.

Wenqi Xian, Zhengqi Li, Matthew Fisher, Jonathan Eisenmann, Eli Shechtman, and Noah Snavely. UprightNet: Geometry-aware camera orientation estimation from single images. In *ICCV*, 2019. doi: 10.1109/ICCV.2019.01007. URL https://openaccess.thecvf.com/content_ICCV_2019/papers/Xian_UprightNet_Geometry-Aware_Camera_Orientation_Estimation_From_Single_Images_ICCV_2019_paper.pdf.

Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. CHORE: Contact, human and object reconstruction from a single RGB image. In *ECCV*, volume 13662 of *Lecture Notes in Computer Science*, pp. 125–145. Springer, Cham, 2022. doi: 10.1007/978-3-031-20086-1_8. URL https://arxiv.org/abs/2204.02445.

Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. CPF: Learning a contact potential field to model the hand-object interaction. *arXiv preprint*, 2021. URL https://arxiv.org/abs/2012.00924.

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 *CVPR*, pp. 20953–20962, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Yang_OakInk_A_Large-Scale_Knowledge_Repository_for_Understanding_Hand-Object_Interaction_CVPR_2022_paper.html.

Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4D: 4D instance segmentation of close human interaction. In *CVPR*, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Yin_Hi4D_4D_Instance_Segmentation_of_Close_Human_Interaction_CVPR_2023_paper.html.

Zhenjun Yu, Wenqiang Xu, Pengfei Xie, Yutong Li, Brian W. Anthony, Zhuorui Zhang, and Cewu Lu. Dynamic reconstruction of hand-object interaction with distributed force-aware contact representation. In *ICCV*, pp. 8590–8599, 2025. doi: 10.1109/ICCV51701.2025.00804. URL https://arxiv.org/abs/2411.09572.

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. In *ICML*, 2026. URL https://arxiv.org/abs/2606.09243.

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 *CVPR*, 2024. URL https://oakink.net/v2/.

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

Yiming Zhao, Taein Kwon, Paul Streli, Marc Pollefeys, and Christian Holz. EgoPressure: A dataset for hand pressure and pose estimation in egocentric vision. In *CVPR*, pp. 27727–27738, 2025. doi: 10.1109/CVPR52734.2025.02582. URL https://arxiv.org/abs/2409.02224.

Danyan Zhou, Jinxuan Lu, Jiawei Lin, Tianxing Chen, Chuqiao Lyu, and Wenbo Ding. TouchSight: Bare-handed tactile prediction from egocentric video via generative visual augmentation. *arXiv preprint*, 2026. doi: 10.48550/arXiv.2609.20414. URL https://arxiv.org/abs/2609.20414.

APPENDIX ROADMAP

The appendix presents dataset composition and label construction (Appendix A), additional results (Appendix B), model and training details (Appendix C), and evaluation protocols and coverage (Appendix D).

A DATASET AND LABEL CONSTRUCTION

A.1 DATASET INVENTORY

Table 4 lists the 24 curated sources and their label types. We derive contact and proximity on the canonical MANO surface from source geometry. Sources with motion-derived force labels are grouped by whether the object motion used to construct them is rigid or articulated.

Counts describe our converted collection, not the full dataset releases or the number of training samples. Frames are counted across all captured camera views. Except for ObMan and MOW ($*$), sources contain ordered sequences; for those two datasets, "sequences" are groups of independent images.

For GigaHands (†), we select 15 objects because the remaining objects have less reliable tracking. The table counts all camera views of sequences containing at least one selected object, including frames in those sequences that are excluded from training. HOCap (‡) contributes to the collection total but is excluded from training. PROX contributes 178 distinct source frames captured at 5 Hz; the table reports this original frame count. HO3Dv2/v3 refer to HO-3D release versions, and OakInk-v1/v2 to OakInk and OakInk2.

Table 4: Data sources and labels in the curated collection.

Dataset	Sequences	Frames
Force labels: articulated objects		
ARCTIC (Fan et al., 2023)	300	1 903 869
HOI4D (Liu et al., 2022)	2 973	891 900
Force labels: rigid objects		
GigaHands† (Fu et al., 2025)	1 424	9 449 724
Show3D (Rim et al., 2026)	1 682	6 931 884
TACO (Liu et al., 2024)	2 120	4 287 490
HOGraspNet (Cho et al., 2024)	13 575	4 270 708
OakInk-v2 (Zhan et al., 2024)	611	3 898 190
HOT3D (Banerjee et al., 2025)	198	701 519
HOCap‡ (Wang et al., 2025)	64	653 175
DexYCB (Chao et al., 2021)	1 000	581 968
H2O (Kwon et al., 2021)	184	453 855
OakInk-v1 (Yang et al., 2022)	792	235 604
HO3Dv3 (Hampali et al., 2020)	17	103 722
HO3Dv2 (Hampali et al., 2020)	17	77 953
H2O3D (Hampali et al., 2022)	15	67 525
Contact and proximity labels only		
InterHand2.6M (Moon et al., 2020)	255 378	12 469 860
ReInterHand (Moon et al., 2023)	250	886 470
DECAF (Shimada et al., 2023)	16	778 817
Hi4D (Yin et al., 2023)	1 600	191 120
ObMan* (Hasson et al., 2019)	23	118 133
RICH (Huang et al., 2022)	111	62 969
HIC (Tzionas et al., 2016)	11	1 824
MOW* (Cao et al., 2021)	512	512
PROX (Hassan et al., 2019)	1	178
Total	**282 874**	**49 018 969**

Sources for which force labels can be constructed account for approximately 70% of curated frames; this does not mean that every frame contributes force supervision. GigaHands contributes contact and proximity supervision, while its motion-derived force targets are omitted because of noise in object poses. HOCap is reserved for evaluation. The rollout loss uses eligible rigid-object episodes only; articulated episodes provide vertex-force supervision.

A.2 Geometric Contact-Label Comparison

Following HACO (Jung & Lee, 2025), we use distance thresholds of 10 mm for hand–object sources, 35 mm for MOW, and 5 mm for InterHand2.6M, HIC, and ReInterHand; all remaining sources use 10 mm. Equation 3 further applies normal filtering and inside-mesh inclusion, as illustrated in Figure 6.

Multiray inside–outside test. To determine whether a hand vertex lies inside a closed object mesh, we transform it into object coordinates and cast several oblique rays against the original, unrepaired triangles. For each ray, an odd number of intersections indicates that the vertex is inside; a majority vote across rays sets the sign of its nearest-surface distance. A small offset in the ray test avoids ambiguous intersections at edges and vertices. For open or unreliable surfaces, we retain the unsigned distance and skip inside-mesh inclusion.

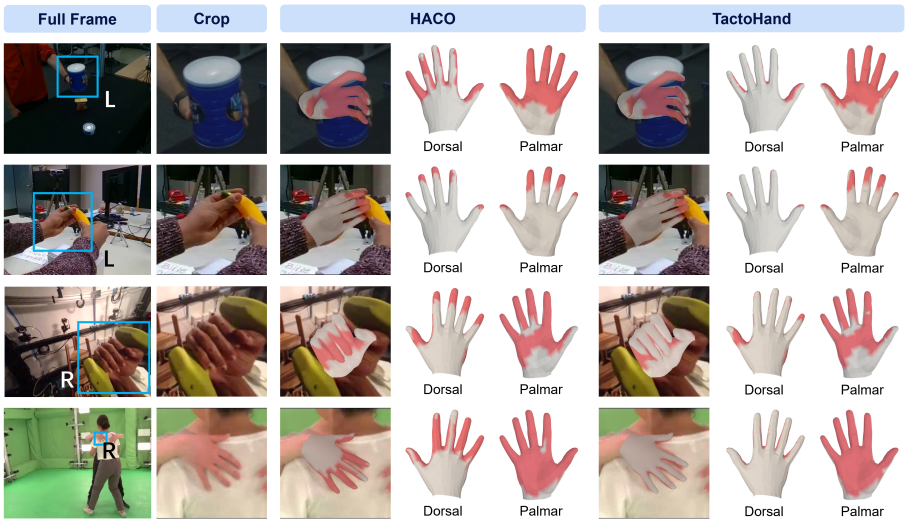

Figure 6: Comparison of contact labels generated by HACO's distance-only rule and our geometric rule. Normal filtering removes back-facing false positives, while inside-mesh inclusion recovers penetrating vertices missed by distance thresholding.

A.3 Motion-Derived Force Labels

A.3.1 Trajectory Processing

Temporal sampling and smoothing. We interpolate observed object trajectories onto a fixed 30 Hz timeline for both rigid- and articulated-object force labeling. Translation is smoothed with a symmetric, truncated Gaussian kernel. For rotation, we project each input matrix onto $\mathrm{SO}(3)$ by SVD, then compute a weighted intrinsic mean. We initialize the rotation estimate at the center rotation and update it iteratively as follows:

$$\boldsymbol{\delta}_t^{(k)} = \sum_{j=-r_R}^{r_R} g_j^R \,\mathrm{Log}\Big((\bar{\mathbf{R}}_t^{(k)})^\top \mathbf{R}_{t+j}\Big), \qquad \bar{\mathbf{R}}_t^{(k+1)} = \bar{\mathbf{R}}_t^{(k)} \,\mathrm{Exp}(\boldsymbol{\delta}_t^{(k)}), \tag{11}$$

where g_j^R are normalized Gaussian weights. We use eight iterations and bandwidths of $\sigma_p = 0.075\,\mathrm{s}$ for translation and $\sigma_R = 0.150\,\mathrm{s}$ for rotation. At solver rate f, either kernel has radius $r(\sigma, f) = \max\{\mathrm{round}(0.30f), \lceil 3\sigma f \rceil\}$ frames. We clamp indices outside the trajectory to the nearest endpoint when smoothing, but supervise a transition only when the full smoothing windows around both states fall within the observed trajectory.

A.3.2 Rigid-Object Force Allocation and Validation

Trajectory-derived target wrench. For a rigid object, let $(\bar{\mathbf{x}}_t, \bar{\mathbf{R}}_t)$ denote its smoothed center-of-mass pose and let $\Delta t = 1/f$. The displacement over transition $t \to t+1$ gives the outgoing linear velocity:

$$\mathbf{v}_{t+1} = \frac{\bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t}{\Delta t}. \tag{12}$$

We obtain the outgoing world-frame angular velocity from $\bar{\mathbf{R}}_{t+1}\bar{\mathbf{R}}_t^\top$ using the Cayley map. With $\boldsymbol{\phi}_t = \mathrm{Log}(\bar{\mathbf{R}}_{t+1}\bar{\mathbf{R}}_t^\top)$ and $\theta_t = \|\boldsymbol{\phi}_t\|_2$,

$$\boldsymbol{\omega}_{t+1}^w = \frac{2\tan(\theta_t/2)}{\Delta t\,\theta_t}\boldsymbol{\phi}_t, \tag{13}$$

with the continuous limit $\boldsymbol{\phi}_t/\Delta t$ as $\theta_t \to 0$. The incoming velocity is the preceding transition's outgoing velocity; for the first transition, we copy its outgoing velocity. Dividing the velocity change by Δt yields the acceleration used in target wrench $\mathbf{w}_t^\star$. We express angular velocity and acceleration in the body frame when computing the torque in Equation 4.

We obtain gravity direction from dataset metadata where available, UprightNet RGB estimates (Xian et al., 2019), or a manually inferred scene or world-down axis. Camera-frame estimates are transformed into the solver world frame. The inverse-dynamics target uses a $1\,\mathrm{kg}$ reference mass and the corresponding scaled inertia, without rollout optimization.

Candidate vertices and grasp map. We consider force candidates for a hand–object pair only when at least one hand vertex is labeled as contact under the 10 mm rule in Section 4. Within that pair, vertices with $s_{t,i} < 20\,\mathrm{mm}$ are candidates, weighted by Equation 5.

Let $\mathbf{p}_i$ and $\mathbf{f}_i$ denote the world position and hand-on-object force of an active vertex. The grasp map

$$\mathbf{G}_i\mathbf{f}_i = \begin{bmatrix} \mathbf{f}_i \\ (\mathbf{p}_i - \mathbf{x}) \times \mathbf{f}_i \end{bmatrix} \tag{14}$$

maps its force to a wrench about the object's center of mass $\mathbf{x}$. Candidate normals $\mathbf{n}_{t,i}$ are normalized, and each force is constrained to a six-facet unilateral Coulomb cone with $\mu = 1.2$.

Environment support. In Equation 6, $e_t \geq 0$ denotes upward support at an estimated tabletop contact point, and $\mathbf{G}_t^{\mathrm{env}}$ maps it to a wrench about the object center of mass. For tabletop sequences, we estimate this point from mesh vertices within 2 mm of the object's lowest point. We infer whether the object rests on the table or is held above it by comparing its height with nearby hand-free frames. For each hand-contact interval, median heights from up to 15 such frames on either side provide reference heights before and after contact. Before identifying these intervals, we bridge gaps of at most five solver frames in the hand-contact signal. A rise of more than 7 mm for three frames marks pickup; a sustained return to the post-contact reference height marks placement. If the object is already held at the start of an interval, we use the post-contact reference when available; without reference frames, we treat it as off-table.

These support states are estimates rather than observed labels. We accept a frame without nearby active hand force only if table support alone can explain its motion-derived target wrench. When hand and environment contact coexist, the QP jointly allocates hand forces and nonnegative support at the estimated point.

QP implementation. We solve Equation 6 as a sparse sequence QP. The force-magnitude, force-temporal, and environment-temporal coefficients default to one; the environment-magnitude coefficient and priority denominator ϵ are 10^{-4}. Torque equations are scaled by $0.05\ \mathrm{N\,m}$. Temporal differences use the union of adjacent active-vertex sets, treating absent vertices as zero. OSQP (Stellato et al., 2020) uses absolute and relative tolerances of 10^{-6}. Partitioning, cached factorizations, and warm starts leave the objective unchanged.

Validity and replay. We reconstruct the fitted wrench from the solved vertex and environment forces and record its force and torque residuals. Independently, we replay that wrench from the stored smoothed state in Newton Physics (The Newton Contributors, 2025) and compare the one-step pose with the next smoothed state.

A.3.3 Articulated-Object Force Labels

For an articulated object with J joints, we retain its kinematic tree on the same 30 Hz timeline. We distribute the 1 kg reference mass among links in proportion to mesh volume and compute each link's center of mass and inertia at its assigned mass. Multibody inverse dynamics on the smoothed trajectory gives the required generalized effort. Subtracting the generalized gravity contribution yields the contact target

$$\boldsymbol{\eta}_t^\star = \boldsymbol{\eta}_t^{\mathrm{ID}} - \boldsymbol{\eta}_t^g \in \mathbb{R}^{6+J}, \tag{15}$$

whose first six components are the floating-base force and torque, expressed in the solver world frame with torque about the system center of mass; the remaining components are joint efforts. The target can be computed from smoothed base and joint trajectories or, equivalently, by projecting per-link Newton–Euler wrenches onto these generalized coordinates.

We assign each candidate hand vertex to its nearest object part. A force $\mathbf{f}_{t,i}$ applied at $\mathbf{p}_{t,i}$ contributes $\mathbf{f}_{t,i}$ to the base force and $(\mathbf{p}_{t,i} - \mathbf{c}_t) \times \mathbf{f}_{t,i}$ to torque about the system center of mass $\mathbf{c}_t$. For a joint j ancestral to the contacted part, its joint-effort contribution is

$$\gamma_{t,i,j} = \begin{cases} \mathbf{a}_{t,j}^\top \left((\mathbf{p}_{t,i} - \mathbf{o}_{t,j}) \times \mathbf{f}_{t,i} \right), & \text{revolute,} \\ \mathbf{a}_{t,j}^\top \mathbf{f}_{t,i}, & \text{prismatic,} \end{cases} \tag{16}$$

where $\mathbf{a}_{t,j}$ and $\mathbf{o}_{t,j}$ are the joint axis and pivot in world coordinates. Joints outside the part's ancestor chain receive zero contribution. This part-aware map replaces the rigid grasp map in a sequence-level allocation with the same signed-distance priority, unilateral friction constraints, and temporal force penalty.

Permitted environment support contributes to base balance. Joint effort unexplained by contact forces remains an internal, non-hand contribution instead of being assigned to hand vertices. Hand-free frames have zero hand force. If a hand-contact allocation fails, we record that the force estimate is invalid or that environment support was used as a fallback, rather than treating the failed estimate as a valid hand force. We check fitted generalized effort against its target and use valid articulated vertex forces for field-level supervision, without the rigid rollout loss.

B Additional Results and Visualizations

Table 5 extends the main reconstruction comparison with joint errors and hand–object F-scores. Across all six method–dataset pairs, TactoHand also improves these metrics over the base and HACO variants. Figure 7 illustrates the resulting hand–object alignment on HOCap, while Figure 8 separates the effects of the contact and proximity terms on individual examples.

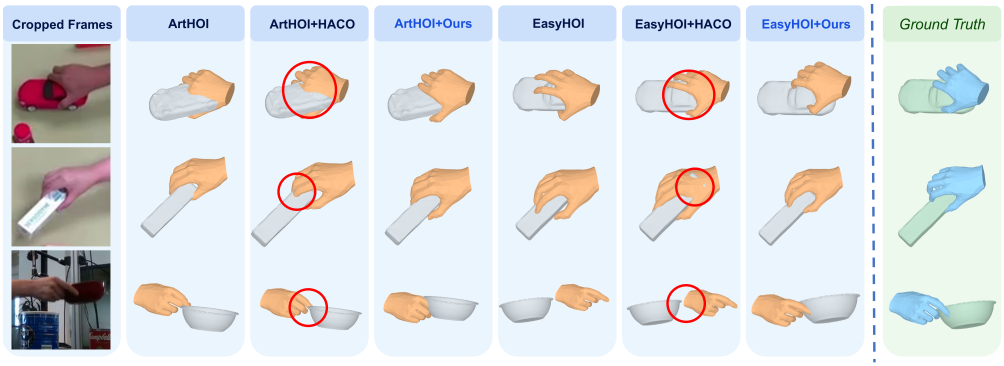

Figure 7: HOCap (Wang et al., 2025) reconstructions. Red circles highlight hand–object contact artifacts in the HACO variants.

Table 5: Full reconstruction results. PA-VPE/JPE: mm; $\mathrm{CD}_{\mathrm{ho}}$: $10^{-3}\,\mathrm{m}^2$; F-5/F-10: 5/10 mm; ACC: $\mathrm{mm/s}^2$. Bold marks the best per method–dataset block; shading highlights PA-VPE and $\mathrm{CD}_{\mathrm{ho}}$.

Methods	PA-VPE ↓	PA-JPE ↓	$\mathrm{CD}_{\mathrm{ho}}$ ↓	F-5_{ho} ↑	F-10_{ho} ↑	ACC ↓
DexYCB						
EasyHOI	9.937	10.483	24.140	0.3609	0.5359	–
EasyHOI + HACO	9.516	10.032	12.552	0.3609	0.5392	–
EasyHOI + Ours	**9.077**	**9.559**	**2.004**	**0.3817**	**0.5606**	–
ArtHOI	7.317	7.520	11.587	0.3930	0.5240	4311.6
ArtHOI + HACO	6.027	6.190	13.627	0.3970	0.5300	3978.9
ArtHOI + Ours	**4.626**	**4.747**	**1.431**	**0.4540**	**0.6200**	**3892.7**
HOCap						
EasyHOI	9.179	9.790	59.416	0.2306	0.3991	–
EasyHOI + HACO	9.051	9.623	68.943	0.2216	0.3824	–
EasyHOI + Ours	**7.592**	**8.075**	**11.461**	**0.2456**	**0.4319**	–
ArtHOI	6.984	7.243	5.234	0.3215	0.5085	5176.9
ArtHOI + HACO	6.684	6.892	5.199	0.3249	0.5118	4666.3
ArtHOI + Ours	**6.073**	**6.264**	**1.104**	**0.3672**	**0.5869**	**4599.7**
MOW						
EasyHOI	14.832	15.860	69.080	0.2262	0.3728	–
EasyHOI + HACO	14.276	15.215	72.638	0.2195	0.3617	–
EasyHOI + Ours	**13.939**	**14.715**	**15.266**	**0.2370**	**0.3888**	–
ArtHOI	13.572	14.054	73.993	0.2330	0.3640	–
ArtHOI + HACO	13.261	13.727	73.147	0.2340	0.3670	–
ArtHOI + Ours	**13.100**	**13.545**	**32.413**	**0.2610**	**0.4060**	–

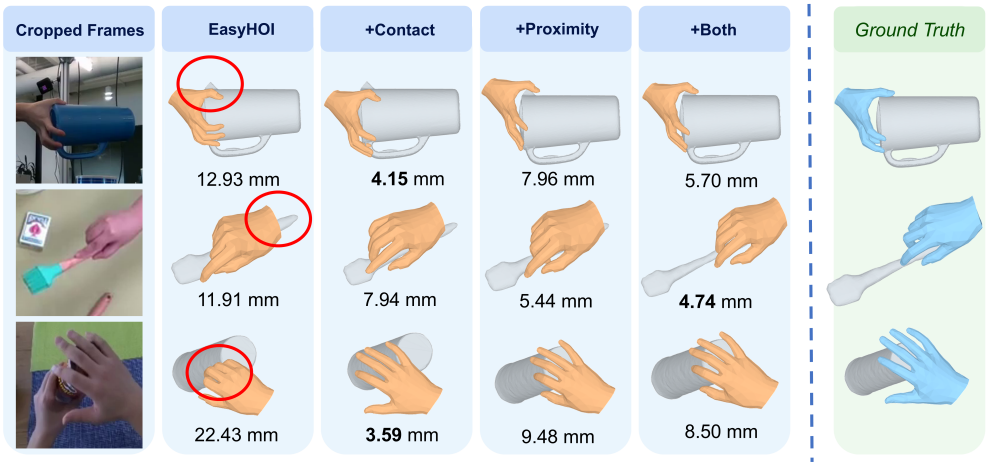

Figure 8: EasyHOI (Liu et al., 2025) reconstruction ablation. + Contact: contact-normal term; + Proximity: distance term; + Both: both terms. Values are per-example PA-VPE (mm).

C Model and Training Details

C.1 Network Architecture

A trainable DINOv3 ViT-L encodes each 256×256 RGB frame independently. We average its final patch tokens, excluding class and register tokens, to obtain one 1024-dimensional feature vector per frame ($D = 1024$ in Figure 1). A six-layer temporal transformer with eight attention heads processes these features with learned positional embeddings. A shared two-layer, eight-head query decoder processes three learned queries, one each for force, contact, and proximity. At each image timestamp, each decoded query is fused with the corresponding frame feature to predict its field for both hands. Transformer feed-forward layers have width $4D$, and dropout is zero.

The four variants combine two temporal modes (SF and T) with rollout supervision (Off and On). T processes the full clip through the temporal transformer and shared query decoder. SF applies

the same modules to each frame as a length-one sequence, preserving the architecture and output shapes while removing cross-frame attention. All variants share the force, contact, and proximity head designs; On enables the rollout loss, whereas Off sets its weight to zero.

C.2 Force Coordinates, Temporal Sampling, and Supervision

Vertex-local force coordinates. Figure 3 (lower left) shows the local basis used by the force head. Its normal is the outward normal of the posed MANO mesh; one tangent points approximately toward the wrist, and the other completes an orthonormal basis. We expand the predicted force components along these axes in camera coordinates before force supervision and wrench aggregation. Contact and proximity retain canonical vertex indexing.

Vertex-force loss. Before computing the camera-frame force loss, we cap predicted and target magnitudes at 20 N by rescaling each vector without changing its direction. We then separate targets by capped magnitude: active targets above 0.05 N receive a vertex-weighted Huber penalty with a 0.05 N transition, while the remaining targets receive mean squared error weighted by 0.1.

Differentiable rigid-body rollout. The model predicts fields at 10 Hz. Each 100 ms visual interval spans three 30 Hz force intervals: the first uses predicted vertex forces, and the next two use detached reference forces. Each force interval contains four 120 Hz integration steps of size $h = 1/120$ s, following Equation 17. The object state propagates continuously, without resets to ground-truth poses.

For one step, let $(\mathbf{x}, \mathbf{v}, \mathbf{q}, \boldsymbol{\omega}^w)$ denote the center-of-mass position, linear velocity, orientation quaternion, and world-frame angular velocity, and let $(\mathbf{F}, \boldsymbol{\tau})$ be the net world-frame wrench about the center of mass. Define $\mathbf{R} = \mathbf{R}(\mathbf{q})$, $\boldsymbol{\omega}^b = \mathbf{R}^\top \boldsymbol{\omega}^w$, $\boldsymbol{\tau}^b = \mathbf{R}^\top \boldsymbol{\tau}$, and $\bar{\boldsymbol{\omega}}^b = (\boldsymbol{\omega}^b + \boldsymbol{\omega}^{b,+})/2$. With body-frame inertia $\mathbf{I}$, the update is

$$\begin{aligned} \mathbf{v}^+ &= \mathbf{v} + \tfrac{h}{m}\mathbf{F}, \\ \mathbf{x}^+ &= \mathbf{x} + h\mathbf{v}^+, \\ \boldsymbol{\omega}^{b,+} &= \boldsymbol{\omega}^b + h\mathbf{I}^{-1}\big(\boldsymbol{\tau}^b - \bar{\boldsymbol{\omega}}^b \times \mathbf{I}\bar{\boldsymbol{\omega}}^b\big), \\ \boldsymbol{\omega}^{w,+} &= \mathbf{R}\boldsymbol{\omega}^{b,+}, \\ \mathbf{q}^+ &= \mathrm{normalize}\big(\mathcal{Q}(h\boldsymbol{\omega}^{w,+}) \otimes \mathbf{q}\big)\,. \end{aligned} \tag{17}$$

We solve the body-frame angular-velocity update with six unrolled Newton iterations. The pre-update rotation $\mathbf{R}$ then maps the result to the world frame for the orientation update; $\mathcal{Q}$ maps an axis-angle vector to a unit quaternion. All state updates are differentiable, allowing the rollout loss to backpropagate through wrench aggregation to predicted vertex forces. Each visual interval contains one force interval driven by predictions; the other two use fixed reference forces while preserving gradients through the evolving object state.

Supervision eligibility. The available annotations determine which losses a sequence contributes to. Geometry-only sequences contribute contact and proximity losses. Rigid and articulated sequences with force labels also contribute force loss at image timestamps where those labels are valid. A field-level loss is evaluated only when the corresponding hand and label are available, regardless of whether that hand is occluded in the image. Invalid force labels do not contribute to the force loss.

Rigid rollout further requires a trajectory whose hand force can be attributed to the target object. We exclude a trajectory if, at any force-valid time step, the hand applies nonzero force to another object while the target receives none. Repeated boundary frames used to complete a visual clip do not contribute to force or rollout losses; both use only time intervals within the recorded sequence.

C.3 Training Configuration

The four variants share the final-stage settings in Table 6. Dataset sampling weights are proportional to the number of available 32-frame windows, multiplied by the listed dataset-specific factors.

Table 6: Training settings and dataset sampling multipliers. Unlisted training sources use a multiplier of 1.

Setting	Value
Hardware and batch	One node with eight A800 GPUs; 7 samples per GPU; two accumulation steps (112 samples per update)
Optimizer	AdamW; learning rate 2×10^{-5} for backbone and heads; weight decay 10^{-4}; cosine decay with two-epoch warmup
Window sampling	32 frames, stride 32; each epoch samples 20% of the windows allocated to each dataset
Rollout-loss weight	10^{-4} with physics; 0 without physics
Dataset sampling multipliers	4: MOW; 2: DexYCB; 0.5: HOI4D, OakInk-v2; 0.1: DECAF, ReInterHand; 0.05: HOGraspNet, ObMan, GigaHands; 0.02: InterHand2.6M

Each 150-epoch final-stage run initializes model weights from an earlier checkpoint but resets the optimizer, scheduler, and epoch state. In Equation 10, $\lambda_c = 1$, $\lambda_d = 2$, and $\lambda_p = 10$, with $\mathcal{L}_{\text{contact}} = \mathcal{L}_{\text{conditional}} + 0.25\mathcal{L}_{\text{engagement}}$. No separate direct wrench loss is used. For variants with physics supervision, λ_r increases from zero in the second epoch to 10^{-4} in the sixth epoch; it remains zero for variants without physics supervision. In Equation 9, $\mathcal{H} = \{1, \ldots, 12\}$ spans four 10 Hz intervals, $\lambda_x = [2(0.01\,\text{m})^2]^{-1}$, and $\lambda_R = [2(0.1\,\text{rad})^2]^{-1}$. Rollout losses are also averaged over eligible windows. At test time, contact is predicted when $\hat{c} > 0.5$.

We report the checkpoints with the best validation performance. Our validation sets are the nominal DexYCB test split and the ARCTIC P2 validation split, and the reported results are measured on these same sets. HOCap is excluded from training. Its results characterize cross-dataset transfer.

D Evaluation Protocol and Coverage

D.1 Contact and Proximity Comparisons

Table 1 re-scores released checkpoints on our geometry-derived annotations and nominal dataset splits. Input pipelines can differ. SF/T denote single-frame/temporal fusion; a dash denotes an unavailable result or undefined acceleration for static MOW. MOW uses SF for proximity and T for contact. For MOW, we form the temporal input by repeating each image across the 32-frame clip; no additional temporal observations are available.

Metric definitions. We decode predicted proximity $\widehat{r}$ as $\widehat{d} = -\log(\widehat{r})/\tau$ in metres and report mean absolute distance error (ADE) in mm over valid vertices. ADE-100 clips both prediction and target at 100 mm before computing the error; all valid vertices remain in the average. This matches InterField's 100 mm output range, whereas TactoHand's target extends to 250 mm. For distance error $e_{t,i} = \widehat{d}_{t,i} - d_{t,i}$, Accel. averages $|e_{t+1,i} - 2e_{t,i} + e_{t-1,i}|$ multiplied by the squared frame-sampling rate, in m/s^2. It measures temporal variation in distance error. We exclude any three-frame group containing a repeated frame inserted to pad a clip.

For Macro-F1, we compute F1 over the vertices of each hand in each frame, then average over frame–hand pairs with contact in either the ground truth or the prediction. Failing to predict any ground-truth contact on a hand, or predicting contact on a hand with none, gives a score of zero. Pairs with no contact in either are excluded from this average. Micro-F1 pools vertex-level true positives (TP), false positives (FP), and false negatives (FN) over all valid frame–hand pairs and computes $2TP/(2TP + FP + FN)$. Correctly predicted non-contact vertices do not increase either score.

Input protocols. ARCTIC uses its official P2 validation split (subject s05, egocentric view, all labeled frames); DexYCB, HOCap, and MOW use their nominal test splits. On ARCTIC, InterField uses its released P2 egocentric checkpoints with the official preprocessing of raw fisheye images and crops around the interaction bounding box. Its reported ADE against its original reference labels is 8.83 mm for SF and 8.40 mm for LSTM. On DexYCB, HOCap, and MOW, InterField instead uses P1 allocentric checkpoints on hand crops, making these out-of-domain evaluations. EgoPHI

is trained only on ARCTIC and uses its released weights and deployed contact threshold of 0.3. Following its training protocol, it uses native full frames except for MOW, where it uses hand crops. TactoHand uses full-frame images on DexYCB, HOCap, and ARCTIC, and hand crops on MOW, where the hand occupies a small fraction of the frame.

D.2 Downstream Reconstruction Protocol

We evaluate TactoHand's contact and proximity fields in EasyHOI (Liu et al., 2025) and ArtHOI (Wang et al., 2026) on DexYCB (Chao et al., 2021), HOCap (Wang et al., 2025), and MOW (Cao et al., 2021). DexYCB and MOW are included in training; HOCap is excluded. MOW contains independent images, while the other two provide video. Within each reconstruction method, variants share the model, initialization, images, and optimization budget. Table 5 compares base, HACO (Jung & Lee, 2025), and TactoHand variants; Figure 8 isolates the EasyHOI contact and proximity terms.

Shared interaction terms. We hold the predicted contact and proximity values at the 778 MANO vertices fixed throughout reconstruction refinement. Proximity is decoded as $\widehat{d} = -\log(\widehat{r})/\tau$ with $\tau = 20\,\mathrm{m}^{-1}$; values saturated at $0.25\,\mathrm{m}$ specify a lower bound on distance. The proximity term applies a Huber penalty ($\delta = 1$) to signed distance residuals normalized by $5\,\mathrm{mm}$. In the contact-normal term, predicted contact probability weights the factor $\max(0, \mathbf{n}_h^\top \mathbf{n}_o + 0.5)^2$, where $\mathbf{n}_h$ and $\mathbf{n}_o$ denote the hand and object surface normals. The proximity and contact-normal terms both have coefficient 1. They replace the original contact objective, while reconstruction, penetration, and regularization terms remain.

Table 7: Reconstruction stages and optimization settings shared across the compared variants.

Stage	Updates	Optimized variables	Learning rate
ArtHOI alignment	250	Hand translation	$10^{-2} \to 5 \times 10^{-4}$
ArtHOI hand refinement	399	Finger pose, hand translation	$10^{-3} \to 10^{-4}$
ArtHOI penetration correction	300	Hand translation, global orientation	10^{-3}
ArtHOI smoothing	300	Hand translation	5×10^{-4}
EasyHOI global alignment	20×100	Hand-global parameters, object translation	5×10^{-2}
EasyHOI hand refinement	1,000	Pose residual, object translation; MANO shape	10^{-2}; 10^{-4} for shape

EasyHOI refinement. We fix metric CAD geometry and ground-truth object orientation, optimizing translation from its image-based initialization. Table 7 lists 3,000 updates, excluding image-based initialization and auxiliary PCA fitting, with constant learning rates within each stage. Hand refinement starts from the aligned hand parameters, without backpropagating through the preceding alignment stage; the hand's root rotation remains fixed during refinement. During global alignment, interaction losses are enabled under the same conditions as in the original method. The proximity term ramps up over the first 50 updates in which those losses are evaluated, while the contact-normal term is off. During refinement, proximity is fully enabled and the contact-normal term ramps up over 10 such updates.

ArtHOI refinement. We fix metric object geometry, skip scale optimization, and use Adam with $(\beta_1, \beta_2) = (0.9, 0.999)$. Table 7 lists the 1,249 refinement updates performed after initialization. For the first two stages, learning rates are taken from step 151 onward of the original 800-step schedule. An L1 penalty with coefficient 0.4 discourages deviation from the initial pose. The proximity term ramps up over 50 alignment iterations, and the contact-normal term over 10 hand-refinement iterations. Both remain active during penetration correction; only proximity remains during smoothing.

PA-VPE and PA-JPE are Procrustes-aligned mean vertex and joint errors, respectively, in mm. $\mathrm{CD}_{\mathrm{ho}}$ sums bidirectional mean squared nearest-neighbor distances in $10^{-3}\,\mathrm{m}^2$; $\mathrm{F\text{-}5}_{\mathrm{ho}}$ and $\mathrm{F\text{-}10}_{\mathrm{ho}}$ are

hand–object F-scores at 5 and 10 mm. ACC is acceleration error in $\mathrm{mm/s^2}$, available only for ArtHOI on DexYCB and HOCap; dashes in Table 5 mean unavailable. These are reconstruction outcomes, not direct contact or proximity scores.

D.3 Force and Rollout Evaluation

Table 2 evaluates four single-frame/temporal $\times$ physics-off/on variants on the DexYCB test split. They share data, backbone, and optimization settings (Appendix C.1); physics-on adds rigid rollout loss. Force metrics compare predicted and motion-derived reference vectors in the camera frame at image timestamps with valid labels, excluding repeated frames used for padding. Vertex MAE averages component-wise absolute error over all supervised vertices, including zero-force targets. Active MAE and force angle use only targets above 0.05 N; angle is the mean predicted-to-reference vector angle. MAEs are in mN and angles in degrees.

Rollout uses a 0.4 s horizon: four 10 Hz visual intervals or twelve 30 Hz force intervals. Within each visual interval, predicted forces drive the first force interval and detached reference forces drive the next two; four 120 Hz updates integrate each force interval. Sliding windows start from observed object states and include specified gravity and estimated environment-support forces. We score only intervals within the recorded sequence, excluding those added to pad a clip. Position ADE averages object-center displacement error over the horizon, and rollout angle averages rotation geodesic error. A window is excluded if any state is non-finite or its linear or angular speed exceeds 100 m/s or 1000 rad/s, respectively. Reported errors therefore describe the non-divergent windows of this reference-assisted rollout.