EgoTac: In-the-wild Tactile Prediction from Egocentric Vision

Wenkang Zhang[1,2] **Chengbo Yuan**[3] **Zicheng Zhang**[4] **Zhengxue Cheng**[2] **Yang Gao**[1,3*]
[1]Shanghai Qi Zhi Institute [2]Shanghai Jiao Tong University
[3]Tsinghua University [4]Fudan University
conquer.wkzhang@sjtu.edu.cn, gaoyangiiis@mail.tsinghua.edu.cn

Abstract

Touch is fundamental to dexterous manipulation, yet most egocentric human data increasingly used for robot learning lacks tactile information. Directly collecting large-scale tactile data is challenging due to sensor limitations, while human video data is abundant, contact-rich, and easily scalable. This motivates a natural question: can tactile signals be inferred purely from vision? To address this, we introduce *EgoTac*, a generalizable model that predicts rich tactile information directly from egocentric human videos. EgoTac is trained on a unified corpus of over 5.7M image–tactile pairs, covering both continuous force measurements and binary contacts. By learning from this diverse dataset, EgoTac captures nuanced touch dynamics across varied interactions. Experiments demonstrate strong performance: in-domain prediction achieves an average force error below 0.06N. On out-of-domain contact prediction benchmarks, EgoTac consistently outperforms the state-of-the-art contact estimator. It also captures the rise and fall patterns of real tactile data and enables zero-shot predictions on unconstrained real-world videos. Scaling analyses further reveal that both data diversity and volume improve performance steadily. Overall, EgoTac provides a scalable pathway to extract tactile priors from egocentric human videos, enabling broadly applicable tactile-aware robot learning. Related resources can be found at https://egotac.github.io/.

1 Introduction

Dexterity is central to physical intelligence and robot learning [20, 2], while touch is essential to dexterity. Everyday skills such as wiping a vase, opening a door, or inserting a socket require robots not only to perceive the scene, but also to continuously regulate physical contact: where to touch, how much pressure to apply, and when to release as the interaction evolves. However, learning such tactile-aware behavior requires large amounts of interaction data that capture rich manipulation patterns. Since collecting this data directly on robots remains difficult and costly, researchers increasingly turn to human data as a scalable source of manipulation experience [48, 33, 23].

Large-scale egocentric human datasets [18, 44, 30, 10] capture diverse behaviors and offer a promising substrate for learning physical skills, but they typically stop at pixels, hand poses, and object trajectories. The tactile variables are absent largely due to the difficulty of capturing tactile signals at scale. Unlike vision, which can be recorded passively and remotely, touch must be measured at the physical contact interface, making data collection sensitive to sensor placement, calibration, spatial coverage, durability, and wearability.

This difficulty motivates a complementary approach: instead of collecting touch everywhere, can we recover tactile information from the human videos that already exist? If dense hand tactile states can

*Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

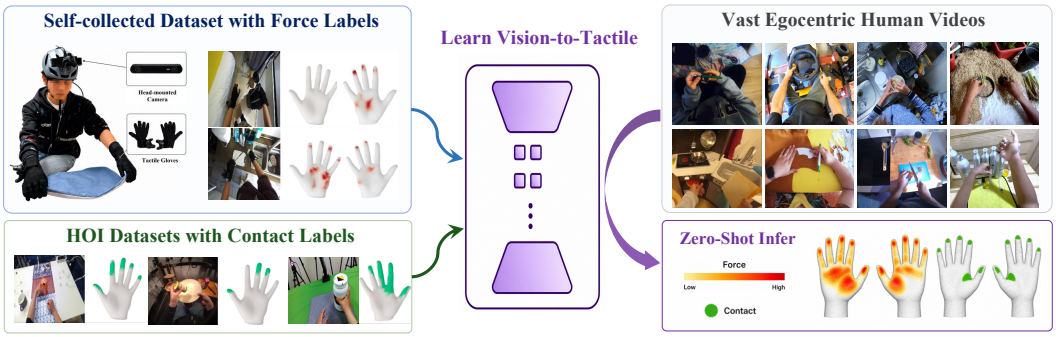

Figure 1: **Overview of EgoTac.** EgoTac unifies self-collected tactile data with contact-labeled egocentric hand-object interaction datasets into a shared MANO-aligned representation, and learns a vision-to-tactile model that predicts dense force and contact from egocentric RGB videos. The learned mapping enables zero-shot tactile labeling of vast in-the-wild human video corpora, providing scalable physical supervision for robot learning and beyond.

be inferred from in-the-wild egocentric videos, existing visual datasets of human manipulation can be augmented with contact and force-like physical supervision without requiring tactile sensors at collection time. In this paper, we study this novel **vision-to-tactile prediction problem**, aiming to turn large-scale human video corpora into a scalable source of tactile supervision.

To address this, we introduce ***EgoTac*, the first generalizable model for vision-to-dense tactile prediction from in-the-wild egocentric human videos.** To support unified training and evaluation, we first construct *EgoTac-Dataset*, a large-scale dataset that contains over **5.7 million** image-tactile pairs. It combines our self-collected egocentric dataset *EgoTac-SC* and eight prior hand-object interaction datasets, covering both single-hand and bimanual interactions across a wide range of real-world tasks and scenes. To fully use these heterogeneous sources with differing spatial coverage, we map all annotations into a shared MANO [38] hand representation [45].

Building on this corpus, we train *EgoTac* as a unified and generalizable vision-to-touch model. *EgoTac* follows a concise encoder–decoder architecture and jointly predicts dense continuous tactile values and contact classification labels from temporal vision input. With enough data scale for training, *EgoTac* learns transferable vision-to-touch knowledge across diverse tasks, scenes, and annotation types. An overview of the framework is shown in Figure 1.

Experiments demonstrate EgoTac's effectiveness along three complementary axes. First, in-domain evaluation shows accurate force prediction ($\text{MAE}_{\text{act}} < 0.06$ N) and strong contact estimation (F1 $>$ 0.70). Second, **out-of-domain and zero-shot transfer** experiments highlight its robust generalization. EgoTac consistently outperforms state-of-the-art methods on unseen contact benchmarks (OAKINK2 [44], FPHA [12]) and preserves temporal force dynamics on OpenTouch [39]. Furthermore, it enables **plausible zero-shot tactile predictions on unconstrained datasets**, including EgoDex [18], EPIC-KITCHENS [5], and Ego4D [16]. Third, data scaling analysis confirms consistent performance improvements as the training corpus grows. Collectively, these findings show that EgoTac delivers accurate in-domain tactile predictions and transferable contact and tactile-dynamics estimates across diverse out-of-domain settings.

Our contributions can be summarized as follows:

- **A unified visual-tactile dataset for egocentric hand-object interaction.** We introduce a large-scale MANO-aligned dataset that merges our newly collected EgoTac-SC dataset with eight prior hand-object interaction datasets, yielding over 5.7 M samples with real force or dense contact annotations.
- **The first in-the-wild tactile prediction model from egocentric vision.** We propose EgoTac, the first model to jointly predict dense tactile force and contact over both hands from egocentric RGB videos. EgoTac enables effective training from heterogeneous data sources and exhibits clear data-scaling benefits across diverse real-world scenarios.
- **Comprehensive empirical evaluation of tactile prediction and generalization.** We evaluate EgoTac on in-domain prediction, out-of-domain benchmarks, and zero-shot transfer

Table 1: **Comparison of egocentric hand-object interaction datasets.** We summarize the key properties of existing representative egocentric datasets and our newly collected *EgoTac-SC* dataset.

Dataset	# Frames	# Clips	Bimanual	Hand Pose	Contact	Real Force	# Envs	In-the-wild
Ego4D(HOI) [16]	21M	89K	✓	×	×	×	900	✓
EPIC-KITCHENS [5]	11.5M	432	✓	×	×	×	32	✓
EgoPAT3D [27]	1.1M	150	×	Estimation	×	×	15	✓
HOI4D [31]	2.4M	4K	×	Estimation	×	×	610	✓
FPHA [12]	105K	1.2K	✓	MoCap	×	×	3	×
TACO [30]	5.2M	2.5K	✓	MoCap	×	×	1	×
HOT3D [1]	1.5M	425	✓	MoCap	×	×	4	×
HO-Cap [41]	656K	64	✓	Estimation	×	×	1	×
OakInk2 [44]	4M	627	✓	MoCap	×	×	1	×
EgoDex [18]	90M	338K	✓	Estimation	×	×	1	×
H2O [24]	571K	164	✓	Estimation	Analytical	×	3	×
ARCTIC [10]	2.1M	339	✓	MoCap	Analytical	×	1	×
ActionSense [6]	1.4M	574	✓	Glove	Pressure	✓	1	×
PressureVisionDB [14]	3.5M	12K	×	×	Pressure	✓	1	×
EgoPressure [47]	4.3M	1.3K	×	Estimation	Pressure	✓	1	×
OpenTouch [39]	550K	2.9K	×	Glove	Pressure	✓	14	✓
FEEL [7]	3M	-	✓	Estimation	Pressure	✓	1	✓
EgoTac-SC(Ours)	3.9M	857	✓	Glove	Pressure	✓	24	✓

to in-the-wild egocentric scenes. Results demonstrate consistent improvements over prior work and robust tactile prediction under diverse scenes.

2 Related Work

2.1 Egocentric Datasets for Hand-Object Interaction

Egocentric datasets are essential for computer vision [29, 25] and robot learning [22, 42, 23, 48]. Early foundational collections primarily focused on visual observations [16, 5], enabling large-scale understanding of daily activities but lacking fine-grained physical interaction annotations required for dexterous manipulation. To address this, subsequent multimodal datasets [27, 31, 30, 12, 1, 44, 41, 18, 10, 24] introduced 3D hand and object poses. However, these datasets typically rely on analytical contact labels inferred computationally from mesh proximity rather than genuine physical contact. More recently, efforts have shifted towards incorporating direct tactile signals [47, 39, 7]. Yet, these resources often face significant limitations: they are typically constrained to specific interaction surfaces [14, 47], limited in scale [39], or lacking real-world scene diversity [6, 7]. To overcome these critical bottlenecks, we introduce a large-scale, unified visual-tactile dataset yielding over 5.7 M samples. By aggregating 8 well-established egocentric datasets featuring analytical mesh-based contacts with our newly collected RGB-force dataset *EgoTac-SC*, we provide a standardized MANO-aligned format that bridges the gap between simulated contacts and real-world high-resolution force signals. We summarize and compare these representative egocentric datasets in Table 1.

2.2 Tactile Prediction from Vision

Vision-based tactile prediction has steadily progressed from localized texture recognition to the inference of full-hand force distributions. Early investigations largely relied on optical-tactile sensors like GelSight [43] to build cross-modal mappings from RGB images [28, 11]. However, these approaches primarily captured geometric and textural cues, neglecting the complex force dynamics inherent in dexterous manipulation. On the other hand, contact-centric methods [3, 15, 21] successfully employ thermal imaging or mesh-distance heuristics to localize interaction areas, but they fail to output the continuous force values critical for precision tasks. To bypass the severe scarcity of dense ground-truth tactile labels, simulation-based methods [8, 17] often supervise tactile inference with physics engines, but they suffer a persistent sim-to-real gap. While recent efforts have attempted real-world force prediction, they are predominantly confined to controlled laboratory setups [47, 26, 14, 13] or specifically tailored to localized regions such as fingertips [35, 9, 4, 36]. By leveraging our proposed unified dataset, this work takes a substantial step towards achieving robust, full-hand tactile prediction from egocentric vision across unconstrained daily environments.

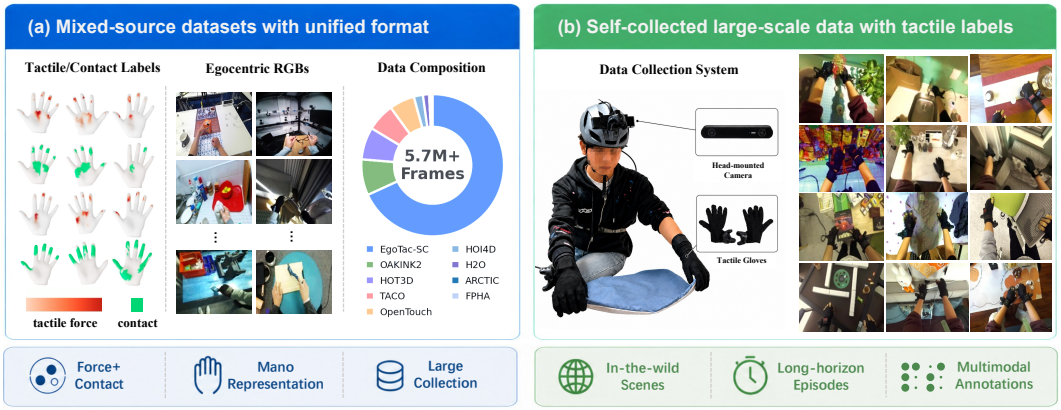

Figure 2: **Dataset overview.** (a) We aggregate mixed-source hand-object interaction datasets containing real tactile measurements or mesh-based contact labels, all unified under a shared MANO topology. (b) The largest portion of our corpus is the self-collected *EgoTac-SC* dataset, which utilizes a custom tactile sensing hardware setup to capture diverse long-horizon manipulation data.

3 Unified Dataset for Visual-Tactile Learning

3.1 EgoTac Dataset Overview

We aggregate 9 datasets into a unified visual-tactile dataset with more than 5.7M samples: our self-collected EgoTac-SC dataset and 8 established hand-object interaction datasets. Figure 2(a) summarizes the dataset composition and annotation types. We next detail the unified format, custom collection process, and dataset integration.

3.2 Unified Data Format

We use a unified annotation format for all datasets. Let t denote the number of time steps and let $V = 778$ be the number of vertices of one MANO hand [38]. For two-hand interactions, the total number of vertices is $2V$. For each sequence, we represent annotations as

$$\mathbf{F} \in \mathbb{R}^{t\times(2V)\times 1}, \qquad \mathbf{C} \in \{0,1\}^{t\times(2V)\times 1}. \tag{1}$$

Here, $\mathbf{F}$ stores continuous tactile intensity values, and $\mathbf{C}$ stores binary contact labels (0 for non-contact, 1 for contact). This representation enables seamless joint training with tactile-rich and contact-only data sources.

3.3 Custom Tactile Data Collection

As a core component of the unified dataset, we collect a new tactile dataset, *EgoTac-SC*, that captures high-resolution force signals during natural daily interactions. Participants wear custom flexible tactile gloves with 264 force sensors together with a commodity hand-tracking glove for 3D hand pose capture. Visual context is recorded by a head-mounted egocentric camera at 30 FPS (hardware setup shown in Figure 2(b)). Participants perform more than 70 contact-rich manipulation tasks (e.g., folding clothes, wiping tables, and opening doors) in 24 diverse environments, including kitchens, offices and more. During post-processing, raw tactile readings are filtered and mapped to MANO vertices to match the unified format in Sec. 3.2. Additional details are shown in Appendix B.

3.4 Integration of Existing HOI Datasets

To further improve scale and coverage, we process and integrate existing egocentric HOI datasets. These include OpenTouch [39], which provides real tactile measurements but mainly single-hand interactions, and contact-only datasets including HOT3D [1], TACO [30], HOI4D [31], H2O [24], ARCTIC [10], OAKINK2 [44], and FPHA [12]. For these contact-only datasets with hand and object meshes, we derive vertex-level contact labels following [21]: a MANO vertex is marked as contact when its minimum distance to the object mesh is below a threshold. More processing details are provided in Appendix C.

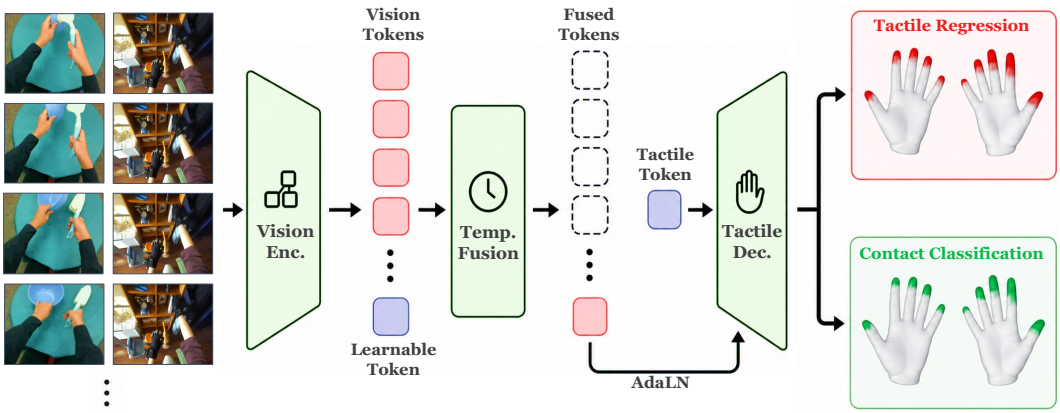

Figure 3: **Model architecture of EgoTac.** Given a short history of egocentric RGB frames, a pretrained vision encoder computes spatial-semantic tokens, and a temporal fusion module aggregates global context over time. An AdaLN-based tactile decoder then jointly conducts dense force regression and contact classification on the shared MANO hand topology.

4 Vision-to-Tactile Prediction Framework

Predicting dense tactile states from unconstrained egocentric vision is fundamentally challenging. Prior vision-based methods were predominantly confined to controlled laboratory setups [47], focused only on localized regions (e.g., fingertips) [35], or relied purely on analytical mesh-distance heuristics [21]. To address these bottlenecks, we introduce a unified prediction framework that leverages the large-scale heterogeneous dataset described in Sec. 3. Our key design is a joint force-contact training scheme that cohesively aggregates continuous force measurements with discrete contact annotations into a shared MANO representation. We detail our problem formulation in Sec. 4.1, model architecture in Sec. 4.2, and training strategy in Sec. 4.3.

4.1 Problem Formulation

We aim to infer dense tactile states of both hands from egocentric RGB observations. Given an egocentric RGB clip $\mathbf{I}_{t-T+1:t} \in \mathbb{R}^{T\times 3\times H\times W}$ with an observation horizon T, our model concurrently predicts continuous tactile force fields $\hat{\mathbf{F}}$ and discrete contact states $\hat{\mathbf{C}}$ over a prediction horizon K, all anchored on the MANO hand mesh [38]:

$$(\hat{\mathbf{F}}_{t:t+K-1}, \hat{\mathbf{C}}_{t:t+K-1}) = f_\theta(\mathbf{I}_{t-T+1:t}), \qquad \hat{\mathbf{F}}, \hat{\mathbf{C}} \in \mathbb{R}^{K\times 2V}, \tag{2}$$

where $V = 778$ designates the number of vertices on the MANO mesh, and the factor of two accounts for both the left and right hands. Specifically, $\hat{\mathbf{F}}$ denotes the scalar normal force at each vertex, while $\hat{\mathbf{C}}$ represents the corresponding per-vertex contact probability logits.

4.2 EgoTac Model Architecture

The architecture of EgoTac is illustrated in Figure 3. Our model consists of a vision encoder, a temporal fusion module, and an AdaLN-based tactile decoder. Each egocentric RGB frame within the observation horizon is encoded by a pretrained DINOv2 ViT [34] into spatial-semantic vision tokens. To aggregate information over time, we use a learnable temporal token that attends to the visual sequence via cross-attention, producing a global context vector $z \in \mathbb{R}^d$.

Conditioned on z, the tactile decoder performs direct prediction in a single forward pass. It maintains a set of learnable tactile query tokens $q \in \mathbb{R}^{K\times d}$, where each token corresponds to tactile estimations at one prediction step. These tokens are iteratively updated through Transformer layers modulated by Adaptive Layer Normalization (AdaLN):

$$\mathrm{AdaLN}_\ell(h; z) = (1 + \gamma_\ell(z)) \odot h + \beta_\ell(z), \tag{3}$$

where $h \in \mathbb{R}^{K\times d}$ denotes the token features at layer ℓ, and the scale $\gamma_\ell(z)$ and shift $\beta_\ell(z)$ are dynamically generated from the context vector z and broadcast over the token dimension. This conditioning mechanism preserves computational efficiency while enabling the temporal visual context to guide predictions over the entire MANO topology.



Figure 4: **Attention maps from EgoTac.** Across casual in-the-wild videos, EgoTac emphasizes hand-object interaction regions that are informative for tactile prediction.

4.3 Force-Contact Joint Training

Dense tactile inference requires predicting both where interaction occurs and how strong it is. We therefore attach two task-specific heads to the shared decoder: (1) a tactile regression head for continuous force values, and (2) a contact classification head for per-vertex contact logits. This dual formulation enables mixed-dataset learning: force-supervised samples provide both targets, while contact-only samples provide contact targets without force magnitudes.

Specifically, our training set mixes the force-supervised egocentric tactile source with geometry-derived contact-only datasets. For force-supervised samples, contact labels can be derived by thresholding the continuous force; for contact-only samples, they come from mesh-based collision proximity. For each sample, the dataloader provides a force-valid mask $\mathbf{M}^F$ and a contact-valid mask $\mathbf{M}^C$ for coherent supervision even with missing annotations. The total objective is defined as

$$\mathcal{L} = \lambda_F \mathcal{L}_F + \lambda_C \mathcal{L}_C + \lambda_A \mathcal{L}_A + \lambda_{\text{cons}} \mathcal{L}_{\text{cons}}. \tag{4}$$

Here, $\mathcal{L}_F$ is an MSE loss on normalized tactile values, applied only where continuous force supervision is available. The contact loss $\mathcal{L}_C$ is a BCE loss over contact-valid vertices with positive class reweighting. To avoid trivial all-zero predictions, we apply an active-region loss $\mathcal{L}_A$ as a Smooth-L_1 penalty on vertices with active ground-truth force. Finally, the consistency loss $\mathcal{L}_{\text{cons}}$ aligns contact probabilities derived from the force branch with detached contact-head predictions via BCE. Formal definitions are provided in Appendix D.3.

For contact-only samples, $\mathbf{M}^F = 0$, so neither $\mathcal{L}_F$ nor $\mathcal{L}_A$ contributes. Nevertheless, $\mathcal{L}_C$ updates the shared encoder and decoder, while $\mathcal{L}_{\text{cons}}$ constrains where and when the force branch is active. Contact thus supplies a structural constraint, while EgoTac-SC provides force magnitudes. We examine these complementary roles in Section 5.5.

4.4 Implementation Details

We optimize the full model end-to-end with AdamW [32] on 4 A10 GPUs, using a total batch size of 128, a cosine-decayed learning rate schedule from 1e-4 to 1e-5 and 5% warm-up of the total training steps. Unless otherwise noted, the vision encoder is fine-tuned jointly with the decoder using a reduced learning-rate multiplier of 0.1, and we maintain an EMA model for evaluation. Additional details, including decoder depth and width, dataset sampling, loss weights, and augmentation settings, are provided in the Appendix D.4.

5 Experiments

5.1 Experimental Overview

We evaluate EgoTac through four research questions:

- **In-domain performance and ablations:** On seen domains, how accurately does EgoTac predict continuous force and binary contact, and how much do key design choices contribute? (Section 5.2)
- **Out-of-domain generalization:** Under domain shift, how does EgoTac compare with prior SOTA contact estimators, and does it preserve force dynamics on real tactile data? Further, is it possible to zero-shot infer tactile labels on diverse HOI scenes? (Section 5.3)

Table 2: **In-domain tactile evaluation and ablation results.** Averaged metrics on the EgoTac-SC test set are reported. Ablations validate effects of the temporal visual context, the active-region loss, and explicit force supervision.

Method	MAE ↓	MAE_{act} ↓	Temporal Pearson ↑
EgoTac	**0.004**	**0.056**	**0.506**
w/o temporal context	0.005	**0.056**	0.463
w/o active-region loss	**0.004**	0.097	0.482
w/o force supervision	0.322	0.108	0.240

Table 3: **In-domain contact evaluation.** Results are averaged across timestamps on all test splits of each component dataset. EgoTac maintains strong contact localization accuracy despite the additional force regression objectives.

Benchmark	Method	AUROC ↑	IoU ↑	Precision ↑	Recall ↑	F1 ↑
HOT3D [1]	EgoTac-i	**0.976**	0.587	0.656	**0.771**	0.702
	EgoTac	**0.976**	**0.593**	**0.667**	0.767	**0.706**
H2O [24]	EgoTac-i	**0.989**	0.647	0.725	0.846	0.777
	EgoTac	**0.989**	**0.743**	**0.804**	**0.904**	**0.850**
ARCTIC [10]	EgoTac-i	0.972	**0.676**	0.716	**0.905**	0.790
	EgoTac	**0.977**	0.669	**0.732**	0.897	**0.795**

- **Data scaling effects:** How do training data scale and diversity affect EgoTac's prediction performance? (Section 5.4)
- **Supervision sources:** How do measured-force and contact-only data contribute to contact and force prediction? (Section 5.5)

Datasets. Training uses EgoTac-SC together with HOI4D [31], H2O [24], HOT3D [1], ARCTIC [10] and TACO [30]. For in-domain evaluation, each dataset is split 9:1, and we report results on the held-out 10%. For out-of-domain (OOD) evaluation, we use OpenTouch [39], OAKINK2 [44], and FPHA [12] without further splitting.

Baselines. For in-domain evaluation, we mainly compare EgoTac with EgoTac-i, an image-based variant that removes temporal context. For out-of-domain contact benchmarks, we compare against three representative RGB-based 3D contact estimators: BSTRO [19], DECO [40], and HACO [21]. BSTRO and DECO predict vertex-level 3D contact from RGB observations and are therefore directly comparable to our MANO-aligned contact prediction setting. HACO is the state-of-the-art dense contact estimator, which first uses a hand detector [37] to crop the right hand and then performs contact inference only on the right hand.

Metrics. For in-domain continuous tactile prediction, we report global MAE, active-region MAE (MAE_{act}) and temporal Pearson correlation between predicted and ground-truth force. Unless otherwise stated, active vertices are defined by ground-truth force $> 0.05\,\text{N}$. For contact benchmarks, we report Precision, Recall, F1, IoU, and AUROC from contact logits. For the OOD force-sensitive benchmark, we report relative force dynamics metrics including cosine similarity, rise F1 and fall F1. Additionally, we provide metric details in Appendix E and extensive frequency-domain diagnostics in Appendix F.

5.2 In-Domain Performance and Ablations

In-domain tactile-based evaluation. We first evaluate calibrated force prediction on held-out EgoTac-SC episodes. Because non-contact vertices dominate the MANO surface, global averages can underestimate errors on physically meaningful regions; therefore, we prioritize MAE_{act} and use global MAE and temporal correlation as complementary signals. Table 2 shows three main trends. First, incorporating temporal visual context improves temporal consistency. Second, adding the active-region loss enhances force fidelity in interacting areas. Third, relying on contact-only

Table 4: **Out-of-domain contact evaluation.** Results are averaged across timestamps on OAKINK2 [44] and FPHA [12]. EgoTac yields superior performance compared with previous methods.

Benchmark	Method	AUROC ↑	IoU ↑	Precision ↑	Recall ↑	F1 ↑
Oakink2 [44]	BSTRO [19]	0.537	0.076	0.172	0.139	0.134
	DECO [40]	0.655	0.144	0.291	0.329	0.243
	HACO [21]	0.772	0.259	0.356	**0.568**	0.400
	EgoTac	**0.844**	**0.285**	**0.498**	0.441	**0.435**
FPHA [12]	BSTRO [19]	0.535	0.017	0.080	0.033	0.032
	DECO [40]	0.635	0.041	0.161	0.092	0.075
	HACO [21]	0.699	0.160	0.187	**0.716**	0.268
	EgoTac	**0.758**	**0.178**	**0.304**	0.401	**0.290**

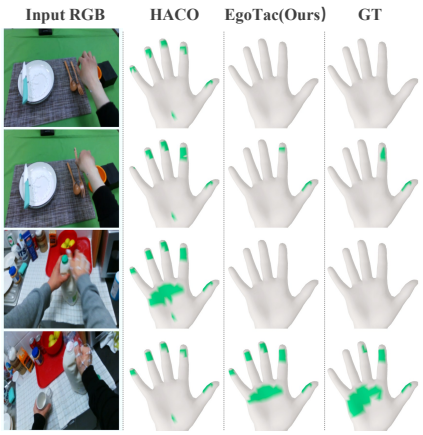

Figure 5: **Qualitative comparison of OOD contact estimation (right-hand).** EgoTac produces cleaner contact status separation than HACO [21] on OAKINK2 [44] and FPHA [12].

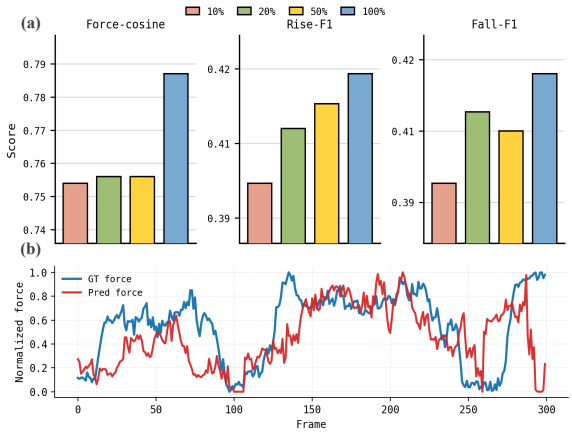

Figure 6: **Data scaling and OOD force prediction.** (a) Scaling training data volume improves force dynamics prediction on OpenTouch [39]. (b) Comparison of normalized total force over time, showing EgoTac effectively tracks ground-truth temporal trends.

supervision is insufficient for realistic force estimation, confirming that contact logits do not directly encode the force scale.

In-domain contact-based evaluation. Since EgoTac jointly predicts force and contact on a shared MANO topology, we also evaluate contact quality on in-domain HOI test splits. As shown in Table 3, EgoTac maintains strong contact prediction performance (AUROC > 0.97, F1 > 0.70), indicating that improved force prediction is not achieved by sacrificing contact localization. Together, the force and contact evaluations provide a stricter in-domain validation of both physical intensity prediction and spatial interaction accuracy.

5.3 Out-of-Domain Generalization

OOD contact-based benchmarks. Most public HOI datasets do not provide dense force supervision, so OOD evaluation is primarily contact-centric. We compare EgoTac with three representative RGB-based 3D contact estimators, BSTRO [19], DECO [40], and HACO [21], on OAKINK2 [44] and FPHA [12]. As shown in Table 4, EgoTac achieves the best AUROC, IoU, Precision, and F1 on both benchmarks. HACO achieves higher Recall in some settings; however, this improvement largely stems from over-predicting contact regions, resulting in greater coverage at the cost of more false positives. Figure 5 qualitatively supports this trend: EgoTac yields cleaner separation between in-contact and out-of-contact states and produces more accurate contact maps, whereas HACO tends to predict contact even when no interaction occurs.

OOD force-sensitive benchmark on OpenTouch. To test OOD transfer beyond binary contact, we evaluate force dynamics on OpenTouch [39], which uses a tactile sensing system different from

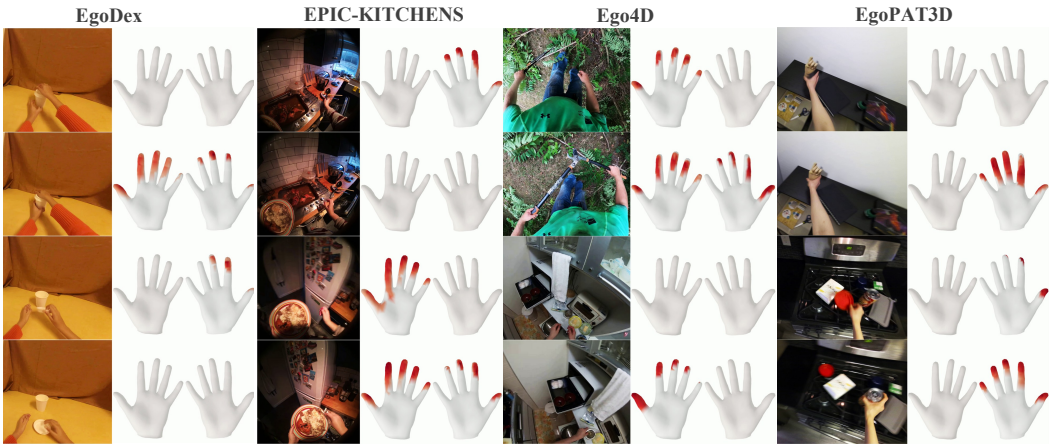

Figure 7: **Qualitative results in diverse, in-the-wild scenes.** We visualize zero-shot tactile prediction results for both hands on clips from multiple egocentric datasets, including EgoDex [18], EPIC-KITCHENS [5], Ego4D [16], and EgoPAT3D [27]. EgoTac produces spatially coherent tactile patterns and smooth temporal transitions.

Table 5: **Data diversity scaling on OOD contact estimation.** We increase the number of data sources used for training and report contact metrics on the OAKINK2 [44] benchmark. Consistent improvements are shown with increasing data diversity.

Data Use	AUROC ↑	IoU ↑	F1 ↑
1 dataset	0.603	0.065	0.120
4 datasets	0.832	0.263	0.407
6 datasets	**0.844**	**0.285**	**0.435**

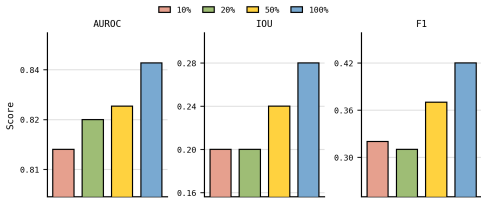

Figure 8: **Data volume scaling on OOD contact estimation.** Scaling training data volume from 10% to 100% improves contact prediction performance on OAKINK2 [44].

EgoTac-SC. The public OpenTouch annotations do not provide the per-taxel physical calibration and effective sensing areas required for a reliable conversion to Newton level forces. Therefore, a direct absolute-force MAE in Newtons would require unsupported cross-sensor calibration assumptions.

Instead, we normalize prediction and ground truth to $[0, 1]$ using calibration-split statistics before evaluation. This protocol measures whether EgoTac preserves relative tactile intensity and temporal structure under sensor and domain shift. In Figure 6(a), EgoTac reaches force-cosine similarity above 0.7 and rise/fall F1 above 0.4. Figure 6(b) further shows that predicted temporal force traces track the ground-truth trend. Overall, these results indicate that EgoTac can reliably predict continuous force dynamics even under out-of-domain conditions.

Qualitative results on diverse HOI scenes. Figure 7 shows qualitative predictions on clips from EgoDex [18], EPIC-KITCHENS [5], Ego4D [16], and EgoPAT3D [27]. Across diverse real-world scenes, EgoTac produces spatially coherent tactile patterns and smooth temporal transitions. In examples from EgoDex and EPIC-KITCHENS, EgoTac not only identifies which hand is actively engaged in manipulation, but also captures fine-grained contact and force changes over long-horizon episodes. In EgoPAT3D examples, EgoTac further adapts its predicted tactile patterns according to the objects being interacted with, indicating sensitivity to object-specific geometry. Additionally, we use EgoTac to infer tactile from casual recorded videos, demonstrating the model's performance on custom data. Figure 4 presents attention maps in these self-recorded scenes, showing that EgoTac consistently focuses on hand-object interaction regions that determine tactile patterns.

5.4 Effects of Data Scaling

In this part, we analyze the effects of data scaling from two perspectives: (1) increasing the number of diverse datasets used in training, and (2) scaling the overall data volume by using different proportions of the full training set. First, Table 5 shows that leveraging more data sources (from 1 to 6 datasets) steadily improves out-of-domain contact estimation on OAKINK2 [44]. Second, scaling the total

Table 6: **Ablation of supervision sources.** We compare training on EgoTac-SC, five contact-only HOI datasets, and their union (Mixed). "–" denotes the absence of a force-supervised model. Tables X4 and X5 report the complete contact and force-dynamics metrics.

	OAKINK2: contact			OpenTouch: force dynamics		
Training data	AUROC ↑	IoU ↑	F1 ↑	Force-Cosine ↑	Rise-F1 ↑	Fall-F1 ↑
EgoTac-SC only	0.60	0.07	0.12	0.745	**0.404**	0.385
Contact-only HOI	**0.84**	**0.29**	**0.44**	–	–	–
Mixed	**0.84**	**0.29**	**0.44**	**0.777**	0.403	**0.403**

data volume from 10% to 100% yields improvements across multiple OOD metrics. Specifically, it enhances contact AUROC, IoU, and F1 on OAKINK2 [44] (Figure 8), while significantly boosting force-cosine similarity and rise/fall F1 scores on the OpenTouch [39] benchmark (Figure 6(a)). These consistent performance gains suggest that EgoTac benefits from both source diversity and data volume, demonstrating the scalability of our formulation and indicating that further improvements can be expected as more tactile-supervised egocentric data becomes available.

5.5 Understanding Heterogeneous Supervision

EgoTac combines two types of supervision with complementary physical meanings: continuous force measurements specify interaction intensity, while contact annotations provide large-scale information about where and when interactions occur. To understand the contribution of each supervision source, we compare three training configurations: (1) EgoTac-SC only, (2) contact-only HOI datasets only, and (3) their combination. The contact-only sources include HOT3D [1], HOI4D [31], TACO [30], H2O [24], and ARCTIC [10]. Results are evaluated under the OOD protocols described in Section 5.3.

Complementary roles of force and contact supervision. The two supervision sources provide distinct and complementary benefits. As shown in Table 6, on OAKINK2, training with contact-only HOI datasets substantially improves contact generalization over EgoTac-SC alone, increasing F1 from 0.12 to 0.44. This demonstrates that diverse bare-hand HOI data provide valuable visual and interaction priors for contact localization. In contrast, EgoTac-SC provides the only continuous force supervision and is therefore essential for learning tactile intensity. On OpenTouch, combining EgoTac-SC with contact-only datasets further improves force cosine from 0.745 to 0.777, indicating that contact supervision can benefit force prediction through shared visual-temporal representations while measured-force data remain necessary for recovering continuous tactile dynamics.

Interaction between contact and force supervision. The two supervision sources are not independent: contact annotations provide a structural prior over interaction occurrence, while force annotations capture the continuous intensity of physical interactions. Through shared representation learning and the force–contact consistency objective, contact supervision can improve the generalization of force prediction even without directly providing force magnitudes. Detailed ablations of this interaction, including the effect of removing $\mathcal{L}_{\text{cons}}$, are provided in the Appendix G.

6 Conclusion

We introduce EgoTac, a generalizable framework tailored for predicting dense tactile states directly from in-the-wild egocentric RGB videos. Our contributions are threefold: (1) we construct a massive 5.7M-sample unified dataset by lifting our newly collected real-force EgoTac-SC dataset and eight existing hand-object interaction datasets onto a shared MANO topology; (2) we propose a scalable architecture with a mixed force-contact objective that effectively learns from this heterogeneous supervision; and (3) we demonstrate through comprehensive evaluations that EgoTac achieves accurate in-domain predictions, strong out-of-domain generalization, and impressive zero-shot scaling capabilities on unconstrained real-world scenes. Ultimately, EgoTac provides a foundational step toward extracting rich physical priors from ubiquitous human video corpora.

Acknowledgements

This research was conducted with the support of the Shanghai Qi Zhi Institute & Spirit AI Innovation Program. Funding and support for this work were also provided by the Tsinghua University - Keystone Electrical (Zhejiang) Co.,Ltd Joint Research Center for Embodied Multimodal Artificial Intelligence (JCEMAI). Additionally, we would like to extend our thanks to the Xiongan AI Institute. This project is also supported by Beijing Natural Science Foundation(JR25019).

References

[1] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7071, 2025.

[2] Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

[3] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. ContactDB: Analyzing and predicting grasp contact via thermal imaging. In *CVPR*, 2019.

[4] Nutan Chen, Göran Westling, Benoni B Edin, and Patrick Van Der Smagt. Estimating fingertip forces, torques, and local curvatures from fingernail images. *Robotica*, 38(7):1242–1262, 2020.

[5] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018.

[6] Joseph DelPreto, Chao Liu, Yiyue Luo, Michael Foshey, Yunzhu Li, Antonio Torralba, Wojciech Matusik, and Daniela Rus. Actionsense: A multimodal dataset and recording framework for human activities using wearable sensors in a kitchen environment. *Advances in Neural Information Processing Systems*, 35:13800–13813, 2022.

[7] 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. *arXiv preprint arXiv:2603.15847*, 2026.

[8] Kiana Ehsani, Shubham Tulsiani, Saurabh Gupta, Ali Farhadi, and Abhinav Gupta. Use the force, luke! learning to predict physical forces by simulating effects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 224–233, 2020.

[9] Navid Fallahinia and Stephen A Mascaro. Real-time tactile grasp force sensing using fingernail imaging via deep neural networks. *IEEE Robotics and Automation Letters*, 7(3):6558–6565, 2022.

[10] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12943–12954, 2023.

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

[12] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018.

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

[14] 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*, pages 328–345. Springer, 2022.

[15] Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp. ContactOpt: Optimizing contact to improve grasps. In *CVPR*, 2021.

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

[17] Ryo Hanai, Yukiyasu Domae, Ixchel G Ramirez-Alpizar, Bruno Leme, and Tetsuya Ogata. Force map: Learning to predict contact force distribution from vision. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3129–3136. IEEE, 2023.

[18] Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025.

[19] 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.

[20] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. pi0.5: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[21] Daniel Sungho Jung and Kyoung Mu Lee. Learning dense hand contact estimation from imbalanced data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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

[23] Simar Kareer, Karl Pertsch, James Darpinian, Judy Hoffman, Danfei Xu, Sergey Levine, Chelsea Finn, and Suraj Nair. Emergence of human to robot transfer in vision-language-action models. *arXiv preprint arXiv:2512.22414*, 2025.

[24] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10138–10148, 2021.

[25] Dayou Li, Lulin Liu, Bangya Liu, Shijie Zhou, Jiu Feng, Ziqi Lu, Minghui Zheng, Chenyu You, and Zhiwen Fan. Egocentric world model for photorealistic hand-object interaction synthesis. *arXiv preprint arXiv:2603.13615*, 2026.

[26] Yichen Li, Yilun Du, Chao Liu, Francis Williams, Michael Foshey, Benjamin Eckart, Jan Kautz, Joshua B Tenenbaum, Antonio Torralba, and Wojciech Matusik. Learning to jointly understand visual and tactile signals. In *The Twelfth International Conference on Learning Representations*, 2023.

[27] Yiming Li, Ziang Cao, Andrew Liang, Benjamin Liang, Luoyao Chen, Hang Zhao, and Chen Feng. Egocentric prediction of action target in 3d. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20971–20980. IEEE, 2022.

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

[29] 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*, 2025.

[30] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21740–21751, 2024.

[31] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022.

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[33] Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. Being-h0: vision-language-action pretraining from large-scale human videos. *arXiv preprint arXiv:2507.15597*, 2025.

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

[35] Tu-Hoa Pham, Abderrahmane Kheddar, Ammar Qammaz, and Antonis A Argyros. Towards force sensing from vision: Observing hand-object interactions to infer manipulation forces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2810–2819, 2015.

[36] Tu-Hoa Pham, Nikolaos Kyriazis, Antonis A Argyros, and Abderrahmane Kheddar. Hand-object contact force estimation from markerless visual tracking. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2883–2896, 2017.

[37] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12242–12254, 2025.

[38] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM TOG*, 2017.

[39] Yuxin Ray Song, Jinzhou Li, Rao Fu, Devin Murphy, Kaichen Zhou, Rishi Shiv, Yaqi Li, Haoyu Xiong, Crystal Elaine Owens, Yilun Du, et al. Opentouch: Bringing full-hand touch to real-world interaction. *arXiv preprint arXiv:2512.16842*, 2025.

[40] 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.

[41] 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. *arXiv preprint arXiv:2406.06843*, 2024.

[42] Chengbo Yuan, Rui Zhou, Mengzhen Liu, Yingdong Hu, Shengjie Wang, Li Yi, Chuan Wen, Shanghang Zhang, and Yang Gao. Motiontrans: Human vr data enable motion-level learning for robotic manipulation policies. *arXiv preprint arXiv:2509.17759*, 2025.

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

[44] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 445–456, 2024.

[45] Chi Zhang, Penglin Cai, Haoqi Yuan, Chaoyi Xu, and Zongqing Lu. Unitachand: Unified spatio-tactile representation for human to robotic hand skill transfer. *arXiv preprint arXiv:2512.21233*, 2025.

[46] Xinyu Zhang, Ziyi Kou, Chuan Qin, Mia Huang, Ergys Ristani, Ankit Kumar, Lele Chen, Kun He, Abdeslam Boularias, and Li Guan. Glove2hand: Synthesizing natural hand-object interaction from multi-modal sensing gloves. *arXiv preprint arXiv:2603.20850*, 2026.

[47] Yiming Zhao, Taein Kwon, Paul Streli, Marc Pollefeys, and Christian Holz. Egopressure: A dataset for hand pressure and pose estimation in egocentric vision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27727–27738, 2025.

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

A Appendix Overview

This supplementary material provides additional details on dataset construction, model implementation, and evaluation for *EgoTac*. It is organized as follows:

- B. Details of the self-collected dataset (EgoTac-SC)
- C. Details of other datasets
- D. Details of EgoTac implementation
- E. Details of metrics
- F. More experimental results
- G. More discussions
- H. Limitations and societal impacts

B Details of the self-collected dataset (EgoTac-SC)

This section provides additional details on the construction, calibration, post-processing, and organization of our self-collected *EgoTac-SC* dataset. EgoTac-SC is collected with a wearable multimodal platform that synchronizes egocentric visual observations, hand-pose tracking, and full-hand tactile measurements. We convert the raw streams into the unified MANO-aligned representation used throughout the paper, enabling joint training with both force-supervised and contact-only datasets.

B.1 Dataset collection

Hardware. Figure X1 illustrates the capture hardware used for our data collection. The platform contains three synchronized sensing components: an egocentric RGB-D camera, wearable motion-tactile gloves, and wrist-pose tracking devices. The visual stream is captured by a head-mounted ZED 2i camera at 1080p resolution, with optional point-cloud export. For hand tracking and tactile sensing, each hand is equipped with an HTC VIVE Tracker and an in-house motion-tactile glove. Each glove includes 11 high-precision nine-axis inertial modules and a dense force-sensing matrix with more than 120 sensing points. The inertial modules and VIVE Tracker provide hand-pose-related signals (20 hand joint poses plus one wrist pose), while the force matrix records distributed tactile measurements over fingers and palm. Together, these sensors capture synchronized egocentric vision, hand motion, and full-hand tactile interaction during natural manipulation.



Figure X1: **Data-collection hardware.** Our capture platform includes three synchronized components: an egocentric RGB-D camera, wearable motion-tactile gloves, and wrist-pose trackers.

Collection procedure. Before recording, the head-mounted camera and wearable hand devices are connected to the acquisition terminal. The collection process follows a fixed protocol: we first launch the transmission and acquisition software, then perform pose calibration for the wearable trackers. After calibration, participants perform assigned manipulation tasks while the system records synchronized visual, pose, and tactile streams. Each sequence is terminated through the acquisition interface, after which the raw streams are saved for offline synchronization and conversion.

Raw tactile channel organization. Each hand contains a dense force-sensing matrix. In the raw byte layout, finger and palm force channels are interleaved with several bend-related channels. We exclude bend channels from supervision and retain only force channels. Specifically, each hand uses 60 finger-force channels and 72 palm-force channels, resulting in a 132-dimensional tactile vector $s \in \mathbb{R}^{132}$. The five bend-related channels are not used for tactile-force annotation.

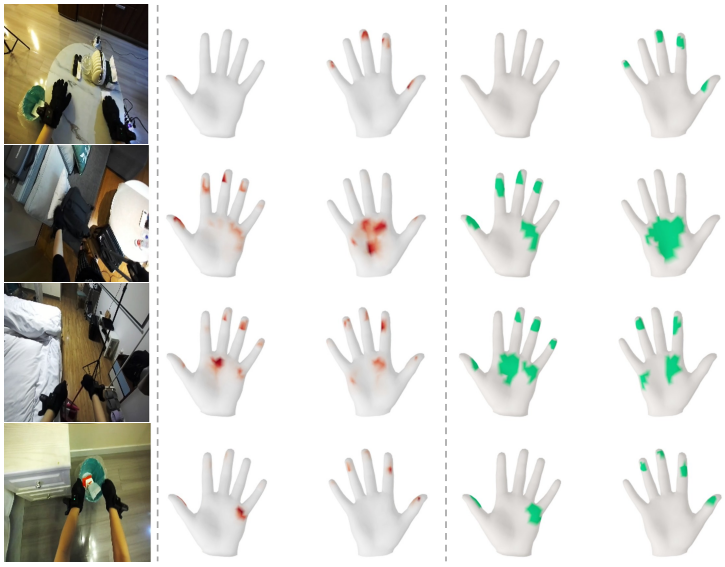

Figure X2: **Representative samples from EgoTac-SC.** Each example shows egocentric RGB together with MANO-aligned tactile-force maps and binary contact labels.

B.2 Tactile post-processing and MANO hand mapping

Temporal alignment and interpolation. Raw streams are stored with individual timestamps. We parse wrist, hand-pose, and tactile streams and resample them to a unified timeline. For pose signals involving rigid transformations or rotations, we use interpolation suitable for SE(3)-type signals. For tactile force readings, we use linear interpolation. This yields frame-level temporal alignment across visual, pose, and tactile observations.

RGB decoding and synchronization. Egocentric video frames are decoded from the recorded camera stream and synchronized to the unified timeline. The aligned RGB frames are then resized and cropped to the training resolution.

Coordinate normalization. Because the camera, wrist trackers, and glove sensors use different coordinate conventions, we apply a fixed geometric transformation chain. This step normalizes camera, tracker, and hand-pose signals into a consistent convention before mapping tactile values to the MANO surface.

Tactile-to-MANO mapping. After temporal alignment and coordinate normalization, we map the per-hand tactile vector to the MANO mesh. Each MANO hand contains $V = 778$ vertices. Following [45], we distribute sparse glove-force readings to dense MANO vertices through predefined region patches and bilinear sensor weights.

Let $s \in \mathbb{R}^{132}$ denote the force vector of one hand at a single frame. For each MANO vertex v, let $\mathcal{W}(v)$ be the set of neighboring tactile sensors used to interpolate the force value at that vertex. The mapped force value is computed as

$$F_v = \sum_{k \in \mathcal{W}(v)} w_{v,k}\, s_k, \qquad \sum_{k \in \mathcal{W}(v)} w_{v,k} = 1, \tag{5}$$

where $w_{v,k}$ is the interpolation weight between MANO vertex v and tactile sensor k. This produces a dense tactile field $F \in \mathbb{R}^V$ for each hand. Binary contact labels are then obtained by thresholding the mapped tactile force:

$$C_v = \mathbb{1}[F_v > \tau_c], \tag{6}$$

where τ_c is set to 0.05 N during preprocessing. After mapping both hands, each frame contains dense tactile and contact annotations over $2V = 1556$ MANO vertices:

$$F \in \mathbb{R}^{2V}, \qquad C \in \{0,1\}^{2V}. \tag{7}$$



Figure X3: **Examples from eight auxiliary HOI datasets.** Contact or tactile annotations are converted to a shared MANO topology to enable joint training across datasets.

This representation is consistent with the unified data format described in Sec. 3.2 and allows EgoTac-SC to be trained jointly with contact-only datasets.

B.3 Visualization of samples

Representative EgoTac-SC samples are shown in Figure X2. They illustrate diverse contact locations and force magnitudes under natural manipulation. Force annotations are visualized as continuous heatmaps on the MANO surface, while contact annotations are shown as binary active regions obtained by thresholding tactile force.

C Details of other datasets

This section describes how auxiliary datasets are normalized into a unified supervision protocol so that heterogeneous sources can be trained jointly without changing the model interface.

C.1 MANO hand contact label deriving

For datasets without native dense force labels (HOT3D [1], TACO [30], HOI4D [31], H2O [24], ARCTIC [10], OAKINK2 [44], and FPHA [12]), we derive MANO-vertex contact labels from

geometric proximity between reconstructed hand and object meshes. For frame t, vertex v, and object vertex set $\mathcal{O}_t$, we use:

$$C_{t,v} = \mathbb{1}\left[\min_{u \in \mathcal{O}_t} \left\| \mathbf{x}_{t,v}^{\text{hand}} - \mathbf{x}_{t,u}^{\text{obj}} \right\|_2 \leq \delta\right], \tag{8}$$

where δ is a fixed contact threshold set to 1 cm following [21]. This produces binary contact targets on the same MANO topology as our force supervision.

Unified conversion protocol. Although source datasets differ in annotation style and storage format, we apply a consistent conversion pipeline:

- Decode egocentric RGB frames and timestamps.
- Reconstruct hand meshes in a canonical MANO space.
- Reconstruct or load object geometry and align it to frame coordinates.
- Compute contact labels using nearest-distance thresholding.
- Export a standardized episode-level format with RGB, MANO contact fields, timestamps, and episode boundaries.

C.2 OpenTouch dataset process

For OpenTouch [39], which provides real tactile measurements with a different sensor layout, we follow the official processing instructions to convert raw tactile grids into MANO-aligned tactile vectors. The conversion includes force normalization, spatial alignment from sensor coordinates to MANO vertices, and temporal synchronization with egocentric RGB.

C.3 Visualization of samples

Examples from external datasets are shown in Figure X3, highlighting diversity in interaction type, contact sparsity, hand-viewpoint shifts, and scene composition.

D Details of EgoTac implementation

This section summarizes the implementation choices used in our main experiments, including data processing, optimization, and objective design.

D.1 Data augmentation

We apply temporally consistent RGB augmentation within each observation window:

- Color jitter and resized random crop.
- View-scale augmentation with scale range $[0.9, 1.1]$, applied with probability 0.8.
- Left-right mirror augmentation with probability 0.5, applied to the selected asymmetric data source HOI4D [31].

D.2 Dataset sampling

We use mixed-source temperature sampling to avoid source starvation during multi-dataset training. For source i with pool size n_i, the sampling weight is

$$w_i = n_i^{\tau}, \qquad p_i = \frac{w_i}{\sum_j w_j}, \tag{9}$$

where $\tau = 0.5$ in our default setup. We additionally enforce per-source minimum quotas to prevent under-training of small datasets. Let E denote epoch size (default: full active training pool size). The initial target sample count of source i is

$$\tilde{m}_i = p_i E. \tag{10}$$

Table X1: Main training hyperparameters used in the primary setting.

Item	Value
Observation horizon T	4
Tactile horizon K	1
Input RGB size	224×224
Output dimension	1556 (two-hand MANO vertices)
Batch size (per step)	32×4
Gradient accumulation	1
Optimizer	AdamW
Learning rate	1×10^{-4}
Final learning rate	1×10^{-5}
Weight decay	1×10^{-5}
LR scheduler	cosine
Warm-up ratio	0.05
Max grad norm	5.0
Mixed precision	bf16
Decoder layers / heads / width	12 / 12 / 768
Vision encoder	DINOv2-Base [34]
Temperature sampling τ	0.5
Force loss weight λ_F	1.0
Contact loss weight λ_C	1.0
Active-region loss weight λ_A	0.2
Consistency loss weight λ_{cons}	0.05

With minimum quota $q_i \in [0, 1]$, the lower-bound count is

$$m_i^{\min} = \lfloor q_i E \rfloor, \tag{11}$$

and the final per-source target m_i satisfies

$$m_i \geq m_i^{\min}, \qquad \sum_i m_i = E. \tag{12}$$

In practice, we use conservative non-zero quotas for low-resource sources to stabilize optimization and preserve source diversity.

D.3 Training objective

Let $\hat{\mathbf{F}}, \mathbf{F} \in \mathbb{R}^{K \times 2V}$ denote predicted and target force, and $\hat{\mathbf{Z}}, \mathbf{C} \in \mathbb{R}^{K \times 2V}$ denote predicted contact logits and target contact labels. Let $\mathbf{M}^F, \mathbf{M}^C \in \{0, 1\}^{K \times 2V}$ be force/contact valid masks. We define masked mean as

$$\langle \mathbf{A} \rangle_{\mathbf{M}} = \frac{\sum_{t,v} A_{t,v} M_{t,v}}{\sum_{t,v} M_{t,v} + \epsilon}. \tag{13}$$

Force loss (MSE) is

$$\mathcal{L}_F = \left\langle (\hat{\mathbf{F}} - \mathbf{F})^2 \right\rangle_{\mathbf{M}^F}. \tag{14}$$

Active-region loss uses active mask $\mathbf{M}^A$:

$$\mathbf{M}^A_{t,v} = \mathbb{1}\left[\mathbf{F}^{\text{raw}}_{t,v} > \delta\right] \mathbf{M}^F_{t,v}, \tag{15}$$

where $\delta = 0.05$ in our default setting, and

$$\mathcal{L}_A = \left\langle \text{SmoothL1}(\hat{\mathbf{F}}, \mathbf{F}) \right\rangle_{\mathbf{M}^A}. \tag{16}$$

Contact loss with positive reweighting α is

$$\mathcal{L}_C = \left\langle \text{BCEWithLogits}\Big(\hat{\mathbf{Z}}, \mathbf{C}; \alpha\Big) \right\rangle_{\mathbf{M}^C}, \tag{17}$$

where $\alpha = 2.0$ in our default setting.

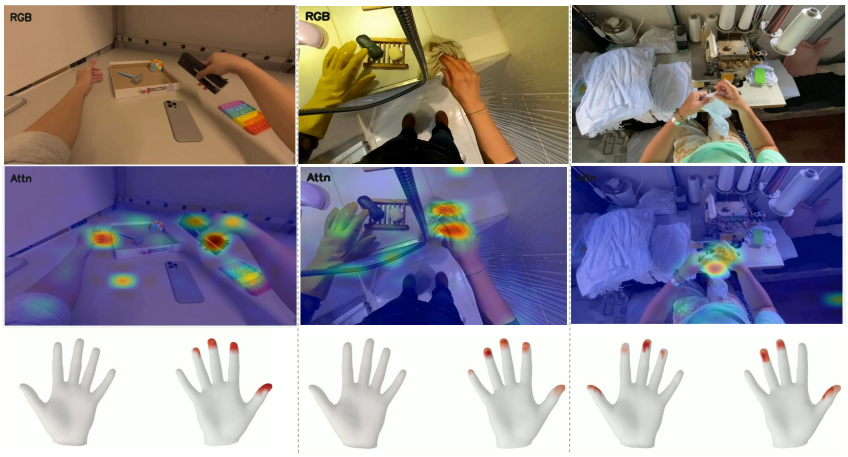

Figure X4: **Attention maps from the DINO encoder [34].** Across diverse tasks, EgoTac emphasizes hand-object interaction regions that are informative for tactile prediction.



Figure X5: **Additional qualitative results in diverse in-the-wild scenes.** We show predictions on EPIC-KITCHENS [5], EgoDex [18], Ego4D [16], EgoPAT3D [27], and self-collected videos.

For consistency, we derive force-branch contact logits

$$\hat{\mathbf{Z}}^F = \frac{\hat{\mathbf{F}}^{\text{raw}} - \theta}{\tau_c}, \tag{18}$$

with $\theta = 0.05$ and $\tau_c = 0.03$, and use detached contact-head probabilities as soft targets:

$$\mathcal{L}_{\text{cons}} = \left\langle \text{BCEWithLogits}\left(\hat{\mathbf{Z}}^F, \sigma(\hat{\mathbf{Z}})^{\text{stopgrad}}\right) \right\rangle_{\mathbf{M}^C}. \tag{19}$$

The total objective is

$$\mathcal{L} = \lambda_F \mathcal{L}_F + \lambda_A \mathcal{L}_A + \lambda_C \mathcal{L}_C + \lambda_{\text{cons}} \mathcal{L}_{\text{cons}}, \tag{20}$$

with $(\lambda_F, \lambda_A, \lambda_C, \lambda_{\text{cons}}) = (1.0, 0.2, 1.0, 0.05)$ for the main setting.

D.4 Hyperparameter settings

Our model uses a pretrained ViT visual encoder and a 12-layer AdaLN Transformer decoder (d=768, 12 heads). Training uses AdamW [32] with cosine learning-rate decay and warm-up, with mixed-precision distributed optimization and gradient clipping. Key hyperparameters are listed in Table X1.

E Details of metrics

This section defines the metrics reported in Sec. 5. We group them into four categories: (1) in-domain continuous force metrics, (2) contact benchmark metrics, (3) OpenTouch OOD force-dynamics metrics, and (4) frequency-domain force-dynamics metrics.

E.1 Tactile-related metrics

For in-domain continuous tactile evaluation, the main paper reports three metrics: global MAE, active-region MAE and temporal Pearson correlation.

Let $F_{t,v}^{\mathrm{pred}}$ and $F_{t,v}^{\mathrm{gt}}$ denote predicted and ground-truth force at frame t and MANO vertex v, and let Ω be the valid evaluation set.

Global MAE.

$$\mathrm{MAE} = \frac{1}{|\Omega|} \sum_{(t,v)\in\Omega} \left| F_{t,v}^{\mathrm{pred}} - F_{t,v}^{\mathrm{gt}} \right|, \tag{21}$$

which measures overall force error across the full hand surface. This metric is sensitive to the strong class imbalance between contact and non-contact regions.

Active-region MAE. We define active vertices by a force threshold $\theta = 0.05$ N:

$$\Omega_{\mathrm{act}} = \{(t,v) \in \Omega \mid F_{t,v}^{\mathrm{gt}} > \theta\}. \tag{22}$$

Then

$$\mathrm{MAE}_{\mathrm{act}} = \frac{1}{|\Omega_{\mathrm{act}}|} \sum_{(t,v)\in\Omega_{\mathrm{act}}} \left| F_{t,v}^{\mathrm{pred}} - F_{t,v}^{\mathrm{gt}} \right|, \tag{23}$$

which focuses on physically meaningful contact areas and is therefore emphasized in our analysis.

Temporal Pearson correlation. For each frame, we compute total hand force

$$S_t = \sum_v F_{t,v}. \tag{24}$$

Temporal Pearson is

$$\mathrm{TempPearson} = \rho\Big(\{S_t^{\mathrm{pred}}\}_t, \{S_t^{\mathrm{gt}}\}_t \Big), \tag{25}$$

which measures whether the predicted force trajectory follows the temporal evolution of real interaction intensity.

E.2 Contact-related metrics

For contact benchmarks (in-domain and OOD), the main paper reports Precision, Recall, F1, IoU, and AUROC for MANO contact prediction.

Given binary prediction $\hat{y}$ and ground truth y, with TP/FP/FN/TN defined at vertex level:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \quad \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{26}$$

$$\mathrm{F1} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}, \quad \mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}, \tag{27}$$

and AUROC is computed from continuous contact scores before thresholding. Precision emphasizes false-positive control, Recall emphasizes missed-contact control, F1 balances the two, and IoU provides a stricter overlap measure under sparse positives.

E.3 OpenTouch OOD force-dynamics metrics

For the OOD force-sensitive benchmark on OpenTouch [39], we report **force-cosine**, **rise F1**, and **fall F1**, as in Sec. 5.3. Since sensor calibration differs across datasets, we evaluate relative dynamics after normalizing force signals to $[0, 1]$.

Force-cosine. Let $\mathbf{s}^{\mathrm{pred}}, \mathbf{s}^{\mathrm{gt}} \in \mathbb{R}^T$ be predicted and ground-truth total-force time series (after normalization and masking). Force-cosine is

$$\text{Force-Cosine} = \frac{\langle \mathbf{s}^{\mathrm{pred}}, \mathbf{s}^{\mathrm{gt}} \rangle}{\|\mathbf{s}^{\mathrm{pred}}\|_2 \|\mathbf{s}^{\mathrm{gt}}\|_2 + \epsilon}. \tag{28}$$

It measures trend-shape consistency independent of global scale.

Rise F1 and Fall F1. Define first-order differences

$$\Delta s_t = s_t - s_{t-1}. \tag{29}$$

With event threshold $\eta = 1e^{-4} > 0$, rise and fall events are:

$$e_t^{\uparrow} = \mathbb{1}[\Delta s_t > \eta], \qquad e_t^{\downarrow} = \mathbb{1}[\Delta s_t < -\eta]. \tag{30}$$

We compute F1 between predicted and ground-truth event sequences separately:

$$\text{Rise F1} = \mathrm{F1}(e^{\uparrow,\mathrm{pred}}, e^{\uparrow,\mathrm{gt}}), \quad \text{Fall F1} = \mathrm{F1}(e^{\downarrow,\mathrm{pred}}, e^{\downarrow,\mathrm{gt}}). \tag{31}$$

These two metrics evaluate whether the model captures *when force increases* and *when force decreases*, which is crucial for contact transition dynamics and manipulation timing.

E.4 Frequency-domain metrics

In addition to the time-domain metrics above, we evaluate whether the predicted tactile signals preserve the temporal structure of real interactions in the frequency domain. These metrics provide a complementary characterization of tactile dynamics beyond point-wise magnitude errors and local rise/fall events.

For each episode, we first aggregate the dense MANO tactile field into a scalar force trace:

$$s_t = \sum_v F_{t,v}. \tag{32}$$

Given a force sequence $\mathbf{s} = \{s_t\}_{t=1}^T$, we remove its DC component and apply a Hann window before computing a single-sided real discrete Fourier transform (DFT):

$$\widetilde{s}_t = w_t \left(s_t - \frac{1}{T} \sum_{\tau=1}^{T} s_\tau \right), \qquad \mathcal{F} = \mathrm{rFFT}(\widetilde{\mathbf{s}}), \tag{33}$$

where w_t is the Hann-window coefficient. We apply the same procedure to the predicted and ground-truth force traces and denote their spectra as $\mathcal{F}^{\mathrm{pred}}$ and $\mathcal{F}^{\mathrm{gt}}$, respectively.

DFT log-magnitude Pearson correlation. We first compute the log-magnitude spectrum

$$A(f) = \log\left(1 + |\mathcal{F}(f)|\right). \tag{34}$$

The DFT-Pearson metric is then defined as

$$\text{DFT-Pearson} = \rho\left(A^{\mathrm{pred}}, A^{\mathrm{gt}}\right), \tag{35}$$

where $\rho(\cdot,\cdot)$ denotes the Pearson correlation across frequency bins. This metric measures the similarity between the predicted and ground-truth spectral envelopes while being relatively insensitive to a global amplitude scale. Higher is better.

Spectral convergence. To additionally measure spectral magnitude mismatch, we compute

$$\text{SpecConv} = \frac{\left\| |\mathcal{F}^{\mathrm{pred}}| - |\mathcal{F}^{\mathrm{gt}}| \right\|_2}{\| |\mathcal{F}^{\mathrm{gt}}| \|_2 + \epsilon}. \tag{36}$$

Unlike DFT-Pearson, spectral convergence is sensitive to differences in the magnitude of frequency components. Lower is better.

Table X2: Temporal and frequency-domain evaluation on the in-domain EgoTac-SC evaluation set.

Method	Force-Cosine ↑	Rise-F1 ↑	Fall-F1 ↑	DFT-Pearson ↑	Spec. Conv. ↓	Domin. Freq. Error ↓
EgoTac-i	0.829	0.396	0.391	0.831	0.684	0.206
EgoTac	**0.844**	**0.424**	**0.397**	**0.860**	**0.645**	**0.184**

Table X3: Temporal and frequency-domain evaluation on the OOD OpenTouch benchmark under different training-data scales. (We use EgoTac-i for this ablation)

Data Use Ratio	Force-Cosine ↑	Rise-F1 ↑	Fall-F1 ↑	DFT-Pearson ↑	Spec. Conv. ↓	Domin. Freq. Error ↓
10%	0.754	0.397	0.394	0.754	0.683	0.127
20%	0.756	0.408	0.409	0.767	0.656	**0.108**
50%	0.756	0.413	0.405	0.744	0.699	0.131
100%	**0.787**	**0.419**	**0.417**	**0.771**	**0.641**	0.119

Dominant-frequency error. We further compare the locations of the strongest non-DC frequency components. Let

$$f_{\mathrm{dom}} = \arg\max_{f>0} |\mathcal{F}(f)| \,. \tag{37}$$

The dominant-frequency error is defined as

$$E_{\mathrm{dom}} = \left| f_{\mathrm{dom}}^{\mathrm{pred}} - f_{\mathrm{dom}}^{\mathrm{gt}} \right| . \tag{38}$$

This metric measures whether the dominant interaction rhythm of the predicted tactile signal occurs at a frequency similar to that of the ground truth. We report the error in Hz, and lower is better.

Together, these frequency-domain metrics evaluate whether EgoTac preserves the spectral structure of tactile dynamics, complementing the time-domain force-cosine and rise/fall metrics.

F More experimental results

This section provides additional results omitted from the main paper due to space constraints.

Frequency-domain evaluation. To complement the time-domain force-dynamics metrics, we further evaluate the spectral fidelity of predicted tactile signals using the frequency-domain metrics defined in Sec. E.4, including DFT log-magnitude Pearson correlation, spectral convergence, and dominant-frequency error.

We first evaluate on the in-domain EgoTac-SC test set. As shown in Table X2, EgoTac consistently outperforms the image-based EgoTac-i variant. In particular, DFT-Pearson improves from 0.831 to 0.860, while dominant-frequency error decreases from 0.206 Hz to 0.184 Hz, indicating that temporal visual context improves both time-domain and spectral tactile dynamics.

We further evaluate OOD transfer on OpenTouch under different training-data scales. Table X3 shows that EgoTac consistently preserves the coarse spectral structure of the ground-truth tactile trajectories. DFT-Pearson ranges from 0.744 to 0.771 and dominant-frequency error remains between 0.108 and 0.131 Hz. With the full training set, EgoTac achieves the best DFT-Pearson of 0.771 and spectral convergence of 0.641, together with the strongest overall temporal-dynamics performance.

Together, these results show that EgoTac captures tactile dynamics in both the temporal and frequency domains. On EgoTac-SC, calibrated MAE measures force-magnitude accuracy within the same sensing system, whereas the temporal and spectral metrics characterize dynamics fidelity. On OpenTouch, where the sensing hardware and calibration differ, these metrics evaluate cross-sensor transfer of normalized tactile dynamics rather than Newton-calibrated absolute-force accuracy.

EgoTac-SC dataset ablation. We conduct additional ablations to isolate the contribution of the self-collected EgoTac-SC dataset from the existing contact-only HOI datasets. Specifically, we

Table X4: Dataset ablation on the OOD contact benchmark OAKINK2. All results use the same whole-hand evaluation protocol as the cross-method comparison in the main paper.

Training setting	AUROC ↑	IoU ↑	Precision ↑	Recall ↑	F1 ↑
EgoTac-SC only	0.60	0.07	0.35	0.08	0.12
Contact-only HOI	**0.84**	**0.29**	**0.51**	**0.45**	**0.44**
EgoTac-SC + Contact-only HOI	**0.84**	**0.29**	0.50	0.44	**0.44**

Table X5: Dataset ablation on the OOD force-dynamics benchmark OpenTouch. Contact-only HOI datasets do not contain continuous force targets and therefore cannot independently train the force predictor.

Training setting	Force-Cosine ↑	Rise-F1 ↑	Fall-F1 ↑	DFT-Pearson ↑	Spec. Conv. ↓	Domin. Freq. Error ↓
EgoTac-SC only	0.745	**0.404**	0.385	**0.767**	**0.656**	0.142
Contact-only HOI	–	–	–	–	–	–
EgoTac-SC + Contact-only HOI	**0.777**	0.403	**0.403**	0.756	0.693	**0.127**

compare three training configurations: (1) EgoTac-SC only, (2) contact-only HOI datasets only, and (3) the full mixed-data setting. Here, the contact-only HOI datasets include HOT3D, HOI4D, TACO, H2O, and ARCTIC. The mixed setting, EgoTac-SC + Contact-only HOI, is exactly the training configuration used by the full EgoTac model and does not introduce any additional training data.

We first evaluate OOD contact prediction on OAKINK2 using the same whole-hand protocol as the main paper, where all 778 MANO vertices are included in evaluation. As shown in Table X4, training only on EgoTac-SC results in substantially weaker OOD contact generalization, whereas the contact-only HOI datasets provide strong cross-domain contact performance. The mixed model maintains this strong contact generalization while additionally benefiting from the continuous force supervision provided by EgoTac-SC.

We further evaluate the same training configurations on the OOD force-sensitive OpenTouch benchmark. As shown in Table X5, contact-only HOI data cannot independently train a continuous force predictor because it contains only binary contact supervision. In contrast, EgoTac-SC provides the continuous tactile annotations required for learning force dynamics. Combining EgoTac-SC with contact-only HOI data retains comparable force-dynamics performance while improving several metrics, suggesting that the two supervision sources are complementary.

Overall, these results highlight the complementary roles of the two data sources. EgoTac-SC uniquely provides **real continuous-force supervision** that enables tactile-intensity and force-dynamics learning, while heterogeneous bare-hand contact datasets substantially improve **visual diversity and OOD contact generalization**.

$\mathcal{L}_{\text{cons}}$ ablation. To further understand why contact-only supervision can affect force prediction, we isolate the contribution of the force–contact consistency loss $\mathcal{L}_{\text{cons}}$. Specifically, we train an additional mixed-data model using the same EgoTac-SC and contact-only HOI datasets, while removing $\mathcal{L}_{\text{cons}}$ and keeping all other training settings unchanged.

As shown in Table X6, mixed-source training without $\mathcal{L}_{\text{cons}}$ already improves several force-dynamics metrics over EgoTac-SC-only training. For example, force-cosine increases from 0.745 to 0.759 and dominant-frequency error decreases from 0.142 to 0.131 Hz. This suggests that the broader visual coverage provided by contact-only HOI data can benefit force prediction through the shared visual-temporal representation, even without an explicit consistency constraint.

Adding $\mathcal{L}_{\text{cons}}$ further improves force-cosine from 0.759 to 0.777, fall F1 from 0.400 to 0.403, and dominant-frequency error from 0.131 to 0.127 Hz. However, the improvement is not consistent across all metrics. We therefore interpret $\mathcal{L}_{\text{cons}}$ as providing a modest structural regularization effect rather than being the primary source of the mixed-data improvement.

Table X6: Ablation of the force–contact consistency loss $\mathcal{L}_{\text{cons}}$ on the OOD OpenTouch benchmark.

Training setting	Force-Cosine ↑	Rise-F1 ↑	Fall-F1 ↑	DFT-Pearson ↑	Spec. Conv. ↓	Domin. Freq. Error ↓
EgoTac w/o $\mathcal{L}_{\text{cons}}$	0.759	0.398	0.400	0.753	**0.672**	0.131
EgoTac	**0.777**	**0.403**	**0.403**	**0.756**	0.693	**0.127**

Table X7: **More results of in-domain contact evaluation.** Results are averaged across timestamps on all test splits of each component dataset.

Benchmark	Method	AUROC ↑	IoU ↑	Precision ↑	Recall ↑	F1 ↑
TACO [30]	EgoTac-i	**0.974**	**0.657**	**0.732**	0.855	**0.788**
	EgoTac	0.973	0.656	0.729	**0.858**	0.787
HOI4D [31]	EgoTac-i	**0.961**	0.575	0.647	**0.828**	0.720
	EgoTac	0.959	**0.579**	**0.657**	0.818	**0.724**

Additional in-domain results. We report additional in-domain evaluation results on TACO [30] and HOI4D [31] test sets in Table X7, along with additional results for training-data volume scaling on in-domain contact benchmarks in Table X8.

Additional visualizations. Figure X4 presents attention maps in diverse scenes, showing that EgoTac consistently focuses on hand-object interaction regions that determine tactile patterns. Figure X5 provides additional qualitative examples of in-the-wild tactile prediction from egocentric videos.

G More discussions

This section provides further discussion on several aspects of EgoTac, including the visual-domain gap introduced by tactile gloves, the interaction between contact-only supervision and force prediction, and the scope of our architectural contribution.

Visual-domain gap introduced by tactile gloves. EgoTac-SC is collected using wearable tactile gloves, which inevitably introduces a visual appearance gap between the force-supervised training data and unconstrained bare-hand videos. A natural strategy is to remove or translate the glove appearance before training. In a preliminary study, we used Grounded SAM-2 to segment the gloves and recolored the segmented regions toward a skin-like appearance. However, we did not observe a clear improvement in zero-shot tactile prediction on bare-hand videos.

We believe that simple recoloring or naive inpainting is insufficient for this problem. The glove physically occludes the underlying hand appearance, and therefore appearance translation must reconstruct rather than merely recolor the missing visual information. Moreover, artifacts around finger boundaries, hand–object occlusions, and fine-grained contact regions may remove or distort precisely the visual cues that are most informative for tactile prediction.

Instead, EgoTac mitigates this domain gap through heterogeneous mixed-source training. We map multiple bare-hand HOI datasets into the same MANO-aligned representation as EgoTac-SC and jointly train the model using gloved force-supervised data and bare-hand contact-supervised data. As shown by the dataset ablations in Sec. F, the bare-hand contact datasets substantially improve OOD contact generalization, while EgoTac-SC provides continuous-force supervision that is unavailable in these contact-only datasets. We therefore view mixed-domain supervision as a complementary alternative to direct appearance translation. More sophisticated geometry-preserving glove-to-hand translation [46] remains a promising direction for future work.

Why contact-only data can affect force dynamics. Contact-only HOI samples contain no continuous-force targets. Accordingly, their force-valid mask M_F is zero, and they do not directly contribute to the force-regression loss $\mathcal{L}_F$ or the active-region loss $\mathcal{L}_A$. Nevertheless, contact-only data can still influence force prediction through two complementary mechanisms.

First, the force and contact tasks share the same visual encoder, temporal fusion module, and tactile decoder. The contact classification objective therefore improves the shared visual-temporal

Table X8: **Data volume scaling on in-domain contact estimation.** As training data scales from 10% to 100 %, contact estimation performance on held-out in-domain test sets consistently improves. We conduct these experiments on EgoTac-i.

Benchmark	Use Ratio	AUROC ↑	IoU ↑	Precision ↑	Recall ↑	F1 ↑
HOT3D [1]	10%	0.914	0.392	0.586	0.524	0.524
	20%	0.932	0.454	0.603	0.603	0.581
	50%	0.958	0.539	0.634	**0.771**	0.658
	100%	**0.976**	**0.587**	**0.656**	**0.771**	**0.702**
H2O [24]	10%	0.936	0.323	0.565	0.469	0.458
	20%	0.954	0.425	0.626	0.613	0.574
	50%	0.975	0.567	0.708	0.737	0.706
	100%	**0.989**	**0.647**	**0.725**	**0.846**	**0.777**
ARCTIC [10]	10%	0.810	0.236	0.462	0.356	0.381
	20%	0.846	0.330	0.483	0.518	0.480
	50%	0.910	0.475	0.628	0.674	0.636
	100%	**0.972**	**0.676**	**0.716**	**0.905**	**0.790**
TACO [30]	10%	0.933	0.510	0.647	0.689	0.660
	20%	0.949	0.564	0.686	0.743	0.708
	50%	0.968	0.634	0.720	0.830	0.769
	100%	**0.974**	**0.657**	**0.732**	**0.855**	**0.788**
HOI4D [31]	10%	0.934	0.449	0.638	0.629	0.600
	20%	0.943	0.493	0.641	0.700	0.644
	50%	0.955	0.547	0.644	0.785	0.693
	100%	**0.961**	**0.575**	**0.647**	**0.828**	**0.720**

representation using a broader distribution of bare-hand appearances, objects, viewpoints, scenes, and contact transitions. These improved shared features can subsequently benefit the force prediction branch even without direct force supervision.

Second, the consistency objective $\mathcal{L}_{\text{cons}}$ provides an explicit interaction between the two prediction branches. As described in Sec. D.3, $\mathcal{L}_{\text{cons}}$ aligns the contact logits implicitly derived from the predicted force with the detached prediction of the contact head. Therefore, for contact-only samples, the contact prediction can provide the force branch with a binary structural constraint indicating *where* and *when* force should be active, although it does not specify the corresponding force magnitude.

The controlled ablation in Sec. F further shows that removing $\mathcal{L}_{\text{cons}}$ from mixed-data training only moderately changes the OpenTouch force-dynamics metrics. This suggests that the benefit of contact-only data cannot be attributed to the consistency loss alone. Instead, we interpret the improvement as a combination of broader visual-domain coverage through shared representation learning and a smaller regularization effect from $\mathcal{L}_{\text{cons}}$. Importantly, continuous tactile intensity itself remains uniquely supervised by EgoTac-SC.

Architectural novelty. The individual architectural components of EgoTac, including the pre-trained vision backbone, temporal attention, and AdaLN-based Transformer decoder, are established techniques. Our contribution is therefore not intended to introduce a new Transformer primitive or a highly specialized network module. Instead, the technical contribution lies in formulating and enabling dense vision-to-tactile learning from large-scale heterogeneous supervision.

Specifically, EgoTac represents continuous tactile force and binary contact on a shared two-hand MANO topology, allowing annotations from different sources to share a common output space. It jointly learns from force-supervised tactile data and contact-only HOI datasets despite their different annotation availability, using force-valid and contact-valid masks so that each sample supervises only the annotations it provides. The force and contact predictions are further coupled by the consistency objective, which encourages structural agreement between continuous tactile intensity and binary contact.

Together, these design choices allow data sources with fundamentally different label types, spatial coverage, and visual domains to contribute to a single dense tactile predictor. We intentionally adopt a concise encoder–decoder architecture so that the effects of the unified representation, heterogeneous

supervision, and data scaling can be studied without conflating them with gains from a highly specialized backbone. We therefore view the main technical novelty of EgoTac as the unified tactile representation and heterogeneous force–contact learning framework, rather than the introduction of a new backbone architecture.

H Limitations and societal impacts

Limitations. Despite these advances, EgoTac has limitations: it may underperform under severe occlusion, strong motion blur or rare interaction patterns. Future work could address these issues by incorporating occlusion reasoning and motion deblurring, as well as expanding training data diversity to better capture rare interactions. In addition, EgoTac-SC is collected with tactile gloves, introducing a visual-domain gap with bare-hand videos. Beyond methodological improvements, EgoTac-predicted tactile signals offer rich priors for downstream applications, such as pretraining robotic manipulation policies, enhancing simulation realism, and enabling contact-aware action planning in human-robot interaction. These directions collectively advance the integration of visual perception and tactile understanding for more physically informed AI and robotic systems.

Societal impacts. This work can benefit robotics, AR/VR, and assistive systems by enabling richer hand-state estimation from egocentric visual input. Potential misuse includes privacy-invasive behavior inference from first-person recordings. Responsible deployment should require explicit consent, strict data governance, and task-specific risk assessment before real-world use.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction correspond directly to the methodology (Section 4) and are empirically supported by the experiments (Section 5).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendices G and H provide discussions on the visual-domain gap, difficult visual conditions, and the scope of force and downstream evaluations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper does not present theoretical proofs, as it focuses on empirical methodology and dataset construction.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe our model architecture explicitly in Section 4 and detail the complete training protocols and evaluation metrics in the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will open-source the codebase, pre-trained weights, and the EgoTac-SC dataset upon publication. We have provided necessary material in the supplementary.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Key implementation details are stated in Section 4.4, with comprehensive hyperparameters and data-split strategies provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The results are obtained from standard randomly seeded runs without cherrypicking, offering a fair and representative evaluation despite not including error bars.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide experimental details and computer resources both in the Section 4.4 and the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research stringently conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have included a discussion regarding the positive impacts (e.g., embodied AI) and potential negative societal impacts in the supplementary material.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The introduced tactile prediction models and hand-object interaction datasets do not pose a high risk for malicious misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly cite all prior datasets used to construct our mixed corpus, and their respective licenses are respected and detailed in the supplementary material.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The newly introduced EgoTac-SC dataset is extensively documented in Section 3 and the appendix, and we will provide full documentation and scripts upon release.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The dataset collection was conducted locally without utilizing external crowdsourcing platforms.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The local data collection strictly involved consenting volunteers performing mundane daily tasks, which does not pose risks requiring institutional review board (IRB) assessments beyond standard laboratory safety.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We did not use LLMs as a core or original method component in this research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.