FEEL

Force Estimation from Egocentric Learning

CONTROLLED PROBES FOR GROUNDED FORCE PREDICTION

Do Embodied AI Models Have Physical Intuition?

An Egocentric Probing Benchmark

Yilong Zhu[1,2], Xuyang Li[3], Huihua Zhu[4], Jianhao Jiao[5], Shuyang Zhang[4]

[1] The Hong Kong University of Science and Technology · [2] PaXini Tech
[3] Xi'an Jiaotong University · [4] Shenzhen University · [5] The Hong Kong Polytechnic University

ROBOTICS SCIENCE AND SYSTEMS
RSS 2026 | INTERNATIONAL CONVENTION CENTRE AND UNIVERSITY OF TECHNOLOGY SYDNEY
13 - 17 JULY 2026 | AUSTRALIA

WORKSHOP

Project Website
submission.anonymoussubmission.com

FEEL Dataset
hf.co/datasets/FeelAuthors/FEEL-Benchmark

0.763
Overall R^2
Best across 9 evaluated models

4.67 N
Mean Absolute Error
0.56 N lower than ACT

44 M
Parameters
2–95× fewer than larger baselines

7 / 9
Best Material MAE
Top-2 on 8 of 9 conditions

CENTRAL FINDING Scaling semantic and action pre-training does not automatically yield continuous force intuition; compact, physics-aware inductive biases do.

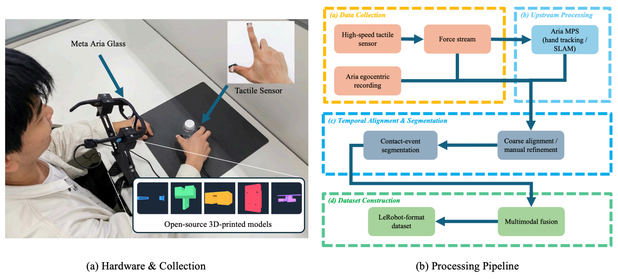
01 MOTIVATION
Physical intuition begins before contact

Expected heavy → Actually light → Force overshoot

Humans use appearance to predict weight, stiffness, and safe grasp force. FEEL asks whether embodied models learn the same continuous physical prior, including its failures under deceptive appearance.

Diagnostic: predict force before tactile correction, then expose the learned prior with a visual-physical mismatch.

02 DATA ENGINE
Reproducible egocentric force probes



(a) Hardware & Collection (b) Processing Pipeline

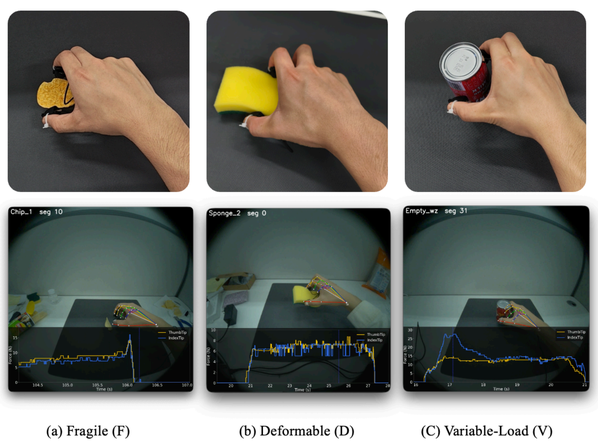
Hardware + pipeline. Meta Aria video and 6-DoF hand pose are synchronized with fingertip tactile force, aligned, segmented, and released as LeRobot episodes.

3.6 h	29	23,627
synchronized capture	raw sessions	labeled frames

For model evaluation, RGB and pose are inputs; synchronized tactile measurements provide the 2-D fingertip normal-force supervision.

Reproducible by design: the release includes 3D-printable mounts, a bill of materials, alignment tools, contact segmentation, and conversion scripts.

03 BENCHMARK DESIGN
Controlled probes, not broad correlations



(a) Fragile (F) (b) Deformable (D) (C) Variable-Load (V)

F / D / V taxonomy. Fragile thresholds, deformable responses, and variable loads demand qualitatively different force strategies.

- **F — Boundary samples** Approach or exceed unsafe force limits, such as crushing a potato chip.
- **D — Water-level priors** Vary fill from 0% to 100% so visible deformation and load change together.
- **V — Visual-physical mismatch** Opaque containers look full but are empty, separating appearance from true mass.

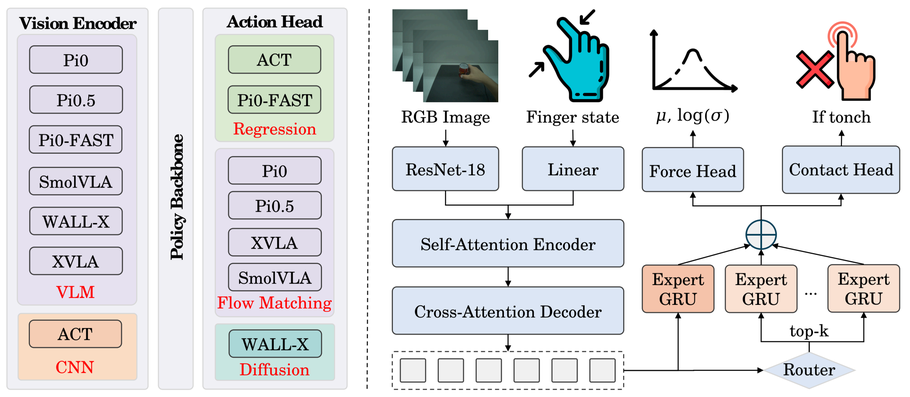
Approach → Grasp → Lift → Carry → Place

317	88	36
train episodes	ID test episodes	OOD episodes

Scope: 7 object families, 9 in-distribution material conditions, and 1 unseen empty-milk category.

The split is stratified at episode level; one expert demonstrator removes inter-subject motor variability from the diagnostic.

04 EGOFORCE
Build the physical assumptions into the model



(a) VLA architecture (b) Our architecture

Architecture. Four RGB frames and a 7-D hand state feed a temporal predictor with uncertainty, contact, and material-adaptive heads.

1. **Temporal context** GRU over 4 observations
2. **Contact phase** Thumb-index distance + auxiliary head
3. **Uncertainty** Clamped Laplace force head
4. **Specialization** 4 experts + shared GRU, top-2

INPUT	SUPERVISION	OUTPUT
4 RGB frames + 7-D hand state	synchronized tactile force labels	2-D fingertip normal force

The 44 M-parameter model focuses capacity on force-relevant structure instead of relying on semantic pre-training scale.

05 LEARNING OBJECTIVE
Uncertainty without scale escape

$$L_{force} = |F_t - \mu_t| / b_t + \log b_t$$

$$\log b_t \le 2.0 \qquad L = L_{force} + 0.1\, L_{contact}$$

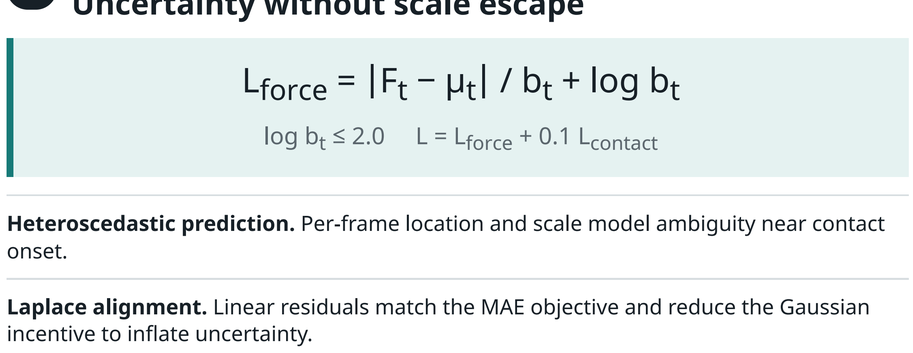
Heteroscedastic prediction. Per-frame location and scale model ambiguity near contact onset.

Laplace alignment. Linear residuals match the MAE objective and reduce the Gaussian incentive to inflate uncertainty.

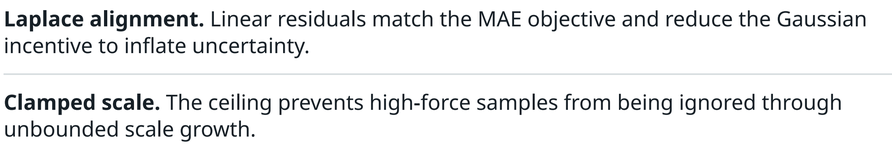
Clamped scale. The ceiling prevents high-force samples from being ignored through unbounded scale growth.

Material-adaptive routing: four specialist GRUs plus one always-active shared expert; top-2 routing is recomputed at every timestep. The shared expert is initialized from the stable single-GRU model.

06 MECHANISM EVIDENCE
What actually moves force prediction?



Temporal behavior. EgoForce follows both the magnitude and time profile across deformable, empty-container, and high-load episodes.

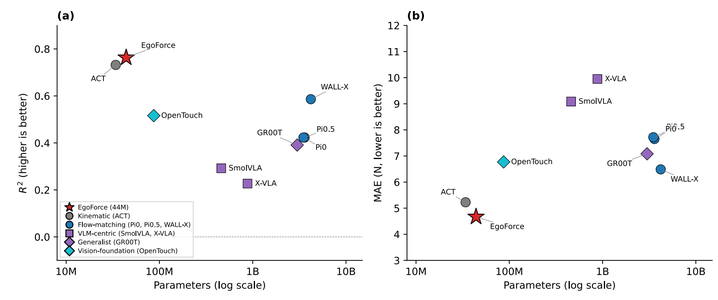


Loss behavior. Laplace NLL + clamp (shown before MoE) balances low-force ambiguity with high-force accuracy.

−0.610 R^2	+0.052 R^2	+0.021 R^2	+0.007 R^2
remove RGB vision	contact auxiliary head	GRU temporal model	MoE-GRU routing

Interpretation: vision supplies the dominant physical cue; explicit contact and temporal objectives turn that cue into calibrated force trajectories, while MoE adds a smaller cross-material gain.

07 MAIN BENCHMARK
Model scale is not force intuition

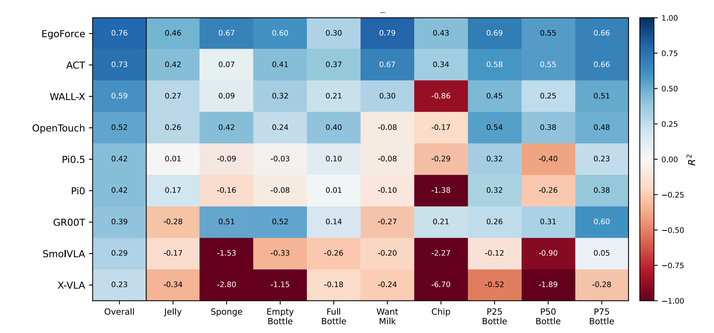


Parameter efficiency. EgoForce is the only model that simultaneously reaches the highest R^2 and lowest MAE.

MODEL	PARAMS	MAE ↓	R^2 ↑
EgoForce	**44 M**	**4.670**	**0.763**
ACT	34 M	5.227	0.732
WALL-X	4.2 B	6.486	0.586
OpenTouch	87 M	6.772	0.516
GR00T	3 B	7.083	0.391
Pi0.5	3.6 B	7.650	0.423
Pi0	3.5 B	7.718	0.423
SmolVLA	460 M	9.086	0.292
X-VLA	880 M	9.949	0.227

All results use 10k optimization steps, 3 seeds, and held-out episodes.

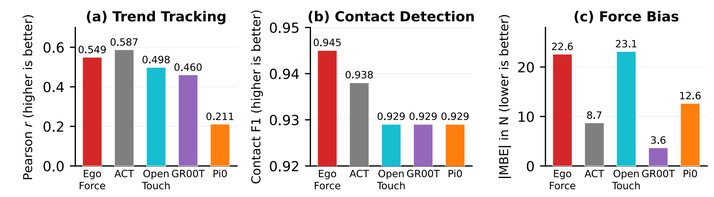
08 CROSS-MATERIAL RESULTS
Consistent gains across force regimes



Per-material R^2. EgoForce and ACT are the only models positive on every condition; cell text shows unclipped values while color is clipped to [−1, 1].

MAE is best on 7/9: OpenTouch leads Full Bottle; ACT leads P75 Bottle by only 0.04 N.

09 OOD COUNTERFACTUAL
Strong visual priors can still mislead



Empty opaque milk. The container looks like the training full-milk condition, but mean ground-truth force falls from about 45 N to 12 N.

0.945	0.549	22.6 N
best contact F1	Pearson r	absolute mean bias

Trade-off: ACT tracks the trend better (r = 0.587) with 8.7 N bias; GR00T has the lowest bias (3.6 N) but weaker trend and contact scores.

Better in-distribution force priors do not guarantee calibrated force under visual-physical mismatch.

10 TAKEAWAYS & LIMITATIONS
Ground force prediction, then stress-test it

1. **Semantic competence is not physical calibration.** Billion-parameter action models trail a compact force-specific baseline.
2. **Inductive bias beats indiscriminate scale.** Temporal, contact, uncertainty, and material structure each add measurable value.
3. **Counterfactual probes are essential.** OOD mismatch reveals the ceiling of vision-only force priors.

Current scope
Offline; one demonstrator; fingertip normal force only.

Open validation
Shear/torque, sensor wear, multi-subject, and cross-embodiment closed loop.

Next test: mismatch-aware calibration with tactile feedback, not stronger visual priors alone.