FEEL (Force-Enhanced Egocentric Learning): A Dataset for Physical Action Understanding

Eadom Dessalene, Botao He, Michael Maynord, Yonatan Tussa, Pavan Mantripragada, Yianni Karabatis, Nirupam Roy, and Yiannis Aloimonos

University of Maryland, College Park, USA
Email: edessale@umd.edu

https://www.cs.umd.edu/~edessale/feel

arXiv:2603.15847v1 [cs.CV] 16 Mar 2026

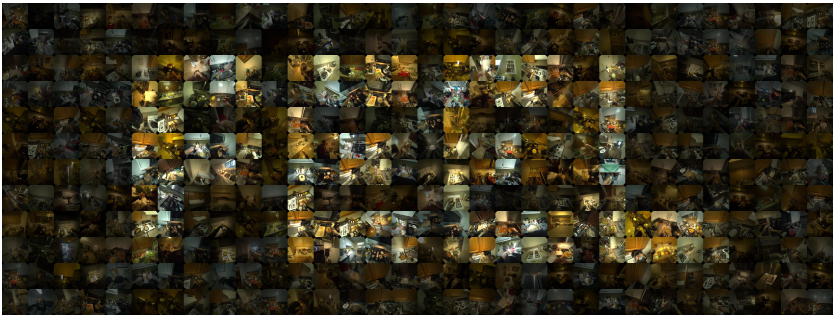

Fig. 1: FEEL pairs egocentric video with force measurements to capture the physical causes—not just visual effects—of hand-object interaction.

Abstract. We introduce FEEL (Force-Enhanced Egocentric Learning), the first large-scale dataset pairing force measurements gathered from custom piezoresistive gloves with egocentric video. Our gloves enable scalable data collection, and FEEL contains approximately 3 million force-synchronized frames of natural unscripted manipulation in kitchen environments, with ∼45% of frames involving hand-object contact. Because force is the underlying cause that drives physical interaction, it is a critical primitive for physical action understanding. We demonstrate the utility of force for physical action understanding through application of FEEL to two families of tasks: (1) **contact understanding**, where we jointly perform temporal contact segmentation and pixel-level contacted object segmentation; and, (2) **action representation learning**, where force prediction serves as a self-supervised pretraining objective for video backbones. We achieve state-of-the-art temporal contact segmentation results and competitive pixel-level segmentation results without *any* need for manual contacted object segmentation annotations. Furthermore we demonstrate that action representation learning with FEEL improves transfer performance on action understanding tasks without *any manual labels* over EPIC-Kitchens, SomethingSomething-V2, EgoExo4D and Meccano.

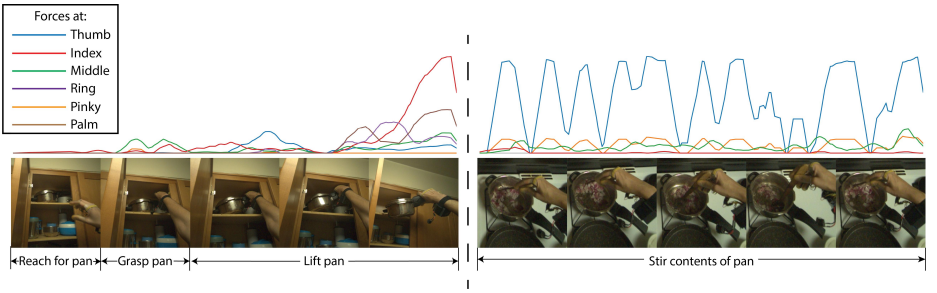

Fig. 2: Forces reveal interaction dynamics. Representative sequences showing how force measurements disambiguate physical interactions. **Left:** Grasping and lifting a pan exhibits distinct force signatures with an onset at the end of the reach and a spike as the pan is lifted. **Right:** Stirring produces rhythmic force patterns reflecting cyclic manipulation.

1 Introduction

Physical manipulation actions are defined by the forces the hands exert on the world. For this reason, force should not be treated as a peripheral signal, but as a primitive for physical action understanding. Vision, in contrast, primarily measures motion and appearance — observable effects of action but not causal in themselves. Learning actions from video alone leaves the underlying action structure implicit in a high dimensional stream of pixels - incorporating force turns the hidden action structure into an intrinsically lower dimensional learnable signal.

Egocentric video datasets have enabled major progress by focusing on the visual modality for action understanding tasks. This data is useful, but it lacks a modality directly capturing the causal dynamics of action: force. Force is absent from existing datasets largely because it cannot be reliably obtained through manual labeling. The physical response of objects to forces applied are often very subtle and difficult for annotators to recognize directly from visual changes. This difficulty includes annotating the making and breaking of contact. For illustration, see Figure 3 for a case where appearance is inconclusive while force disambiguates.

In order to address the shortcomings of existing datasets, we introduce FEEL (Force-Enhanced Egocentric Learning). FEEL is built around a custom force sensing glove (see Figure 4) consisting of 5 Piezoresistive sensors mounted to the pads of the fingers, as well as a long Piezoresistive sensor mounted to the palm. This glove when worn provides a stream of measurements of the forces imparted by hands. We record this stream alongside egocentric captures collected with Meta Aria glasses [15]. The result is time-synchronized egocentric video, audio, and force sensing of natural, unscripted, unstructured manipulation. See Figure 2 for illustration of synchronized video and force sensing streams. Contact is derived from force exchange, yielding temporally precise per-frame contact state and boundaries (see Figure 3). A key benefit to force-derived contact labels as

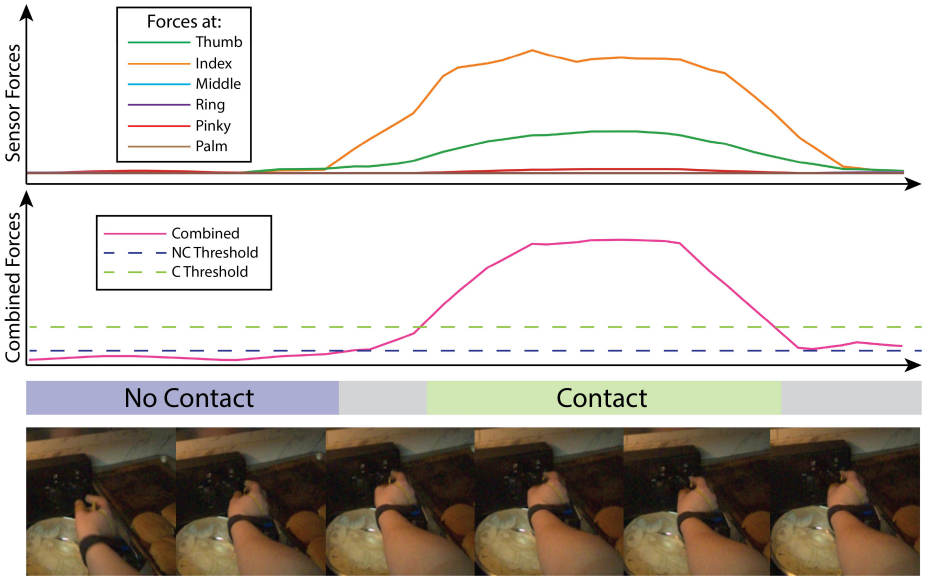

Fig. 3: Contact detection from force measurements. Despite visual similarity across frames, force measurements precisely identify contact boundaries. **Top:** Raw sensor forces (pinky and middle finger dominate this grasp). **Middle:** Consolidated force with dual thresholds—above C threshold indicates contact, below NC threshold indicates non-contact, between is ambiguous (excluded). Forces enable scalable contact supervision without manual annotation of ambiguous visual transitions.

opposed to manual annotation of contact labels over RGB frames is scalability. We focus on kitchens, which provide a broad range of interaction/action types. FEEL contains $\sim$ 3M frames over $\sim$ 27 hours of dense manipulation, with $\sim$ 45% of frames involving contact.

Using force as supervision provides a low-dimensional physically grounded signal that constrains how action is extracted from high-dimensional pixel space. Along these lines we develop learning methods leveraging forces as supervision for the following tasks: Contact Understanding and Action Representation Learning (see Figure 5).

In the **Contact Understanding** task we leverage our force signal in the modeling of contact in egocentric video. This task is defined as jointly: (1) temporal contact segmentation - classifying the per-frame contact state over time for each hand; and, (2) pixel-level segmentation - segmenting the object(s) in contact with each hand. Our force measurements provide temporally precise supervision for the onset and offset of contact, enabling the learning of sharp interaction boundaries which would be difficult to learn over imprecise contact annotations (see the Supplementary Materials for examples).

In the **Action Representation Learning** task, we use FEEL to demonstrate the utility of force for learning action representations over video. Training to

predict forces encourages the network to infer physically grounded action structure from video. We evaluate transfer of the resulting representations to the task of action recognition. We evaluate on established benchmarks, including EPIC-Kitchens, Ego-Exo4D, Something-Something V2, and Meccano.

The primary contributions of this work are as follows:

1. **Release of FEEL:** We release the largest egocentric force–video dataset, capturing the physical causes - not just the visual effects - of hand-object manipulation. FEEL consists of approximately 3 million frames collected during natural, unscripted manipulation.
2. **Scalable force-derived contact labels:** We derive per-frame contact states, releasing $1.35M$ frames of hands and objects in contact (along with object segmentations) and $1.65M$ frames of hands and objects not in contact.
3. **Self-supervised force pre-training pipeline** A demonstration of the benefits of training force-aware video backbones for downstream tasks. Force-aware pretraining over FEEL improves downstream action recognition across nearly all datasets.
4. **A new state-of-the-art contact estimation model:** We train a model to jointly perform temporal contact segmentation and pixel-level object-in-contact segmentation over still images. Our model achieves state-of-the-art temporal contact detection across all three benchmarks.

The rest of this paper is structured as follows: In Section 2 we cover related work, in Section 3 we detail our method, in Section 4 we cover evaluations, and in Section 5 we discuss and conclude.

2 Related Work

Contact Understanding The predominant paradigm for contact understanding relies on large, manually annotated datasets over images [7, 8, 26, 32, 40] or video [8, 10, 13, 23], which incur significant annotation cost and label noise. Some works mitigate this through semi-automatic segmentation workflows that propagate manual annotations using off-the-shelf video tools [10, 13]. Alternative approaches collect data in controlled lab settings with motion capture to precisely track hands and objects in 3D [4, 16, 39], or sidestep real-world collection entirely via synthetic data [20].

Force Understanding Imbuing vision models with physical understanding remains highly challenging and underexplored [14, 17, 29, 44]. Due to the scarcity of real-world force data, most vision-based approaches rely on simulation [14, 21, 44]. In robotics, a growing body of work addresses force-aware visual representations through real-world data collection [9, 11, 42], though these are confined to simple manipulation tasks such as pick-and-place. The closest work to ours is [38]; our dataset is over 5× larger and we explore tasks un-explored within [38].

Applications of Force and Contact Understanding Reliably inferring physical hand-object interactions unlocks a range of important downstream tasks. Contact understanding has been shown to improve action recognition [13, 23, 34]

and to enable robots to learn manipulation trajectories by parsing contacts between hands and objects from human video [27, 36, 37]. Force understanding has seen comparatively fewer applications: [1] trains force-aware robot policies on a small dataset of simplistic tasks, and [38] successfully demonstrates that forces improve grasp classification performance. We are the first to demonstrate zero-shot transfer of our learned representations over our force dataset for both contact understanding and action representation learning.

3 Methods

We introduced two families of learning tasks that leverage force as supervision: **contact understanding** and **action representation learning** across modalities. Contact understanding jointly addresses temporal contact segmentation (per-frame contact state classification) and pixel-level segmentation of objects in contact. For action understanding, we use force prediction as a pretraining objective to learn representations that are more physically grounded than appearance based features alone. Figure 5 illustrates our approach: we train modality-specific backbones on image and video inputs with force prediction as a self-supervised pretraining objective.

The remainder of this section is organized as follows: Section 3.1 describes our data collection pipeline and the custom force-sensing hardware; Section 3.2 details contact understanding; Section 3.3 describes action representation learning. The availability of synchronized force measurements as a new supervisory modality enables both the scalable derivation of contact annotations and the learning of physically grounded action representations — applications that would be difficult or impossible with vision alone.

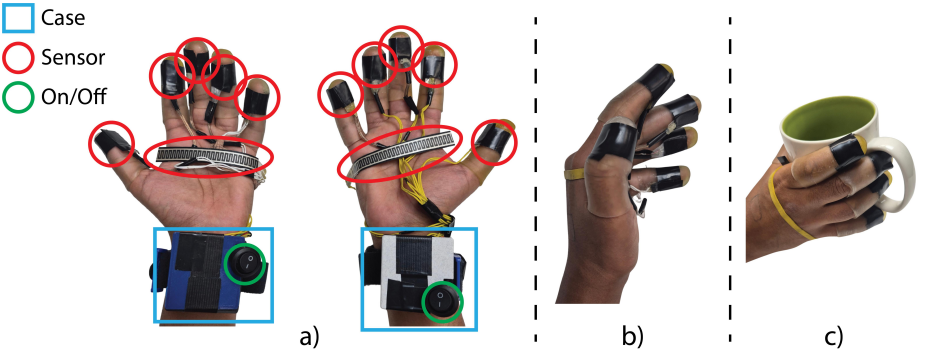

Fig. 4: Force-sensing glove hardware. Leftmost: Palm-facing views of the left and right gloves, showing the six piezoresistive sensors (red) mounted at each fingertip and across the palm, the Arduino microcontroller case (blue) mounted at the wrist, and the on/off switch (green). Middle and right: Side and in-use views demonstrating the glove's low-profile form factor during natural object manipulation.

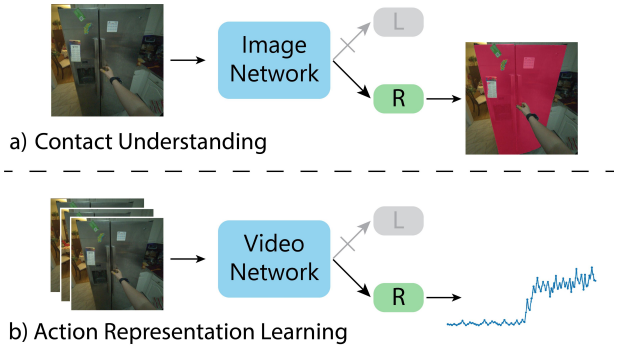

Fig. 5: Learning contacts and action from forces. (a) Contact Understanding Image network trained with force-derived contact labels performs contact detection and segmentation (red mask) for each hand (L/R). **(b) Action Representation Learning** Video network predicts per-hand forces from clips for pretraining. The video backbone is transferred to action recognition after discarding force heads. No manual contact or action labels required during pre-training.

3.1 Dataset Collection

To obtain synchronized visual, acoustic, and force streams, we developed a lightweight sensing system combining a custom piezoresistive glove with Meta Aria glasses for egocentric capture.

Hardware Our sensing system combines two synchronized components: a custom piezoresistive force glove and Meta's Project Aria glasses (Gen 1) . This pairing enables simultaneous capture of manipulation forces, egocentric visual streams, and spatial audio.

Force Glove The force sensing component consists of six piezoresistive sensors (Flexiforce A301, Tekscan) mounted on a glove - one sensor per fingertip and a sensor strip mounted on the palm. See Figure 4 for the glove we construct and use for data collection. Each A301 sensor measures forces up to 45 N (approximately 10 lbs). The sensors connect to an Arduino Nano microcontroller mounted in a case enclosure that also houses a $3.3V$ Lithium Polymer (LiPo) battery and toggle switch for powering the device on and off to start and stop recordings.

Project Aria Glasses For egocentric visual stream and spatial audio capture, we use Meta's Project Aria device. The device captures synchronized multimodal streams: a forward-facing RGB camera (1440×1440 pixels), two monochrome scene cameras (640×480 pixels) positioned on the left and right temples for wide peripheral coverage, a 7-microphone array for spatial audio, and dual IMUs sampled at 800 Hz and 1000 Hz.

Sensor Calibration and Synchronization Synchronizing the force gloves and Aria glasses requires aligning two independent data streams operating on separate internal clocks. See the Supplementary Materials for the three-step calibration procedure employed at the start of each recording session.

Table 1: Comparison of commonly used hand-object interaction datasets. Modality codes: V=video, A=audio, D=depth, F=force, E=eye tracking. Contact annotation for Ego4D (FHO) is partial (only a single critical frame is annotated out of each sequence). FEEL is the only real-world (Real) egocentric dataset providing both contact annotations and force measurements — all other datasets with force lack contact labels or real-world capture, and all datasets with contact lack force.

Dataset	View	Scale	Real	Modalities	Contact	Forces
EgoHands [22]	Ego	1.5 hrs	✓	V	×	×
Arctic [16]	Both	4 hrs	✓	V/D	✓	×
VISOR [10]	Ego	36 hrs	✓	V/A	✓	×
HOI4D [24]	Ego	45 hrs	✓	V/D	✓	×
Ego4D (FHO) [18]	Ego	120 hrs	✓	V/A	1/2	×
100DOH [32]	Both	-	✓	V	✓	×
Zhang et al. [43]	Neither	1 hour	×	V/F	×	✓
Ehsani et al. [14]	Exo	0.5 hrs	×	V/F	✓	✓
OpenTouch. [38]	Ego	5 hrs	✓	V/A/F/E	✓	✓
FEEL (Ours)	**Ego**	**27 hrs**	✓	**V/A/F/E**	✓	✓

Data Processing Egocentric Pose and Hand Tracking We leverage Meta's Machine Perception Services (MPS) to extract foundational geometric and kinematic information from the Aria recordings. MPS processes the side camera streams and IMU data through a visual-inertial odometry pipeline to produce high-frequency (1 kHz) 6-DoF camera trajectories. Additionally, MPS performs 3D hand tracking on the side camera imagery, recovering hand keypoints for each frame in the recordings.

Additional Processing To arrive at nouns of objects of interaction, we adopt a weak labeling strategy: for each recording session, we partition the video into 2-minute temporal windows and select 8 uniformly sampled frames within each window. Over those frames we annotate the set of object nouns interacted with, without specifying temporal order or precise timing. We further compute dense geometric cues to support contacted object segmentation (see Section 3.2): scaled depth maps are predicted (in meters) using MLDepthPro [3], and optical flow fields are densely estimated between consecutive frames using SEA-RAFT [41].

3.2 Contact Understanding

Prior approaches to contact understanding [10, 32] depend on human annotators to label contact events from RGB frames—a process that is both labor-intensive and difficult to scale. Our approach instead exploits direct physical measurements - which, among other advantages, enable us to determine contact states and temporally localize contacted objects without manual annotation. From these force-derived contact states, we introduce a weakly supervised method for generating pseudolabel segmentations of contacted objects. This physically

grounded supervision enables training paradigms that would not be possible with vision-only approaches.

Contact Detection Raw piezoresistive sensors exhibit substantial baseline drift and high-frequency noise that preclude simple thresholding for contact detection. We apply a multi-stage filtering pipeline to each sensor: outlier removal via Hampel filtering, Gaussian smoothing, time-varying baseline estimation via rolling percentile, baseline subtraction with negative clipping, and exclusion of regions with abrupt baseline changes or high RMS magnitudes. We then normalize each sensor by its 99.5th percentile value and compute the geometric mean across all six sensors to produce a unified force signal robust to individual sensor failures. Frames where this signal exceeds an upper threshold indicate contact, frames below a lower threshold indicate non-contact, and frames between thresholds are excluded as ambiguous regions. For full details see the Supplementary Materials.

Segmentation Pseudolabel Generation We develop a weakly supervised approach to generate pseudo-labels for spatially segmenting objects in contact, eliminating the need for manual object segmentation annotations.

The inputs to our segmentation pipeline are: (i) frames designated as containing contact by our force-based detection system; (ii) a set of object mask proposals produced by a set of concept prompts fed to SAM3 (the set of concepts are the only weak labels provided, produced as per Section 3.1); (iii) pre-computed optical flow fields using SEA-RAFT; (iv) camera intrinsics/extrinsics from MPS; and (v) pre-computed depth maps. The process by which these inputs are derived is described in Section 3.1.

Given these inputs, we generate contacted-object pseudo-labels by: **(1)** extracting the masked optical flow for each object proposal; **(2)** computing the fundamental matrix between the current frame and the next frame using the Aria Machine Perception Services (MPS) provided intrinsics/extrinsics; **(3)** evaluating the Sampson epipolar error [25] within each mask using the masked flow and fundamental matrix (where higher error indicates stronger violations of the static-scene (epipolar) constraint due to object motion or manipulation); **(4)** selecting the proposal with the maximum mean epipolar error; and **(5)** accepting this proposal as the contacted object if and only if at least a set number of pixels belonging to the mask lie within a 3D distance threshold of the 3D hand centroid (ensuring that the selected high-motion region is spatially proximate to the manipulating hand). We have $500K$ images with binary contact pseudolabels, $180K$ of which contain binary segmentations of the object(s) in contact. For full details see Supplementary Materials.

Training To demonstrate the utility of FEEL, we train a contact understanding model for comparison against state of the art vision networks trained against large datasets of manually collected contact annotations. As illustrated in the first row of Figure 5, we train this network to process a single RGB image and produce two prediction heads per hand (left and right): (1) a binary contact classifier indicating whether the hand is in contact with an object, and (2) a segmentor of the object in contact with that hand. We supervise the devoted network head for each hand only when that hand is at least partially visible in the

Table 2: Temporal contact detection results evaluated on the validation sets of FEEL, EPIC-VISOR, and HOI4D. We emphasize that while our model was only trained over 1 dataset (FEEL), the 100DOH model and VISOR model were both trained over the EPIC-VISOR dataset [10]. **BC** represents "Binary contact detection" - a prediction is correct iff it captures the binary contact state between the hand and the object, irrespective of which predicted hand side is involved. **BSC** represents "Binary Signed Contact" - a prediction is correct iff it captures both the binary contact state of between the hand and the object, as well as the hand side involved.

Method	**Sup.**	**Training Set #**	**HOI4D**		**VISOR**		**FEEL**	
			BC	**BSC**	**BC**	**BSC**	**BC**	**BSC**
Depth [3]	Self	5	66.6	66.0	52.8	39.9	48.6	30.0
HOIRef [2]	Full	2 (inc. VISOR)	52.1	50.8	55.4	53.7	59.6	57.2
VISOR [10]	Full	1 (VISOR)	55.3	55.0	66.3	65.2	45.2	40.2
100DOH [32]	Full	2 (inc. VISOR)	**90.0**	80.5	81.3	71.9	49.0	40.0
Ours	Self	1 (inc. FEEL)	82.2	**81.3**	**87.3**	**79.7**	**95.6**	**88.9**

frame; network heads corresponding to out-of-view hands receive no supervision. For data augmentation, we apply random crops and random horizontal flips. When flipping, we mirror both the input image and the target outputs in order to maintain correspondence.

3.3 Learning Forces from Video

Traditionally, video networks are pretrained on large-scale datasets of labeled images and videos (e.g., ImageNet [12] and Kinetics [6]), which provide strong appearance- and motion-level priors but little direct supervision about the underlying physical interaction; in contrast, we use FEEL's physically grounded force signals to pretrain video models to capture force dynamics (see Figure 5) and transfer these representations to downstream action understanding tasks. We formulate video-to-force prediction as a self-supervised pretraining task, then transfer the learned representations to downstream action recognition benchmarks. During the video-to-force pretraining, we only select clips for which at least one hand is visible for 50% of the frames, ensuring that a training signal is reflected in the input images.

4 Experiments

We evaluate FEEL across two families of tasks: (1) **contact understanding**, where we perform temporal contact detection and pixel-level object segmentation, and (2) **action representation learning**, where force prediction serves as a self-supervised pretraining objective for video backbones. See Figure 5 for a depiction of the modeling of both tasks.

Table 3: Spatial contact segmentation results evaluated on the validation sets of FEEL, EPIC-VISOR, and HOI4D. We emphasize that while our model was only trained over 1 dataset (FEEL), the 100DOH model and VISOR model were both trained over the EPIC-VISOR dataset. As the 100DOH model produces bounding boxes (not segmentations), we employ in addition a SAM model (trained over Internet-scale data) to segment the 100DOH model predictions. Results reported as intersection-over-union (IOU) ↑.

Method	**Sup.**	**Training Set #**	**Dataset**		
			HOI4D	**VISOR**	**FEEL**
MOVES [19]	Weak	N/A	-	-	0.309
SAM3 [5]	Full	Comp.	0.201	0.085	0.030
VISOR [10]	Full	1 (inc. VISOR)	0.502	0.402	0.111
100DOH [32]+SAM [5]	Full	Comp. (inc. VISOR)	**0.765**	**0.443**	0.218
Ours	Weak	1 (FEEL)	0.751	0.399	**0.625**

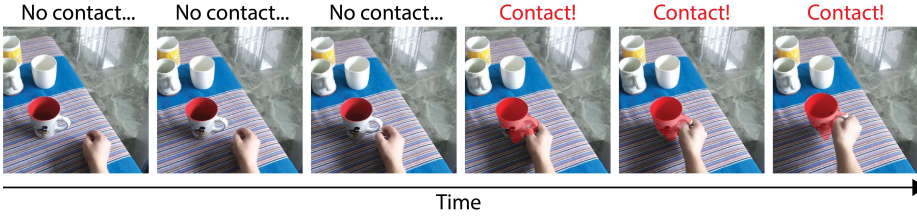

Fig. 6: Temporal contact detection. A representative sequence showing our model's contact predictions over time as a hand reaches toward and grasps a mug. The model correctly predicts no contact during the approach phase and transitions to contact precisely at the moment of grasp.

4.1 Contact Understanding

Task: Contact understanding jointly addresses temporal contact segmentation (per-frame contact state classification) and pixel-level segmentation of objects in contact. The model takes an image as input and produces a prediction as to the binary contact state of each hand as well as a pixel-level segmentation of the contacted object for each hand. See Figure 5.

Model and implementation: Traditional contact understanding models [32] are complex to train due to the non-differentiable operations that happen along the way, along with the additional hyperparameters introduced. For simplicity of implementation, we train a Dense Prediction Transformer (DPT) [30] on top of frozen DINOv3 [35] features. The image is fed as input to DINO, which produces feature maps subsequently fed into DPT, which produces two sets of outputs: one set of outputs as a binary logit based on the contact detected for each of the left and right hand; and a two-channel segmentation map devoted to the left and right hand, returned at the input resolution of the image. For further implementation details, see the Supplementary Materials.

Baselines The contact detections we learn over are derived in a fully self-supervised fashion; the contact segmentations we learn over are derived in a weakly-supervised fashion, as the pseudolabel generation algorithm is provided as input the object categories of all objects interacted with throughout the recording session. At no point is our training dependent on manually provided object segmentations. As such, we compare against other (1) self-supervised approaches, (2) weakly-supervised approaches, and (3) fully-supervised approaches.

Contact detection baselines. To our knowledge, there is no commonly used method for learning to detect contact in a self-supervised fashion. Therefore we introduce a simple zero-shot baseline combining recent advances in depth and segmentation networks. We compute segmentations of the body, and from the depth map of the scene compute a signed distance field of the scene with respect to each of the two hands. If the number of pixels where the distance $d(x, y)$ for all pixels (x, y) exceeds a threshold, the contact prediction is positive, and vice-versa. We also compare against three fully-supervised baselines: HOIRef [2], EPIC-VISOR [10], and 100DOH [32]. See Table 2 for results.

Contact segmentation baselines. As for comparisons to other contact segmentation approaches, the closest self-supervised and weakly-supervised comparisons are [19] and [33], but their methods are not publicly available. We instead compare against a reimplementation of the method for generating ground truth pseudolabels produced within [19] - we compare our network's segmentation predictions directly against those pseudolabels. We also compare against SAM with its concept prompting, as well as two fully-supervised baselines: EPIC-VISOR and 100DOH. See Table 3 for results.

Evaluation Datasets We evaluate our method as well as other method over the contact pseudolabels belonging to FEEL, as well as zero-shot transfer over two popular egocentric video datasets that contain contact annotations - EPIC VISOR [10] and HOI4D [24].

4.2 Action Representation Learning from Video

Action representation learning involves the exploring of pretraining objectives for video backbones before transferring of action representations for downstream tasks. When pre-training over FEEL, the model takes a video as input and produces force predictions. See Figure 5. We then transfer those learned representations to the task of action recognition. Here action recognition takes video as input, and produces a distribution over action categories as output.

Due to computational constraints, rather than jointly pretraining on K710/Ego4D and FEEL, we instead initialize from K710 and Ego4D pretrained weights, fine-tune over FEEL's force data, and transfer the resulting force-aware representations to action recognition.

Downstream Action Recognition Forces are most impactful for the understanding of *physical actions* performed in video. Therefore, after pretraining these video networks over FEEL, we focus our evaluation over downstream datasets centered on manipulation between hands and objects. The datasets

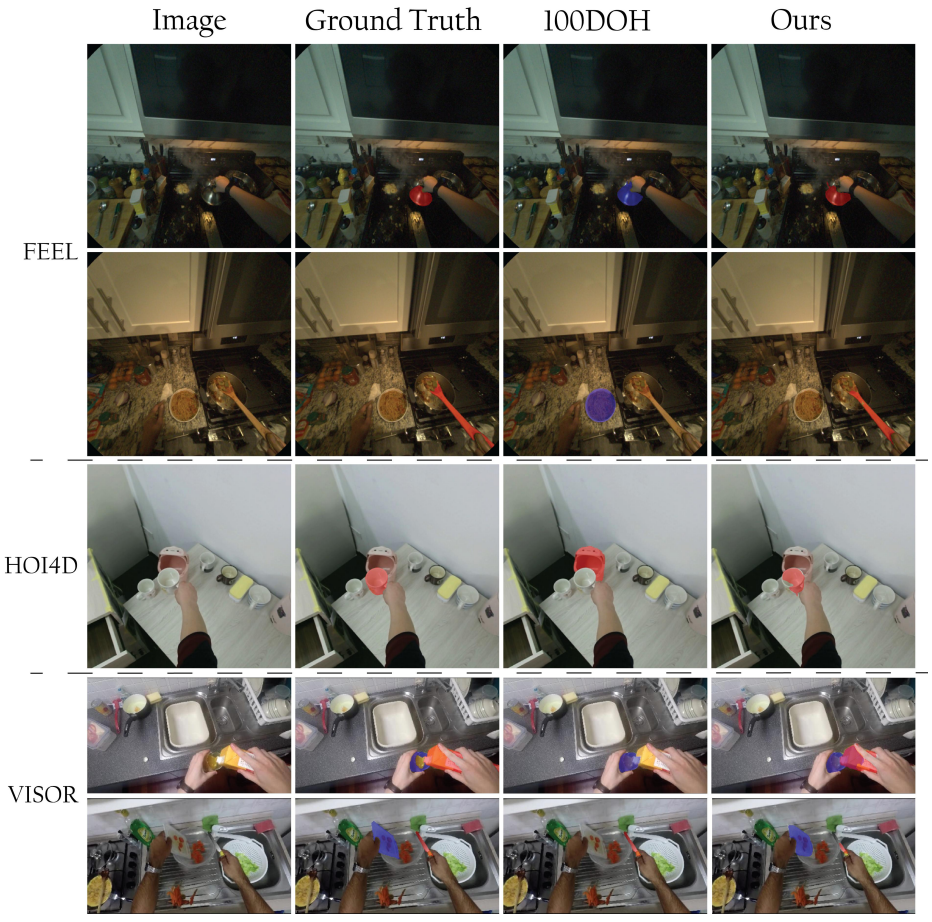

Fig. 7: Qualitative contact segmentation results. Each row shows an input image, ground truth, 100DOH prediction, and our prediction across FEEL, HOI4D, and VISOR. Right and left hand contacts are shown in red and blue respectively. Our model correctly identifies hand side (row 1), avoids confusing proximity for contact (rows 2-3), and handles simultaneous two-hand contacts (rows 4-5). Where our model produces partial segmentations (row 3), 100DOH (combined with SAM) confidently segments the wrong object entirely.

we evaluate over are EPIC-Kitchens-100, SomethingSomething-V2, and Ego-Exo4D. We evaluate each model over the task of action recognition - in the case of EPIC-Kitchens, predicting *verbs* and *nouns*, whereas in Ego-Exo4D, SomethingSomething-V2 and Meccano the task is to predict *actions*. For results, see Table 1.

Table 4: Action recognition results across all benchmarks. Top-1 accuracy for frozen and unfrozen settings across EPIC-Kitchens (verb/noun), SSV2, Ego-Exo4D, and Meccano. For each model, we compare standard pretraining (K710 for Hiera-B, Ego4D for EgoVideo-B) against the same weights further pretrained on FEEL's force data. Bold denotes the better result within each model pair. Models pre-trained over FEEL's data generally outperform the models pre-trained solely over Kinetics and Ego4D. We particularly observe this in the frozen evaluation setting compared to the unfrozen, end-to-end training.

	Model	Pretrain	EPIC-Kitchens		SSV2	Ego-Exo4D	Meccano
			Verb	Noun	Action	Action	Action
Frozen	Hiera-B	K710	41.65	15.12	36.08	13.11	26.42
	Hiera-B	Ours	**46.50**	**21.38**	**39.04**	**20.83**	**29.57**
	EgoVideo-B	Ego4D	59.24	29.42	49.61	26.00	36.59
	EgoVideo-B	Ours	**62.00**	**30.11**	**54.34**	**27.75**	**38.23**
Unfrozen	Hiera-B	K710	**68.34**	48.18	68.88	35.25	38.52
	Hiera-B	Ours	68.24	**49.08**	**70.01**	**41.19**	**41.13**
	EgoVideo-B	Ego4D	70.99	**55.67**	74.55	**40.92**	48.99
	EgoVideo-B	Ours	**71.34**	55.60	**74.91**	40.75	**50.01**

5 Discussion

Temporal Contact Detection Qualitative results show in Figure 6. Note that our method accurately localizes the precise point in time at which contact is established, and is not confounded by mere hand / object proximity, correctly transitioning from no-contact to contact exactly at the onset of grasp formation. Table 2 shows that our force-supervised model substantially outperforms the zero-shot depth baseline and surpasses fully supervised competitors across all three datasets, despite never receiving manually annotated contact labels. The most competitive baseline is 100DOH [32], which benefits from training on two large, fully labeled datasets — the original 100DOH dataset and EPIC-VISOR — to achieve strong generalization to egocentric video. Our model surpasses the 100DOH model while relying solely on physical force measurements for supervision.

A further advantage of our approach is its handling of handedness. Our model experiences a smaller drop between the Binary Contact (BC) and Binary Signed Contact (BSC) metrics than 100DOH, reflecting more reliable left/right hand discrimination. The larger BC–BSC gap in 100DOH may stem from annotation noise: distinguishing hand side is subtle and error-prone for human annotators, potentially introducing label inconsistencies that degrade signed predictions. For ambiguities in annotations of handedness, see the Supplementary Materials.

Contact Segmentation Table 3 demonstrates that our weakly supervised model achieves competitive spatial segmentation performance across all three datasets. On VISOR, our model slightly trails 100DOH, which is expected given that 1) the 100DOH model relies on SAM3 [5], a powerful segmentation model trained over Internet-scale data, and 2) the 100DOH model is fully trained on

the EPIC-VISOR training split while our model is deployed zero-shot. Despite this disadvantage, the gap is narrow.

Our model correctly identifies hand side, avoids confusing proximity for contact, and handles simultaneous two-hand contacts. While 100DOH — which relies on bounding box predictions followed by SAM [5] — produces complete but occasionally over-confident masks of the wrong object, our end-to-end model more consistently identifies the correct contacted object, sometimes at the cost of partial coverage. See Figure 7 for examples.

Action Representation Learning Table 1 reveals two consistent trends across all four downstream benchmarks across Hiera [31] and EgoVideo [28] models.

Force pretraining improves frozen representations more than fine-tuned ones. Performance gains are consistently larger in the frozen-eval setting than under end-to-end fine-tuning. This indicates that force-aware features are more immediately transferable than conventional appearance-based pretraining weights, and suggests that force supervision instills a qualitatively different representational prior.

Gains are largest on smaller downstream datasets. The four evaluation datasets span a wide range of sizes (Meccano smallest, Ego-Exo4D second smallest, EPIC Kitchens second largest, and SomethingSomething-V2 largest), and the relative performance benefit of force pretraining is most pronounced at the smaller end of this spectrum. This is consistent with the hypothesis that physically grounded representations reduce the amount of labeled downstream data needed to learn effective action classifiers — a practically valuable property given the cost of large-scale action annotation.

Limitations. On the hardware side, rubber gel pad mounts were uncomfortable for extended wear, individual sensors were sometimes unreliable due to mechanical drift and bending, and the sensors obstruct the fingertips during tasks requiring fine dexterity - limitations that could be addressed by moving to above-site sensing such as wrist-mounted EMG devices. On the algorithmic side, our pseudolabel generation relies on object motion, meaning we cannot segment stationary contacted objects such as tables or heavy appliances, which will require cues beyond motion to resolve.

Summary Force is the underlying cause that drives physical interaction, it is a critical primitive for physical action understanding. We release FEEL the largest egocentric force–video dataset, capturing the physical causes - not just the visual effects - of hand-object manipulation. We empirically demonstrate the utility of feel over both Contact Understanding, and Action Representation Learning without any need need for manual action label or object segmentation annotations. We hope FEEL serves as a foundation for future work spanning finer-grained hand understanding, force-aware visuomotor control, and manipulation policy learning from human video; and that as wearable sensing hardware continues to improve in form factor and comfort, the community will scale datasets like FEEL to further close the gap between the richness of physical interaction and what can be observed from video alone — with applications reaching from robotics and human-robot interaction to virtual and augmented reality.

References

1. Adeniji, A., Chen, Z., Liu, V., Pattabiraman, V., Bhirangi, R., Haldar, S., Abbeel, P., Pinto, L.: Feel the force: Contact-driven learning from humans. arXiv preprint arXiv:2506.01944 (2025)
2. Bansal, S., Wray, M., Damen, D.: Hoi-ref: Hand-object interaction referral in egocentric vision. arXiv preprint arXiv:2404.09933 (2024)
3. Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second. arXiv preprint arXiv:2410.02073 (2024)
4. Brahmbhatt, S., Tang, C., Twigg, C.D., Kemp, C.C., Hays, J.: Contactpose: A dataset of grasps with object contact and hand pose. In: European Conference on Computer Vision. pp. 361–378. Springer (2020)
5. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
7. Chen, Y., Dwivedi, S.K., Black, M.J., Tzionas, D.: Detecting human-object contact in images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17100–17110 (2023)
8. Cheng, T., Shan, D., Hassen, A., Higgins, R., Fouhey, D.: Towards a richer 2d understanding of hands at scale. Advances in Neural Information Processing Systems **36**, 30453–30465 (2023)
9. Chi, H.G., Barreiros, J., Mercat, J., Ramani, K., Kollar, T.: Multi-modal representation learning with tactile data. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 9660–9667. IEEE (2024)
10. Darkhalil, A., Shan, D., Zhu, B., Ma, J., Kar, A., Higgins, R., Fidler, S., Fouhey, D., Damen, D.: Epic-kitchens visor benchmark: Video segmentations and object relations. Advances in Neural Information Processing Systems **35**, 13745–13758 (2022)
11. Dave, V., Lygerakis, F., Rueckert, E.: Multimodal visual-tactile representation learning through self-supervised contrastive pre-training. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 8013–8020. IEEE (2024)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
13. Dessalene, E., Devaraj, C., Maynord, M., Fermüller, C., Aloimonos, Y.: Forecasting action through contact representations from first person video. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(6), 6703–6714 (2021)
14. Ehsani, K., Tulsiani, S., Gupta, S., Farhadi, A., Gupta, A.: 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. pp. 224–233 (2020)
15. Engel, J., Somasundaram, K., Goesele, M., Sun, A., Gamino, A., Turner, A., Talattof, A., Yuan, A., Souti, B., Meredith, B., et al.: Project aria: A new tool for egocentric multi-modal ai research. arXiv preprint arXiv:2308.13561 (2023)

16. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: Arctic: A dataset for dexterous bimanual hand-object manipulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12943–12954 (2023)
17. Fermüller, C., Wang, F., Yang, Y., Zampogiannis, K., Zhang, Y., Barranco, F., Pfeiffer, M.: Prediction of manipulation actions. International Journal of Computer Vision **126**(2), 358–374 (2018)
18. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18995–19012 (2022)
19. Higgins, R.E., Fouhey, D.F.: Moves: Manipulated objects in video enable segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6334–6343 (2023)
20. Leonardi, R., Furnari, A., Ragusa, F., Farinella, G.M.: Are synthetic data useful for egocentric hand-object interaction detection? In: European Conference on Computer Vision. pp. 36–54. Springer (2024)
21. Li, Z., Sedlar, J., Carpentier, J., Laptev, I., Mansard, N., Sivic, J.: Estimating 3d motion and forces of person-object interactions from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8640–8649 (2019)
22. Lin, F., Price, B., Martinez, T.: Ego2hands: A dataset for egocentric two-hand segmentation and detection. arXiv preprint arXiv:2011.07252 (2020)
23. Liu, S., Tripathi, S., Majumdar, S., Wang, X.: Joint hand motion and interaction hotspots prediction from egocentric videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3282–3292 (2022)
24. Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., Yi, L.: Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21013–21022 (2022)
25. Luong, Q.T., Deriche, R., Faugeras, O., Papadopoulo, T.: On determining the fundamental matrix: Analysis of different methods and experimental results. Ph.D. thesis, Inria (1993)
26. Narasimhaswamy, S., Nguyen, T., Nguyen, M.H.: Detecting hands and recognizing physical contact in the wild. Advances in neural information processing systems **33**, 7841–7851 (2020)
27. Pan, C., Wang, C., Qi, H., Liu, Z., Bharadhwaj, H., Sharma, A., Wu, T., Shi, G., Malik, J., Hogan, F.: Spider: Scalable physics-informed dexterous retargeting. arXiv preprint arXiv:2511.09484 (2025)
28. Pei, B., Chen, G., Xu, J., He, Y., Liu, Y., Pan, K., Huang, Y., Wang, Y., Lu, T., Wang, L., et al.: Egovideo: Exploring egocentric foundation model and downstream adaptation. arXiv preprint arXiv:2406.18070 (2024)
29. Pham, T.H., Kyriazis, N., Argyros, A.A., Kheddar, A.: Hand-object contact force estimation from markerless visual tracking. IEEE transactions on pattern analysis and machine intelligence **40**(12), 2883–2896 (2017)
30. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)
31. Ryali, C., Hu, Y.T., Bolya, D., Wei, C., Fan, H., Huang, P.Y., Aggarwal, V., Chowdhury, A., Poursaeed, O., Hoffman, J., et al.: Hiera: A hierarchical vision

transformer without the bells-and-whistles. In: International conference on machine learning. pp. 29441–29454. PMLR (2023)
32. Shan, D., Geng, J., Shu, M., Fouhey, D.F.: Understanding human hands in contact at internet scale. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9869–9878 (2020)
33. Shan, D., Higgins, R., Fouhey, D.: Cohesiv: Contrastive object and hand embedding segmentation in video. Advances in neural information processing systems **34**, 5898–5909 (2021)
34. Shiota, T., Takagi, M., Kumagai, K., Seshimo, H., Aono, Y.: Egocentric action recognition by capturing hand-object contact and object state. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6541–6551 (2024)
35. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
36. Singh, H.G., Loquercio, A., Sferrazza, C., Wu, J., Qi, H., Abbeel, P., Malik, J.: Hand-object interaction pretraining from videos. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 3352–3360. IEEE (2025)
37. Sivakumar, A., Shaw, K., Pathak, D.: Robotic telekinesis: Learning a robotic hand imitator by watching humans on youtube. arXiv preprint arXiv:2202.10448 (2022)
38. Song, Y.R., Li, J., Fu, R., Murphy, D., Zhou, K., Shiv, R., Li, Y., Xiong, H., Owens, C.E., Du, Y., et al.: Opentouch: Bringing full-hand touch to real-world interaction. arXiv preprint arXiv:2512.16842 (2025)
39. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: Grab: A dataset of whole-body human grasping of objects. In: European conference on computer vision. pp. 581–600. Springer (2020)
40. Tripathi, S., Chatterjee, A., Passy, J.C., Yi, H., Tzionas, D., Black, M.J.: Deco: Dense estimation of 3d human-scene contact in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8001–8013 (2023)
41. Wang, Y., Lipson, L., Deng, J.: Sea-raft: Simple, efficient, accurate raft for optical flow. In: European Conference on Computer Vision. pp. 36–54. Springer (2024)
42. Yu, K., Han, Y., Wang, Q., Saxena, V., Xu, D., Zhao, Y.: Mimictouch: Leveraging multi-modal human tactile demonstrations for contact-rich manipulation. arXiv preprint arXiv:2310.16917 (2023)
43. Zhang, X., Bhatnagar, B.L., Starke, S., Petrov, I., Guzov, V., Dhamo, H., Pérez-Pellitero, E., Pons-Moll, G.: Force: Physics-aware human-object interaction. In: 2025 International Conference on 3D Vision (3DV). pp. 1473–1486. IEEE (2025)
44. Zhu, Y., Jiang, C., Zhao, Y., Terzopoulos, D., Zhu, S.C.: Inferring forces and learning human utilities from videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3823–3833 (2016)