This Week in Robotics

THE LINE

Last week three groups converged on how a predicted future acts on a policy. **This week four groups independently measured where that whole program stops, and every ceiling they found is about data or representation rather than architecture.**

RoboJEPA (preprint, FAIR at Meta + Mila) fits the first compute-scaling law for a robotic world model across 22M→8B predictors and 2.87M trajectories, and reports that its own fits already sit close to data saturation on a fixed corpus — more parameters will not help, more data might. What 30,000 Hours of Ego-centric Video Does Not Teach (preprint, Toyota Research Institute + UIUC + CMU + JHU) runs a 300h→30,000h ladder and finds agent fidelity saturates early while object-interaction fidelity converges *far below* usable — and that skeleton conditioning at 1,000 hours beats the standard recipe at 30,000. LeWAM (preprint, NVIDIA) shows that sampling-based MPC over raw actions collapses **every** world model tested, its own included, to 3–8% closed-loop success, because the planner finds and exploits the dynamics model's error. Long-WAM (preprint, NVIDIA + MIT + HKU + UCSD) shows that giving a model 19.2 seconds of visual history buys 15 points for an autoregressively pretrained backbone and *nothing at all* for a bidirectional one — access to history is not the same as the ability to use it.

The common shape: the architecture debate of the last two quarters was mostly measuring the initialization and the representation underneath it. inference: that is also why this week's two best-evidenced results are not models at all but **data-collection policies** — Mulligan (§6) and A Balanced Data Diet both ask *where should the episode start*, one on a real arm with an operator and one at a million parallel environments, and both beat uniform sampling by large margins.

Second story, and the week's one genuine artifact surprise: cheap full-hand tactile arrived twice — TACROSS at **USD 10.86 for 285 sensing locations** and SoTa at **under USD 10 of materials for 202 taxels** — and **neither has released anything.** I checked both project pages today; see §2.

Window: 2026-10-05 → 2026-10-11. Queue: 5 part files, ~370 scored lines, 32 at 4+.

Coverage caveats carried from the harvester, and they matter more than usual this week. arXiv was never *enumerated* on any of the five sweeps — all discovery was topic-ranked, so recall is unquantified by construction for cs.RO, and the harvester's own gap-fills prove the leak is large (one sweep found 24 items a previous sweep had missed in the same ID band). The band ceiling has not advanced past 2610.12470 for three consecutive sweeps, so Thursday's and Friday's announcements are an unenumerated backlog, not an empty set. 10-10 and 10-11 were Saturday and Sunday and arXiv does not announce on weekends, so *that* silence is real. GitHub release surfaces returned 403 for the whole week; I resolved MuJoCo and LeRobot through PyPI instead (§4). Job boards and lab blogs were unchecked by every sweep; I checked them today (§8), and found one of the harvester's "broken surfaces" is a wrong-URL problem (§4). Treat silence in any category as unmeasured, not empty.

1 Policy learning & VLAs

RoboJEPA — the first scaling law for a robotic world model, and its own ceiling

arXiv:2610.10515 · preprint · FAIR at Meta + Chandar Lab + Mila + Polytechnique Montréal · code, checkpoints and deployment code **stated released** at `github.com/facebookresearch/robo_jepa`

Architecturally new:
Very little, deliberately. It is a family of deterministic, action-conditioned transformer predictors over a **frozen V-JEPA 2.1** encoder — temporal-causal attention, four-axis RoPE over time, image position and camera view, plus embodiment and camera-identity conditioning so one model eats heterogeneous action and state dimensions and variable camera counts. The training objective is two per-token L1 losses: teacher forcing on the next encoder representation, plus an autoregressive rollout loss from a prefix so the model must consume its own predictions. What is new is the *experiment*: holding the representation fixed and sweeping the predictor from 22M to 8B parameters (≈2×10^19 → 9.5×10^22 FLOPs) to see whether prediction error is forecastable.

It is. The compute-optimal frontier is best fit not by a power law but by a **second-order** one, L(C) = E + A·C^(α − γ·lnC). Fitted on 22M–2B and extrapolated to the held-out 4B and 8B models, it lands at 0.6×10^-3 (DROID) and 1.4×10^-3 (RoboCasa) extrapolation error against 2.0×10^-3 and 3.7×10^-3 for a standard power law. Estimated error floors ≈0.2 and ≈0.17.

What the eval actually covers:
Nine-step latent rollout error on held-out DROID (real Franka scenes) and held-out RoboCasa scenes for the law; then planning. Planning is CEM over action sequences scored by predicted terminal latent distance to **a single goal image** — no language, no intermediate subgoals, no learned action proposal, actions sampled uniformly, receding horizon with a rolling KV cache and planning hyperparameters held fixed across embodiments. >50,000 planning episodes across two platforms. The capability thresholds are the result worth keeping: end-effector reaching becomes effective ≈10^20 FLOPs, moving while holding an object ≈3×10^20, static-obstacle avoidance ≈10^21 (and only after the authors extended the AR training rollout from 2 to 10 steps and raised input resolution), and **object pushing — predicting how contact moves a thing — only at ≈10^22**. Real Franka: 8B at 67% grasp / 50% lift / 27% pick-and-place; 4B at 60/30/21.

Released:
Code, checkpoints and deployment code per the paper. I could not verify the artifact from this session — GitHub returned errors all week and no mirror exists under that name on the HF Hub. Treat as stated, not confirmed.

Reproducible on one arm and one GPU?
Not the 8B. But the ladder is the point, and the 22M–300M rungs are squarely one-GPU; the paper tells you, in FLOPs, which capability you are buying. **INFERENCE** `\2192` if you want object-dynamics-through-contact out of a latent world model, the thresholds say you are shopping at ≈10^22 FLOPs, and nothing below that will show it to you — which is a far more useful planning number than a success rate.

Where it bites:
The encoder is frozen, so this is a scaling law *for a dynamics predictor on one fixed representation*, not for a world model. The authors say the fits already sit near data saturation after multiple epochs on a fixed corpus. There is no text conditioning at all. And the π0-FAST / π0.5 comparisons are explicitly contextual, not controlled — different goal specification, different training. π0.5 still beats it on pick-and-place (≈53% vs 27%).

Long-WAM and OpenWAM — two controlled comparisons, and they agree on the backbone

arXiv:2610.10528 (Long-WAM, preprint, NVIDIA + MIT + HKU + UCSD, code link stated but URL unresolved) · arXiv:2610.07922 (OpenWAM, preprint, Stanford, **code released and verified today**)

Long-WAM's central finding is a clean mechanism result rather than a leaderboard delta. On RoboCasa GR-1, extending visual context from 0.0 to 19.2 s takes **both** autoregressively pretrained foundations from 63.3% to 78.7%, while a bidirectional Wan2.2 initialization shows no net gain at all (61.7 at 2.4 s → 64.1 at 9.6 → 61.6 at 19.2). The AR lead *grows* from 3.3 pp with no history to 17.1 pp at 19.2 s. And optimal context is task-dependent, not monotone: LIBERO-Long saturates at 2.4 s (94.5 → 99.5%) and **regresses to 75.2% at 38.4 s**, where 80.4% of sampled history frames are padding. Separately, predict-then-act via an IDM beats

co-denoising (99.5 / 97.8 / 94.5 on LIBERO-Long) with future latents stopped at σ=*0.9 and never decoded to pixels. The deployment number is the one to care about: 107.4 ms per action chunk on an RTX 5090, including the full observation VAE — a 3.3× gain over BF16 eager via NVFP4, streaming VAE, decision-time prefill and KV reuse, and it also runs on DGX Spark and Jetson AGX Thor. On a Unitree G1 it holds 90–100% on conveyor grasping at 3.0–7.5 cm/s where Fast-WAM falls 75→0% and π0.5 15→0%, and takes 19/20 on dynamic cup stacking where both baselines score 0/20*. Soft: pretraining is 30,720 GPU-hours on 64 H100 and the downstream WAMs train on 16 GB200, so inference is reachable and training is not; 8× history costs 3.2× latency (107.4 → 341.0 ms); the training surrogate conditions on forward-noised ground-truth futures while inference starts from pure noise, which the authors flag as an approximation; no seed counts on the real-robot tables.

OpenWAM does what the WAM literature has not: it fixes one causal backbone (Wan2.2-5B plus 10,000+ h of causal robot-video pretraining with **no action labels**) and varies *only* generation order and cross-modal visibility, via a Mixture-of-Transformers action expert that serves joint, video-then-action, action-then-video and decoupled generation **off one shared checkpoint**. LIBERO four-suite means over three seeds: VTA 98.6, Decoupled 98.0, Joint 97.6, ATV 96.8 — i.e. the interaction-order question everyone has been arguing about is worth under two points. The backbone ablation is worth thirty times more: Wan2.2 init 68.4 → robot-video-pretrained 97.8 on LIBERO-Long (+29.4 for VTA, +34.4 for Joint), with MoT worth a further 5.0/3.0 pp over a shared DiT. The reusable part is component surgery: with the visual predictor task-adapted and the action component **frozen**, a local-context IDM trained on mixed counterfactual+demo data reaches 84.0% on four held-out LIBERO-90 tasks against 47.0% full-context and 21.5% local-context-demos-only, while retaining 94.4% source-task success. Counterfactual FDM supervision cuts RGB MSE 34.5% and lifts 16-way outcome identification 21.1% → 71.3%. Real: three bimanual Franka Research 3 tasks at 92.1% (VTA).

Verified today:
The project page confirms `github.com/OpenWAM/OpenWAM` with training, evaluation and composition code plus a quickstart, and pretrained video weights at OpenWAM-Stanford/OpenWAM-Pretraining (exists, updated 2026-10-09, 2 likes). **No licence is stated anywhere**, and no policy or dynamics checkpoints are linked — only the pretraining weights. The paper itself flags a **concurrent, unrelated framework also called OpenWAM**; expect the name collision when you search. Not reproducible at one-GPU scale: 5B backbone, 10k+ hours. What is reusable is the checkpoint, the four interaction programs and the counterfactual-relabelling recipe. Note also that the counterfactual data is simulator-branched, so the IDM-reuse result has no real-robot counterpart yet, and the unified multi-objective checkpoint scores 92.8 against 98.6 for the policy-specialized one.

LeWAM — the negative result to read before you build a planner

arXiv:2610.12407 · preprint · NVIDIA · no code statement

Sampling-based MPC over **raw actions** collapses every model tested — LeWAM and RecWAM included — to **3–8% closed-loop success**, because CEM finds the regions where the dynamics model is wrong and optimizes for hallucinated futures. Moving the decision variable into the deterministic noise space of a flow-matching action head (diffusion steering) recovers and then *exceeds* the policy. The projection step is load-bearing and they show it: without projecting proposals back onto the √A noise shell, DS-CEM on robomimic Can goes 78.8 → 23.0, Square 87.2 → 75.0, Lift 97.6 → 93.0. The model itself is a decoder-free **41M** JEPA WAM — ViT-tiny/14 DINOv2-small architecture trained from scratch at 224 px, SIGReg on a single 384-D CLS latent, no EMA target, no stop-gradient — with one bidirectional transformer (depth 8, width 512, 16 heads) serving forward, backward, inverse and policy modes purely by input masking, every mode on every batch. Removing the policy mode destroys planning entirely (0.0–0.8%). Linear probes: task-state R^2 0.80 against LeWM 0.62, Dreamer 0.68, RecWAM 0.78 — while scoring **0.00 on action-independent distractors** where RecWAM scores 0.73, i.e. it matches reconstruction-based probe accuracy without absorbing the distractors. Evaluation is PushT plus robomimic Lift/Can/Square/ToolHang/Transport with two painted or wobbling distractors, 5 training × 50 evaluation seeds = **250 rollouts per cell**, 300 epochs, baselines reimplemented and size-matched to ~41–43M under one shared protocol. That is the most honest protocol in this week's queue. Soft: simulation only, 200–206 human demos per task, authors name real-world data as the next step; planning is slow and requires a goal latent at inference, retrieved from a demonstration 25 steps ahead; Dreamer has no noise-to-action map so it gets no DS variant and that comparison is asymmetric. 41M at 300 epochs is genuinely one-GPU reproducible **if** code appears.

RACE — where chunk errors actually live

arXiv:2610.05719 · preprint · KAIST + GIST + NVIDIA + POSTECH · code stated on GitHub but **no URL in the paper**

The claim is narrow and testable: action-chunk errors concentrate at **subskill transitions**, so predicting *when* a transition is coming lets you execute ~4× more of each chunk safely. π0.5 backbone, VLABench 33.0 against a fine-tuned π0.5 at 28.1 with 20 executed actions; real AgileX PiPER pick-and-place at 66% vs 48% at the same chunk length from 30 demos over 50 trials; and with asynchronous real-time chunking, idle time 2.11 s → 0.41 s. Training is post-training of the action expert only — 20–40K steps, batch 64, four RTX A6000s — so this is close to a one-GPU recipe. Soft: transition labels are PELT change-point **pseudo-labels**, not annotations; success declines steeply past 20 executed actions and falls *below* baseline at 40, so the window is real but bounded; gains shrink on VLAs without a denoising process; the real-robot study is one pick-and-place task with no seed count.

Also in this cluster and worth a line each: Juno (preprint, USTC + ShanghaiTech, no code) makes the sharp observation that **failed rollouts are poor policy targets but their consequences are still valid dynamics supervision** — it adapts a 15M JEPA world model on all observed transitions, freezes it, then re-aligns the policy on verified executions only, getting a real PiPER from 40% → 100% in-domain on 60 demos and 70–75% under background, height and combined shifts where the baseline collapses to 0/20. Its components are tightly coupled, though: removing fusion alone drops *below* the base VLA (53.9 vs 60.9). PerturBot (preprint, Zhejiang + HKU, **code released**) names modality shortcuts, breaks them with data interventions and ships an offline shortcut-reliance diagnostic. RealtimeWAM (preprint, NTU + Beihang + SenseTime, **code and checkpoints released**) distils a WAM to one-step action generation at 12.2 ms on an H100 — but 27.2 ms on an RTX 4090D, roughly half the headline speedup at consumer scale, and its real-world evaluation is **qualitative only** (a T-shirt-folding rollout with no success rate).

---

2 Manipulation, dexterity & sensing

Full-hand tactile got cheap twice, and shipped zero times

TACROSS, arXiv:2610.11945 · preprint · HKU + CUHK-Shenzhen + Ocean U + Wuhan + INFIFORCE · SoTa, arXiv:2610.02338 · preprint · Stanford + Tsinghua + Cambridge

TACROSS is a **USD 10.86** glove — controller 6.58 plus materials 4.28 — carrying **285 sensing locations** (12×15 palm plus five 3×7 finger arrays) over ~200 cm² at >85% coverage, 256 values per frame at ~150 Hz through a 12-bit ADC. That is **USD 0.038 per sensing location**, against the authors' figures of 0.76 for FlexGlove, 6.83 for Wuji and 28.65–42.98 for a Tekscan Grip 4256E. It is characterized, which is rarer than it is cheap: 0–20 N with no saturation (peak 83.91/255 ADC), repeatability 1.38% FS, nonlinearity 8.26% FS at R^2=0.90, hysteresis 11.65% FS, drift 0.39–0.41% FS per 10 min, <50 g glove and <85 g system, donning under 30 s, wear over 2 h. The method is the interesting half. It aligns **contact events, not raw channels**, through per-sensor canonicalizers with explicit validity masks — so a missing measurement is distinguished from a measured zero — plus residual adapters and a shared causal temporal-and-inter-finger transformer into a 256-D latent. And robot actions are the **sole** ground truth: retargeted human hand targets enter only as confidence-weighted, validity-masked pseudo-labels. 30 robot + 150 human demos gives 91.9% mean over four contact-rich tasks against 51.3% robot-only (+40.6 pp), falling to 73.1% without human tactile and 49.4% without contact-semantic alignment. The acquisition economics are the part to steal: USD 12,000 robot-only at 40 episodes/hour, USD 2,150 Rokoko-Ego at 112/hour, USD 520 Vision-IK-Ego at 140/hour.

SoTa comes at the same target from the materials side: **under USD 10 of materials per skin**, USD 30 of readout at quantity 100, **202 taxels** in an *identical layout on the human hand and the robot hand*, so there is no learned cross-sensor mapping to go wrong. 20 mN–10 N, 97.31% of span retained after 10,000 cycles, 1,280 fist-folding cycles with continuity intact, 11.05% hysteresis against STAG's 17.5%, and it beats copper-polyimide (80 cycles) and silver-TPU (640) on durability by one to two orders of magnitude. Fabrication is 6 h of laser work plus 2 h of assembly; readout is an ESP32-S3 plus PCAP04, 7 h untethered at 30 fps. On Kinova + Wuji hand with an FTP-1 backbone and 150 human demos per task, three tasks and 40 trials per condition: 22.8 → 45.9% mean over eight conditions, improving all five OOD conditions, though the authors state the tactile-specific OOD gains are **mixed**. No seeds stated, one camera and no wrist view, one training object per task.

I checked both project pages today. Neither has released anything. TACROSS's page states that hardware designs, software and the 183.95-hour dataset are *planned* for release, with "repository link to follow" and "download link to follow" and **no licence named** — which corrects the harvester's queue line, logged as "design files stated available." SoTa likewise promises gerbers, firmware and the fab workflow and has delivered none. So the cheapest credible full-hand tactile hardware in this ledger's history is, as of today, two PDFs. The one released ingredient in either paper is AnySkin, which already existed.

Factorized tactile, and the sim ranking that lies to you

arXiv:2610.10510 · preprint · Amazon Frontier Technology Research + UT Austin · no code

Splitting touch into **geometry, force and temporal change** beats learning one representation for all three — a direct answer to the negative result TACTIC left standing in W40 (no universally optimal tactile representation). TSF-85 capacitive taxels on a UR5e with a Robotiq 2F-85, trained in Isaac Lab, 0.48 mm contact localization and 1.69 N force MAE, peg insertion at 63%, with a gated policy that tolerates masking a modality at test time without retraining. Two findings worth keeping beyond the method: **centre-of-pressure features performed worse than proprioception alone**, and **the ranking of representations in simulation did not match the ranking on hardware** — if you are selecting a tactile encoder on a sim benchmark, that is a direct warning. Single lab, no code, A100 training.

EyeRobot 2.0 — active fixation instead of wrist cameras

arXiv:2610.03710 · preprint · UC Berkeley + Amazon FAR + Stanford · code promised

One fixed stereo camera mounted between the arms, no wrist cameras at deployment. Stereo images are re-projected into two converging eye views on a shared 3D fixation point; observations are foveated, multi-resolution centred crops encoded independently as visual tokens over a frozen DINOv3 ViT-S/16; and — the representational move — **end-effector pose and actions are expressed in a fixation-relative SE(3) frame** with fixation distance as a depth anchor. Gaze is learned in two stages without any gaze labels: a goal-conditioned fixation policy trained with PPO on teleoperation frames (SAM3 masks, FoundationStereo depth, reward on distance to the object centroid; ~3.5 cm mean 3D fixation error), then a target *selector* trained in an RL–BC loop whose reward is how well a BC policy predicts demonstrated action chunks under the selector's fixations, scored by Fréchet distance on the more active arm. No manual task-stage annotation anywhere. 1,278 demonstrations, ~288 minutes of teleoperation, 14-DoF bimanual YAM, >1,000 physical trials and 1,800 simulated, 25 configurations per policy per physical task. Results: ≈+40 pp over a gaze-free static stereo baseline in the real world and ≈+20 in simulation, concentrated in precision stages — straw insertion 8% → 44%, marker insertion 4% → 68%. Against ego-plus-wrist it is **comparable when wrist views are clear** (69% vs 64%) and much better when a held object occludes them (48% vs 22%). Ablations: no foveation −20 pp, no stereo −18 pp, world-frame instead of fixation-relative actions −21 pp in sim. Soft: policies are trained per task; an adequate fixed-camera placement is assumed and no head or neck is controlled; gaze targets are object-level; and the authors observe cases where the policy fixates a *static* object as a visual anchor rather than the manipulated one, which is a reminder that the learned gaze is a means to a BC-prediction reward, not an interpretable attention signal. **INFERENCE** \2192 for a one-arm rig, the transferable claim is the fixation-relative action frame plus foveated crops, both of which you can test on an existing static camera without buying anything.

EigenDEXplore — structure the noise, not the action space

arXiv:2610.07681 · preprint · Stanford · **code released, plus pretrained PCA models for six hands**

Across six dexterous hands the paper separates two things the literature confounds: putting human coordination in the **action space** versus using it to **structure exploration noise**. It is the second that pays. The change is one line — add E^T(σ_e ⊙ η_e) to joint-space noise, keep joint-space actions, discard the eigenbasis at deployment — and because the covariance is just diag($σ_q^2$) + E^T diag($σ_e^2$) E, PPO log-probs and entropy need no custom machinery. +34% consecutive successes early in SimToolReal, +40% in DeXtreme; real hardware SimToolReal task progress 33% → 82% over four tool-use tasks; real DeXtreme mean consecutive successes 8.45 → 11.90 over 40 trials with longest run 32 → 41; DexMachina object tracking +28% with **no contact reward shaping and no curriculum**. Two controls make it credible: random orthonormal directions with the same eigenvalue

spectrum do *worse* than plain joint-space noise, so the gain is the human coordination pattern and not correlation per se; and eigen-noise alone is worse than joint-space, so the independent joint noise is load-bearing. Large-scale egocentric video (100k hours) beats GRAB and ARCTIC as the prior source. Soft: five seeds and 95% CIs in simulation, but the real-robot comparisons are **one policy per method** — no seed spread on hardware. Gains shrink on low-DoF hands. The basis is fixed and linear, and noise is sampled independently per timestep.

The cheap force results

Three items, all worth more than their scores suggest. Privileged noise distillation (preprint, NTHU + NYCU, code promised) contains this week's cleanest matched study: in FORGE the same estimated fixture pose *both* enters the observation and anchors the action frame, so the offset is unidentifiable from proprioception before contact, and at σ=5 mm the official state policy falls from 97–99% to 31.8/49.9/60.4% on peg/gear/nut. The fix needs no privileged teacher at all — collect clean demos at σ=0, draw an offset, relabel in closed form (subtract ε from the first three observation dims and ε/b from the first three action dims, b=0.05 m), leaving images and wrench untouched because the physical geometry did not change. Identical student, data budget and trainer, changing **only** demonstration source: 31.5 ± 2.2% (non-privileged) vs 88.7 ± 7.1% (privileged teacher) vs **95.8 ± 1.0% (relabelled)** at σ=5 mm, three seeds, disjoint Wilson intervals, with the evaluation protocol frozen before any result and checkpoints selected on held-out BC loss rather than success. Student is 26.5M params, two ResNet-18 trunks plus a 32×6 raw wrench window at 120 Hz, 7-dim action at 15 Hz, p95 3.4 ms on an **RTX 3090**. The modality ablation is the calibration to keep: vision is worth 34–66 pp at σ=5 mm, while **force is worth 12/16/18 pp on the 0.114 mm-clearance peg, 5–8 on the nut and 0–1 on the gear** — a concrete answer to when force earns its place. Soft: translation only, cross-geometry generalization *fails* (4.6% vs 18.4% for a noise-augmented baseline), and nothing is released yet. PEARS (preprint, HKU + SJTU + Xense, no code) puts a VLM *between* episodes rather than in the loop: it reads the visual outcome plus tactile and wrench history after each episode, names the failure cause, and emits updated contact-force bounds that a high-frequency hybrid force-position controller enforces, on a **frozen** GR00T-N1.7. Real: whiteboard erasing 11/20 → 19/20 and pipette aspiration 8/20 → 18/20 after 40 online episodes, with zero force-related failures remaining. VLM latency is 26.8 s per episode, hidden behind resets. Real-world success is labelled manually by a human operator; single lab. Tactile reconstruction of contact task frames and forces (preprint, Sapienza + Twente, abstract-unverified) recovers interaction force *and* the task frame from tactile alone with no F/T sensor — the cheapest entry point in the hybrid force/motion cluster if it holds up.

And the negative result to keep: SimpleTouch (preprint, Tsinghua + Alibaba, no code) reports a **single-stage** tactile expert on a frozen π0.5 beating two-stage pipelines, i.e. tactile *pretraining* shown unnecessary. Read against OpenViTac (preprint, Fudan + NUS + Neote, website only), which provides the first paired sim–real visuo-tactile benchmark — 11 simulation tasks with 8 real counterparts sharing task semantics *and* success criteria, three tactile sensors and two grippers in both domains on Isaac Sim 5.1.0 + Isaac Lab 2.3.0 with TacEx — and a quantified co-training recipe: with the real budget fixed at 100 demos per task, adding 500 sim demos moves π0.5 on Insert-USB 3/20 → 8/20 and Gear-Assembly 7/20 → 10/20. Its most directly usable finding: **under lighting perturbation, real-only π0.5 falls to 0/20 and 2/20 while a co-trained visuo-tactile policy holds 6/20 and 12/20, with the Gear-Assembly rate completely unchanged by the lighting change.** Touch is the lighting-invariant component. Soft: 20 trials per cell, absolute numbers low throughout (8/20 is the best Insert-USB figure in the table), grippers only, and the authors state a remaining tactile sim-real gap from rendering and sensor modelling.

---

3 Platforms & embodiment

Demonstrated, with numbers. BiGym 2.0 (preprint, Imperial, **open-sourced** at `github.com/swirl-uk/BiGym2` including environments, demos and evaluation traces) is the week's most useful platform artifact and it is simulation: 1,200 human VR demos across 20 tasks on a 29-DoF Unitree G1, with everything — learned policies *and* coding-agent programs — running through the **same frozen GR00T-WBC whole-body controller**. That is the point: a walking controller lags, oscillates and perturbs the torso, and this measures manipulation under that rather than on a floating base. The reproducibility engineering is worth copying outright: the controller keeps internal state across steps, so a seeded reset alone does **not** reproduce a trajectory (9.5×10^-2 max qpos divergence immediately, 2.1×10^-1 within 50 steps) — restoring the controller snapshot captured at first

reset makes evaluation trajectories **bit-identical**. Findings: π0.5 fine-tuning takes the highest nine-task mean, but **cold-start coding agents (GPT-6 Astra, Claude Opus 5.5) beat every demo-driven RL baseline on that mean** and lead on bimanual reaching; classical tooling (IK, ray projection, rate limiters) lifts the agent mean 53 → 76% but unevenly (+67 on dishwasher closing, nothing on reaching); cross-workspace block stacking is unsolved at 19/0/0%. Agent development cost is itemized at USD 8.7–12.1 per session. Soft: one frozen controller, so the results characterise the joint policy-controller stack; agents see one demonstration where policies see 60; session variance is large (33 ± 33% on one task); and **Claude Opus 5.5 recalled the published G1 kinematics from pretraining**, so agent numbers partly reflect a prior that would not exist for an undocumented robot.

Ultra (press, 10-09) published the rarest thing in this section: **96.4% autonomy over one shift** on a π0.6 model in commercial use, on a stationary dual-arm OP1 on casters (5×5 feet, 10 lb per arm, 120 V outlet) across packing, sorting and kitting, with 500k orders packed and deployments in NY, GA, NJ and TX. It is one shift, not a fleet; no intervention rate, no throughput, no fleet size. Still, it is a published autonomy number on a named foundation model in paid work, which is more than any humanoid vendor offered this week.

Agility's investor day (press, 10-07) disclosed a **~USD 150,000 bill of materials** for Digit 5 against a USD 300M, 1,000-robot order that unlocks **skill by skill**. Vendor self-reported, but a per-unit BOM and a skill-gated contract structure are two numbers nobody else published.

Staged, or video-only. Boston Dynamics' Atlas hand rationale (press, 10-09) adds real design content — thirteen *identical* encapsulated actuators, no cables across joints, backdrivability and transparency prioritised, a two-thumb design rejected, the pinky dropped, over 100 lb retained — and **still zero task success rates, zero tactile specs, zero durability data**, three weeks after the hand was announced. Panasonic targets humanoid production by 2029 and Kawasaki a Kaleido rollout around 2030: dated targets, no method, no specs. UBTECH/FAW-Volkswagen expand Walker S Lite from quality inspection into factory logistics — "develop and test," no autonomy ratio, no robot count, no cycle times. BYD and Tesla appear only as **patent drawings**; the Tesla filing describes curved tactile arrays and the reporting explicitly does not confirm use in the new Optimus hand. Worth noting the directional read anyway: BD's GR3 and Tesla's filing both bet on dense fingertip tactile, and in the same week TACROSS delivered 285 characterized sensing points for USD 10.86.

Low-cost, demonstrated-not-scored. PhoneBot (preprint, UCLA RoMeLa, design and code open) is ~USD 400 excluding the phone against USD 3.5K–13.8K for OP3, ToddlerBot, BRUCE, Berkeley Humanoid Lite and HighTorque MiniPi: 1.8 kg, 483 mm, 13 DoF on thirteen USD 23.90 XL430-W250 servos, with a commodity phone serving as IMU, camera *and* policy compute (TF-Lite on device, UDP to an SBC for the servo bus), validated on four cheap handsets with per-function CPU usage tabulated. **No arms, flat terrain only, and walking, human-following and standup are demonstrated rather than scored — no success rate, no distance, no disturbance-rejection number anywhere.** The transferable content is the torque-aware XL430 actuation model in MuJoCo/MJX (torque-speed envelope, asymmetric braking clamp, back-drive stiffness ratio, encoder backlash on the observed position only) and sagittal mirroring, which concatenates a mirrored actor sample per PPO update: five seeds, joint-angle symmetry error 0.254 → 0.003 rad, torque imbalance 0.205 → 0.034. Also surfaced: Q8botOne, a palm-sized open-source quadruped on ESP32-C3 and DYNAMIXEL actuators — crowdfunding, not shipping, no BOM cost or evaluation read.

Still zero, now seven weeks: **no humanoid company has first-authored a method with released code.** This week came closest — LimX Dynamics + SUSTech on a humanoid world-action model with joint state-action generation, RoboParty co-authoring I-BFM, Astribot co-authoring PLaW-VLA, Minerva Humanoids co-authoring iGPC — and none of the four states a release.

---

4 Open source & tooling

Ranked by what reduces build friction this month.

1. **MuJoCo 3.15.0, released 2026-10-05 — and I confirmed it independently today on PyPI, where 3.15.0 is latest.** This closes the thread SEEN W40 opened with a trustworthy 404 on that tag, and it is the first confirmed simulator release in this ledger's history. **It carries breaking changes**: `dampratio` now uses operational-space inertia rather than the previous basis, which will silently change the behaviour of every tuned `dampratio` in your models. Also capsule multicontact,

Stable Neo-Hookean flex, and >2 GiB model serialization. If you have tuned damping against 3.14.x, pin before you upgrade — and note from W40's benchmark audit that an unpinned MuJoCo version was worth 26 pp on one LIBERO task.

2. **LeRobot is still 0.6.1, and this is no longer "unverified."**
   The harvester recorded five consecutive runs against a `/releases` index that renders 2024 dates, and GitHub returned 403 to the `gh` CLI for the whole week. PyPI answers cleanly: **0.6.1, released 2026-08-03**, with 0.6.0 on 2026-07-06 before it. That is **eleven weeks post-deal with no release**, now from a surface that is actually trustworthy. v0.7.0 remains an open community roadmap issue naming eight workstreams.

3. **Isaac ROS 5.0 (ROS Discourse, 10-09)** — NVIDIA's perception stack now installs with Pixi in one command on both Jetson and ordinary Linux desktops. Install friction, removed. No benchmarks published.

4. **Asset interoperability moved from both sides at once.** Draft REP 158 is an actively contested community standard for OpenUSD simulation-asset conventions (21 replies, draft status, **no ratification and no adoption commitment from any named simulator**), and USDCraft (preprint, NVIDIA + NTU + PKU) proposes programmatic rather than mesh-based articulated assets for real-to-sim. If you currently rebuild the same asset per simulator, this is the pair to watch.

5. **Upcoming breaking changes in ros2_control (ROS Discourse, 10-08)** — the deprecated `Handle` and export APIs are being removed. Any custom hardware interface written against them stops building: a strain-gauge bridge, an ADC readout, a tendon-hand driver. No removal date, no target distribution and no migration guide captured.

6. **Dataset tooling had a genuinely good week.** FiftyOne now reads LeRobot (50 embodiments, 497 episodes, one indexed dataset) and a companion guide to LeRobot dataset quality metrics ships concrete metrics plus a plugin. Also: RoboTwin 2.0 end-effector split and ABC-130k-sim both converted to **current LeRobot v3.0** (the latter apache-2.0, re-rendered at 352×256) — conversion work is unglamorous and it is the thing that saves you a day.

7. **SpaceMouse teleop + MuJoCo visual twin + data collection for an AgileX PiPER (ROS Discourse, 10-08)** — close to a complete one-arm episode-collection stack at exactly the budget CONTEXT.md assumes, on the same arm Juno and RACE use for real evaluation. **No repository URL, no licence, no evaluation, single author, unreproduced.** Worth ten minutes to look at and nothing more until it has one of those.

8. Also: a Rokoko Smartglove II ROS 2 bridge (which makes TACROSS's USD 2,150 middle acquisition tier concretely buildable), knoblog (MIT-licensed tracker for silent default-parameter drift in Nav2/PX4/ArduPilot — it sees committed defaults only, not launch or runtime overrides), and an AWS open-source Physical AI Toolchain (trade press only — no repository URL, no licence, no component list was verified, and cloud-vendor toolchains historically assume cloud-scale budgets).

Unverified, and say so:
Isaac Lab, Genesis, ManiSkill and RoboCasa. Every GitHub surface was 403 for the session and no PyPI cross-check was run on them. No claim either way.

One correction to the harvester's surface list. Three sweeps recorded `physicalintelligence.company/blog` as a broken surface serving zero 2026 posts. The live blog is at **pi.website/blog**, and it is not broken — it genuinely has nothing since *VLAs with Long and Short-Term Memory* on **2026-03-03**. INFERENCE \2192 PI has published no public research post in seven months while shipping π0.6 into a customer's commercial deployment (§3). That is a deliberate change in posture, not a dead web server.

5 Money & moves

Two lines each. The interpretation first: **the money moved one layer downstream, from models to the two things that make models deployable — demonstration data and deployment assurance** — and NVIDIA is on both sides of the week's two largest items.

- **Mecka, USD 60M Series B** (press, 10-07) — Sequoia-led with **NVIDIA and M12**, funding human-demonstration data collection at commercial scale. A compute holder and a strategic fund putting a Series B into *collection capacity* says the bottleneck is priced as data, not architecture — which is exactly what TRI's 30,000-hour negative and RoboJEPA's saturation note argue from the research side.
- **NVIDIA reportedly in talks for another USD 1B into Figure** (press, single-source report via The Information, 10-07) — **talks only, no completed deal.** Fifth consecutive week in which the capital behind manipulation is a compute or platform holder rather than an operator. If it closes, NVIDIA is simultaneously the compute vendor, the simulator vendor, a co-author on four papers in this issue, and an equity holder in the best-funded humanoid program.
- **Ultra, USD 50M Series A** (press, 10-09; USD 62M total, Framework Ventures / Y Combinator / NextView) — deepens its Physical Intelligence partnership, running π0.6 in paid warehouse work. The notable structure is that human intervention on the robot *generates training data*, so the customer's labour cost is the model's data pipeline. That is the first commercially-stated instance of the loop every human-in-the-loop paper in §2 and §6 assumes.
- **SafeWorld, USD 12.2M seed** (press, 10-08; Shine Capital, a16z speedrun, CMU endowment; CMU Safe AI Lab director Ding Zhao co-founding) — browser-based safety simulation with reactive human motion, scenarios built from incident logs, six customers. Safety *simulation* funded as a standalone category is new; it implies buyers who already have policies deployed and cannot certify them. **Nothing open-sourced, no engine disclosed, no metrics, no standards named.**
- **Boston Dynamics names Rohit Prasad CEO** (press, 10-07) — former head of Amazon Alexa and AGI, not a hardware executive, now running the Atlas company. **INFERENCE** \2192 this is the clearest single signal of the week about where a hardware incumbent thinks the remaining value is. Read it against §3: the same company shipped a 13-DoF hand with dense tactile and no success rates, so the hardware is considered solved enough to hand to a model person.
- **Teradyne invests in Bright Machines** (press, 10-05) — robotics into AI-infrastructure manufacturing; terms and stake not stated. Compute-adjacent industrial capital again. The circularity is worth naming: robots are being funded to build the data centres that train the robots.
- **FieldAI reportedly raising USD 700M** (press, reported not confirmed) — by an order of magnitude the largest robot-learning round in the window, and **unconfirmed**. Field robotics, not manipulation.
- **Minerva, USD 10M** (press, 10-07) for the Roger humanoid aimed at hazardous work, and **LivSyn, ≥RMB 100M Series A** (press, 10-10, sourced from PEdaily and not independently verified) for a cross-robot platform that turns human demo video into trajectories. Two seed-stage bets on the same thesis as Mecka from opposite ends of the market. Neither has specs, benchmarks or a released artifact.
- **Microagi claims a USD 100M annualized run rate** (press, 10-10) six months after launching Shift. **The sole source is an X post plus a promotional graphic, and the publisher states it did not verify the figure.** Annualized run rate is not revenue earned; there is no methodology, no recurring share, no customers and no split between data and deployment. Logged because Microagi's W40 acqui-hire of Animesh Garg for *post-training* is in this ledger, not because the number means anything.
- Also: a **Meta job listing explicitly referencing its humanoid robot hardware** (press, 10-07) — first hardware reference in Meta's own hiring copy, in the same week Meta FAIR published RoboJEPA. And RobCo at a >USD 1B valuation on a USD 40M round (press, 10-05) for modular cobot cells — industrial automation, adjacent to robot learning rather than in it.

6 Deep read of the week

Mulligan: Performance-Guided Data Collection for Efficient On-Robot Learning · preprint · Stanford (Ankile, Dong, Bhowmik, Muppidi, Yuan, Song, Finn) · code at `github.com/ankile/mulligan`, data at huggingface.co/mulligan, project page mulligan.page

Chosen over RoboJEPA deliberately. RoboJEPA is the week's biggest claim; Mulligan is the week's best *evidence*, it runs on one arm and one operator, and its release is the only one in this issue I verified as actually delivered rather than pledged: **196 dataset repositories on the HF Hub, all uploaded 2026-10-07 — 125 real-robot and 71 simulation — including per-round, per-condition evaluation rollouts** (e.g. `real-square-d2-r05-eval-b02-mulligan-idql-policy-rollouts`), each

with several hundred downloads already. Publishing the *evaluation* rollouts, not just the training data, is what makes a contested claim checkable. No licence is stated on the project page, which is the one gap.

Setup

Task: supervised deployment. An operator places objects, watches, and intervenes; the policy is retrained between batches rather than continuously. Embodiment: **Franka Panda on the DROID platform**, three real tasks — INSERT MARKER, THREAD NUT, ROUTE CABLE — each pairing a broad placement range with a precise insertion or seating requirement, plus two Robosuite square-nut tasks (narrow and broad initial-state ranges) for the controlled study. Data scale is small by design: a base demonstration set, then fixed per-round episode budgets, 1,009 fresh real collection episodes in total. What is held out: **2,550 blinded, interleaved real evaluation episodes on shared held-out initial states**, with five training seeds per condition in simulation.

Method

The genuinely new part is a *collection* policy, not an architecture, and it is two mechanisms. **(1) Where episodes start.** Each round promotes some initial states on which the previous evaluation policy *failed*, prioritized by how early the policy broke down, with a cap on the fraction of the next budget spent on retries. The rest of the budget goes to coverage: base demos use scrambled **Sobol sequences** rather than independent uniform draws, and coverage-fill states are chosen by **farthest-point sampling** against both previously collected states and states already picked for the current round. So it chases known failures without abandoning untested placements. **(2) What a failure teaches.** The paper names a real defect in ordinary corrections: intervene after a substantial mistake and you teach *recovery*, not the desired behaviour from the original start. So when a correction would not show the intended behaviour, the operator resets to the same placement and teleoperates the whole task — a **full counterfactual demonstration**, charged to the budget in place of a coverage-fill episode. Inherited, and acknowledged as such: the actor is an action-chunked Diffusion Policy trained by behaviour cloning; the critic is implicit Q-learning with chunked action values; deployment is best-of-N, sampling candidate chunks and executing the highest-scoring one. The one architectural judgement worth stealing is the **data split**: the actor trains only on human-controlled frames of *successful* episodes, while the critic trains on **all** transitions including failures. Actor and critic training are decoupled, and no gradient flows through the diffusion sampler.

Results

The number that matters is the pooled paired comparison, not the headline. Against uniform HG-DAgger at matched supervised budgets, actor-only Mulligan pooled across rounds gives **+11.7 / +10.9 / +6.0 pp** on INSERT MARKER / THREAD NUT / ROUTE CABLE, each significant. The full method (HiL-IDQL + Mulligan) finishes **+34 / +10 / +16 pp** above HG-DAgger. Baseline: uniform initial-state sampling with the same learner and the same episode count — the right baseline, and the comparison is blinded and interleaved on shared initial states, which almost nothing in this ledger does. Component attributions: Sobol-sampled *demonstrations* alone beat uniform samples before any failure-directed collection begins; counterfactual demos matter most early (round-one INSERT MARKER 14% → 38%) and the gap narrows as corrections accumulate; critic reranking contributes mainly in later rounds once on-policy data exists. And the collection loop stays productive: **98% of 1,009 real collection episodes completed the task** with human help, never below 90% in any round, with higher completed-task throughput per robot-minute than HG-DAgger on all three tasks — including time lost to failed evaluation episodes.

Where it’s soft

- **The final-round headline is weaker than it reads.** The 34-pp INSERT MARKER advantage is significant; the final-round THREAD NUT and ROUTE CABLE comparisons are **not**, by the authors’ own exact McNemar test, despite positive point estimates. The honest claim is the round-pooled one.
- **The reranking increment is individually non-significant on all three tasks and changes sign across ROUTE CABLE rounds.** The authors disclose this. It means best-of-N is not the part to copy first.

- **Three simulation rounds used an evaluation-derived hardness bonus with no real-world counterpart** — disclosed, but it means the sim and real curricula are not the same procedure.
- **Retraining the actor on autonomous successes *lowered* success by 3.4 pp.** That is a useful negative, and it also means the critic's gains stay outside the actor: nothing is distilled back, so you pay best-of-N inference cost forever.
- Eval breadth: three real tasks, one platform, one actor-critic formulation, one operator regime. No cross-policy-class test. And the whole method assumes **initial states can be parameterized and deliberately placed** — false in any uncontrolled environment.
- Compute asymmetry: none worth flagging. This is the rare paper where the baseline and the method cost the same.

Steal list

1. **Sobol, then farthest-point.** Replace uniform random object placement in your base demonstrations with a scrambled Sobol sequence, and pick each new collection state as far as possible from everything already collected. This is a sampler change, costs nothing, and the ablation says it pays *before* any of the clever parts.
2. **Separate your actor and critic datasets on purpose.** Human-controlled frames from successes for the actor; every transition including failures for the critic. The 3.4-pp drop from feeding autonomous successes to the actor is the experiment that justifies it.
3. **Counterfactual demos instead of corrections, early.** When a failure happens far enough in that a correction would only teach recovery, reset to the *same* placement and demonstrate the whole task. Highest value in round one, decaying after.
4. **Blinded, interleaved evaluation on shared held-out initial states, with exact McNemar.** This is the cheapest credibility upgrade available to a one-arm build, it costs only bookkeeping, and it is what lets you tell a 10-pp real gain from a 10-pp seed artifact. It is also, directly, the thing §8 says is still differentiating.

Cost to reproduce a scaled-down version

Hardware: one Franka or equivalent 7-DoF arm, one wrist-free camera rig, one operator with a teleoperation path — DROID-compatible helps, but nothing here requires it. Episodes: the published campaign is ~1,009 collection episodes plus 2,550 evaluation episodes across three tasks; a single-task replication of the *core* claim needs roughly one base set of 50–100 demos, then 3–4 rounds of ~60–100 supervised episodes each, with ~150–250 blinded held-out evaluations per round to resolve a 10-pp effect at the paper's significance levels. Call it **500–800 total episodes on one task**, which at a realistic 40–60 episodes per operator-hour is 10–20 hours of robot time, spread over rounds. Compute: a Diffusion Policy actor plus an IQL critic retrained 3–4 times is **single-GPU work — on the order of tens of GPU-hours per round on one 24 GB card**, not hundreds. The expensive resource is operator time and evaluation discipline, not FLOPs, which is precisely why this is the right paper to copy at small scale.

Read it alongside A Balanced Data Diet (preprint, University of Washington + NVIDIA, no code stated), which makes the same bet in simulation at a million parallel environments: success-guided reset sampling takes Franka nut-and-bolt from 6% (uniform) and 5% (PLR) to **70%**, locomotion 73 vs PLR's 54, a UR5e nut-bolt task 90.04 vs 48.83, with near-goal resets load-bearing, three seeds, no demonstrations and one shared reward — and it still works at 4,096 environments, so the finding is not a million-env artifact. Real UR5e after DAgger distillation to RGB: 37.5 / 61 / 94% across conditions, first-attempt 12.5 / 28.57 / 76. Two papers, two budgets, same conclusion: **choosing where the episode starts is worth more than choosing the learner.**

---

7 Relevant to what you're building

`CONTEXT.md`'s `## Active build` section is still an unfilled placeholder — **fifth week running**, and the harvester has now recorded it on five consecutive sweeps as the reason nothing can score 5 on the "load-bearing for a named project" clause. There are no workstreams to map this week's items onto, so the section is skipped rather than invented. It is also the single highest-leverage thing to fix: four of this week's items (Mulligan's Sobol-plus-farthest-point sampler, the §2 force-value calibration,

WARP-VLA's camera-perturbation adapter, EyeRobot 2.0's fixation-relative action frame) would each have been scored and written against a named subsystem if there were one.

8 Market signal

Boards checked 2026-10-11

no sweep checked them this week; last harvester check was 10-04

- **OpenAI: 19 robotics roles**, against 22 on 10-04 and 23 on 09-27. A 3-role decline is at the edge of the rendering noise this page has shown before (21/26/27 from three sources on one day in W38), so treat the direction as unconfirmed. What is not noise: **fourth consecutive week with zero titles containing policy, VLA, imitation, robot learning or manipulation.** Composition is actuators, motors, dynamometer and actuator test infrastructure, PCB layout, prototyping, thermal-adjacent, commodity management, EHS and TPM, with exactly two software exceptions — *Inference Engineer, Robotics* and *Software Engineer, ML Systems & Training Architecture*. Four weeks of this is a program, not a snapshot: INFERENCE \2192 OpenAI's visible robotics hiring surface is a hardware and actuator program with an inference team attached, and whatever policy work exists is not being hired for publicly.
- **Figure: header reports 96 open roles**, against 97 on 10-04. The board now paginates across three pages and page 1 surfaced only 8 of the AI–Helix roles (Android, Backend, Data Infrastructure, Localization and Mapping, Perception, Training Performance, Video Pretraining, XR), with the Modeling, Pretraining, Reinforcement Learning, Robot Learning and Generative AI roles not on pages 1–2. **I could not enumerate the full Helix list cleanly, so I make no claim about whether the Helix count moved from 13.** The *Reinforcement Learning Engineer, Whole Body Control* role remains under Controls, not Helix.
- **Physical Intelligence: 12 roles**, same total as 10-04 but a visibly different composition — now *ML Infra Engineer, Data Systems*, *ML Infra Engineer, Platform*, *Controls Engineer*, *Forward Deployed Robotics Engineer*, *Build & Release Engineer*, *Robot Build Technician*, *Robot Operator*, *Manufacturing Engineer*, *NPI TPM*, *Robotics Software Engineer*, *Business Operations*, *Mechatronics Intern*. **Still zero policy-research titles.** INFERENCE \2192 a "Forward Deployed Robotics Engineer" plus a build-and-release engineer plus two ML *infra* roles is a deployment org, which is consistent with π0.6 running in a customer's warehouse (§3) and with seven months of research-blog silence (§4).

Comp datapoints, primary postings:

- **OpenAI, ML Engineer, Distributed Data Systems – Robotics: USD 380K–500K + equity**, SF, hybrid 3 days, relocation offered. **Unchanged for four weeks.**
- **Figure, Helix AI Engineer, Modeling: USD 200,000–400,000 base**, San Jose, **5 days/week in office**. Unchanged.
- **Physical Intelligence** publishes exactly one rate, again: **Robot Operator at USD 25/hour + benefits**. Every engineering posting has compensation display switched off.

INFERENCE \2192 unchanged from W40, and the stability is itself the datapoint. Figure is still the only lab hiring modelling and policy people in volume *and* publishing a band for them, and that band still tops out ~USD 100K below OpenAI's band for a data-pipeline engineer. The highest published number in applied robotics remains attached to data infrastructure, not to policy research.

Skills recurring in the work that got attention this week:

1. **Controlled comparison with one thing varied.** OpenWAM holds the backbone, tokenization and objective family fixed and varies only generation order; Long-WAM varies only pretraining type at matched context; the privileged-distillation paper varies only demonstration source at identical student, budget and trainer; LeWAM size-matches every baseline to ~41–43M under one protocol. Four papers, one discipline. This is the skill that distinguishes someone who can be trusted with an ablation budget.
2. **Scaling and threshold reasoning in FLOPs.** RoboJEPA's capability ladder and TRI's 300h → 30,000h ladder both answer "how much do I need to buy for this capability" rather than "did I win." Being able to say "object dynamics through contact

shows up at ~10^22 FLOPs and not below" is a different kind of competence from quoting a success rate.

3. **Statistical hygiene on real hardware.** Mulligan's blinded interleaved evaluation with exact McNemar and paired bootstrap, LeWAM's 5×50 seed grid, the privileged-distillation paper's three seeds with disjoint Wilson intervals and a protocol frozen before results. Still rare enough to be a differentiator.
4. **Deployment-cost accounting at consumer scale.** Long-WAM's 107.4 ms on an RTX 5090 *including* the VAE, RealtimeWAM's 12.2 ms H100 figure degrading to 27.2 ms on a 4090D, RACE's idle-time reduction, PhoneBot's per-function CPU table. Reporting the number on the hardware a reader actually owns.

Table stakes, updated. Released code at submission — cleared this week by RoboJEPA (stated), OpenWAM (verified), BiGym 2.0, Mulligan (verified, data included), PerturBot, RealtimeWAM, EigenDEXplore; **not** cleared by TACROSS, SoTa, TouchScale, LeWAM, QF3, EgoLAP, PEARS, RACE (no URL), Vela, TUCO, InterMimicGen. End-to-end latency rather than head latency. An ID/OOD split. Multi-seed paired evaluation. Knowing which benchmarks are compromised.

Still differentiating. A matched-everything protocol across the configurations you compare; a *blinded* real-robot evaluation with a paired test; publishing the evaluation rollouts and not only the training data (Mulligan is the only instance in this ledger); reporting a capability threshold in compute rather than a win; and seed bands on real hardware, which remain so rare that W39's Res-HIL is still the cleanest example this quarter and **EigenDEXplore, RACE, Long-WAM and WARP-VLA all explicitly lack them on hardware**.

9 Threads

`THREADS.md` updated. Board as it now stands; 12 threads, at cap.

THREAD	STATUS	STATE, AND WHAT WOULD COUNT AS THE NEXT REAL UPDATE
World-action models as unified policy + simulator	**advancing — ceiling now measured**	RoboJEPA fits a second-order compute law (extrapolation error $0.6/1.4\times10^{-3}$ vs 2.0/3.7 for a plain power law) with ordered capability thresholds — reach ~10^{20}, hold-and-move ~3×10^{20}, static obstacles ~10^{21}, object dynamics through contact ~10^{22} FLOPs — and states the fits already sit near **data** saturation on a fixed 2.87M-trajectory corpus. OpenWAM: generation order worth <2 pp, robot-video backbone pretraining worth +29.4/+34.4 pp on LIBERO-Long. Long-WAM: 19.2 s of context worth +15.4 pp to AR-pretrained backbones and ~0 to a bidirectional one, LIBERO-Long saturating at 2.4 s and regressing at 38.4 s. TRI's 30,000-hour ladder: agent fidelity saturates early, object-interaction fidelity converges below usable, skeleton conditioning at 1k h beats standard at 30k h. Carried: Simple-WAM still has no code and still rests on a 5B backbone without embodied pretraining. Contested: Long-WAM's real tables have no seeds; OpenWAM's counterfactuals are simulator-branched; TRI's fidelity metrics are its own and its OOD benchmark is pledged not delivered. **Next:** the W40 question unchanged — the same `t_denoise` sweep on a WAM that *has* embodied pretraining; plus whether the capability ladder survives a different encoder (§10).

THREAD	STATUS	STATE, AND WHAT WOULD COUNT AS THE NEXT REAL UPDATE
Planning on a learned model	**NEW — opened 2026-W41**	LeWAM: sampling-based MPC over raw actions collapses every model tested, its own included, to **3–8%** closed-loop success; moving the decision variable into the flow head's noise space recovers and exceeds the policy, with projection onto the √A shell load-bearing (Can 78.8→23.0 without it). RoboJEPA plans by CEM on one goal image with uniform action sampling and no learned proposal, reaching 67/50/27% on real Franka grasp/lift/pick-place; π0.5 still wins pick-and-place. Future Anchored Verification, R²-WAM and DreamTrue are three independent attacks on action-faithfulness in one window, all abstract-unverified. Contested: LeWAM is simulation only, needs a retrieved goal latent, has no code, and its Dreamer comparison is asymmetric. **Next:** a diffusion-steering result on real hardware; or RoboJEPA's released checkpoints planned with a learned action proposal instead of uniform sampling — an inference-only change on a released artifact.
Where the episode starts	**NEW — opened 2026-W41**	Mulligan (Stanford, **code + 196 HF dataset repos verified delivered, including evaluation rollouts**) beats uniform initial-state sampling at matched budgets by +11.7/+10.9/+6.0 pp round-pooled on three real Franka tasks over **2,550 blinded interleaved held-out episodes**; Sobol base sampling plus farthest-point coverage fill pays before any failure-directed collection; counterfactual demos worth 14→38% in round one; 98% of 1,009 collection episodes complete with human help. A Balanced Data Diet (UW + NVIDIA) reaches the same conclusion in sim at 1M parallel envs: Franka nut-bolt 6% uniform → 70% success-guided, locomotion 73 vs PLR 54, still working at 4,096 envs. Contested: Mulligan's final-round gains are non-significant on two of three tasks and its reranking increment changes sign; the Diet paper states no code and its real UR5e numbers follow DAgger distillation. **Next:** a third-party replication of the Sobol-plus-farthest-point sampler alone, which costs nothing; or either method applied to a policy class other than Diffusion Policy / PPO.
Force as a first-class action	**ADVANCING**	PEARS puts a VLM *between* episodes emitting contact-force bounds a hybrid force-position controller enforces on a frozen GR00T-N1.7: whiteboard erasing 11/20→19/20, pipette aspiration 8/20→18/20 after 40 online episodes, zero residual force failures. The privileged-distillation paper supplies the week's calibration on **when force earns its place**: 12/16/18 pp on a 0.114 mm-clearance peg, 5–8 on a nut, 0–1 on a gear. Tactile reconstruction of task frames and forces (Sapienza + Twente) claims force *and* task frame from tactile alone with no F/T sensor. Carried from W40: Wrench-ACT, FP2, HACo and Impedance Cloning — **none has released code or data, five weeks on**. **Next (unchanged):** any of those four releasing; and the falsifying experiment — a wrench-action policy trained on *position*-teleop data matching one trained on force-feedback teleop data.

THREAD	STATUS	STATE, AND WHAT WOULD COUNT AS THE NEXT REAL UPDATE
Contact sensing without dedicated tactile hardware	**ADVANCING**	Factorized Tactile (Amazon FTR + UT Austin) reports two findings that cut against adding hardware: **centre-of-pressure features performed worse than proprioception alone**, and sim rankings did not match hardware. SimpleTouch claims a single-stage tactile expert on frozen π0.5 beats two-stage pipelines, i.e. tactile pretraining unnecessary. EgoPhys estimates peak contact force and mechanical work from egocentric video with no force sensor; MIM-VLA reads post-contact response off **gripper motor feedback**; both abstract-unverified. Against that, OpenViTac shows why touch is irreplaceable for one specific reason: under lighting perturbation a co-trained visuo-tactile policy holds 6/20 and 12/20 where real-only π0.5 falls to 0/20 and 2/20, with one task rate **completely unchanged** by the lighting change. **Next (unchanged, now seven weeks):** an estimated- or predicted-force policy matching a real-sensor baseline on a task whose labels were *not* collected with that sensor.
Paying at training time, not run time	**ADVANCING**	TACROSS: 30 robot + 150 human demos → 91.9% vs 51.3% robot-only (+40.6 pp); 50 robot + 50 Ego matches 92.5% against 95.0% for 92 robot demos, a 45.7% cut in robot demonstrations. TouchScale: 500 h of paired human vision-and-touch (≈13× the largest comparable corpus) as **action-free mid-training** takes a VTLA from 22.5 → 57.5% real success on four contact-rich tasks with no retargeting. SoTa: identical 202-taxel layout on human and robot, so no cross-sensor mapping is learned at all. EgoLAP (Princeton + TRI + PI): language actions give 16.4% sim / 44.5% real vs 1.6/35.2 for EgoFAST under a matched ablation. Contested: **TACROSS, SoTa and TouchScale have all released nothing**; TouchScale's data was *purchased from a vendor* with no licence stated, and its future-tactile and action prediction errors correlate only weakly (r=0.119), so the mechanism behind its 35-pp gain is unestablished. Carried: ME-Dex's zero-tactile result is still shown only on simulator-generated tactile. **Next:** any cheap-tactile release landing with a licence; and still unrun after five weeks — a matched comparison of one-step action heads (IMLE vs drifting vs consistency vs MeanFlow) on one backbone.
Benchmark validity in physical AI	**ADVANCING**	BiGym 2.0 contributes the most reusable finding: with a stateful whole-body controller a **seeded reset does not reproduce a trajectory** (9.5×10^{-2} qpos divergence immediately, 2.1×10^{-1} within 50 steps) — restoring the controller snapshot from first reset makes evaluation **bit-identical**. CARE proposes *certifying* VLA acceleration claims rather than auditing after the fact, a direct follow-on to W40's deep read. ManiPhysicsBench scores whether a VLA damages objects, which rigid-body success rates never measure. Contested: W40's seven-benchmark audit **still has no resolvable repository URL for its fixes**, two weeks on, and GAM's W39 claim still rests on the LIBERO-PRO column that audit showed was a loader bug. **Next:** the audit's fixes as merged upstream PRs with a URL; and the bar unmet since W38 — a major-lab table with seed bands *and* paired evaluation together. Mulligan is the closest anything has come.
Online correction as the last mile	**advancing (sharply)**	Mulligan is the first result in this thread to clear the evidentiary bar Res-HIL set and then exceed it: blinded interleaved evaluation, exact McNemar, paired bootstrap, 2,550 held-out episodes, **and code plus data actually delivered**. Also RESETTLE (retrieval-based recovery triggered by policy *disagreement* rather than VLM replanning) and UniIntervene++ (intervention strategy adapted to current competence), both abstract-unverified. Carried and still unmet: Res-HIL and ReF-HIL have no code; **EXPO-FT's released code is now five weeks out with no third-party rerun by anyone. Next:** that third-party rerun; or Mulligan's sampler reproduced by someone who did not write it.

THREAD	STATUS	STATE, AND WHAT WOULD COUNT AS THE NEXT REAL UPDATE
Low-cost open hardware as training substrate	**advancing in design, flat in artifacts**	Four new price points, zero new files. TACROSS USD 10.86 for 285 characterized sensing locations (USD 0.038 each); SoTa under USD 10 of materials for 202 taxels with 10,000-cycle durability data; PhoneBot ~USD 400 for a 13-DoF lower body using a commodity phone as IMU, camera and policy compute; Q8botOne, a palm-sized ESP32-C3 quadruped, crowdfunding not shipping. **I verified TACROSS's and SoTa's project pages today: both state release is *planned*, neither gives a repository or download link, neither names a licence.** PhoneBot releases design and code but scores nothing — walking, following and standup are demonstrated with no success rate, distance or disturbance number. Carried: Yeah Hand remains fully released with no evaluation and no independent build; the Antagonistic Tendon-Driven Hand still publishes no design files; still no third-party build of Aero Hand, BRIDGE or Peg-in-Bench. **Next (unchanged):** a clearly independent group reporting success rates on released low-cost hardware with code. TACROSS at USD 10.86 would be the most valuable target the moment its files exist.
Claimed open releases vs actual artifacts	**narrowing broke on hardware, held on models**	Delivered and verified this week: Mulligan (code + 196 HF dataset repos including evaluation rollouts), OpenWAM (repo + pretraining weights on HF, **no licence**), BiGym 2.0, EigenDEXplore (+ pretrained PCA for six hands), PerturBot, RealtimeWAM. Stated but unverifiable from this session: RoboJEPA's code and checkpoints (GitHub 403 all week, no HF mirror under that name). **Pledged and not delivered: TACROSS, SoTa, TouchScale** — the three cheapest and most directly useful artifacts of the week, all hardware or data. RACE states code on GitHub with **no URL in the paper**; Long-WAM states a code link whose URL does not resolve; QF3 and EgoLAP are website-only. Carried: Runway's Praxis-1 open weights still pledged with no licence; **StreamPI official weights now eight weeks pending**; the benchmark audit's fixes still have no URL. **Next:** whether a cheap-tactile release lands with a licence; Praxis-1; StreamPI.
Who owns the open robot-learning stack	**ADVANCING**	**LeRobot is 0.6.1, released 2026-08-03 — eleven weeks post-deal with no release, now confirmed from PyPI rather than inferred from a stale GitHub index.** That upgrades five weeks of "unverified" into a trustworthy negative. Meanwhile NVIDIA appears this week as co-investor in Mecka's data-collection Series B, as a reported USD 1B suitor for Figure, and as co-author on four papers in this issue. Asset interoperability is contested from both the vendor side (USDCraft) and the community side (Draft REP 158, 21 replies, no ratification). Carried: **no governance structure has been named in any of the five platform-holder commitments** — still press quotes, not foundations. **The W35 in-quarter prediction (a logistics incumbent buying an end-effector or policy team) has ~2 weeks left and has not happened once;** five platform/capital moves have. Kill the prediction on lapse, keep the consolidation framing. **Next:** any commitment written into a foundation or steering body; the first post-deal LeRobot release and whether GR00T or Isaac integrations get privileged treatment in it.

THREAD	STATUS	STATE, AND WHAT WOULD COUNT AS THE NEXT REAL UPDATE
Simulation fidelity for transmission and contact	**un-stalled**	MuJoCo 3.15.0 shipped 2026-10-05 — confirmed on the release tag by the harvester and **independently on PyPI by me today** — with capsule multicontact, Stable Neo-Hookean flex, >2 GiB model serialization, and a **breaking change: `dampratio` now uses operational-space inertia**, which silently alters every tuned value. RiCo proposes neural rigid-body simulation via *local contact reasoning* rather than global object dynamics — the first new item aimed squarely at the contact model in three weeks (abstract-unverified). Isaac Lab, Genesis, ManiSkill and RoboCasa remain **unverified**; every GitHub surface was 403. Contested: still **no sim-to-real transfer result crediting the discrete integrator, the IPC flex contact or routed-cable modeling**, so the thread's central claim stays unexercised; MuJoCable still has no repo. **Next:** a tendon-hand or underactuated-finger sim-to-real result attributing transfer to the new flex contact. W40's W43 kill date stands; 3.15.0 buys the thread two more weeks, not a reprieve.

Changes this week. Two threads opened — *Planning on a learned model* and *Where the episode starts* — because this week produced load-bearing, falsifiable results in both that did not fit anywhere existing. To stay at the 12 cap, *Does VLA scale buy language-conditioned control?* and *Humanoid capital vs demonstrated capability* were **closed**, and neither quietly. The first is closed because its own bar (a sub-10M real-hardware multi-task policy) went seven weeks unmoved and the W39 claim it rested on was shown to be a benchmark bug, so the question as posed is no longer answerable from the evidence stream; its live descendant is the viewpoint and shortcut-robustness work (WARP-VLA, PerturBot), which belongs under *Benchmark validity*. The second is closed because it produced the same finding for seven consecutive weeks — vendors report built and shipped, nobody reports working — and because its one useful metric now has a better home: Ultra's 96.4% on a π0.6 model is a *policy* datapoint, not a humanoid-capital one, and the humanoid platform items are reported in §3 without needing a thread to justify them. Both closures are recorded in `THREADS.md` with their final state. One candidate is queued behind the cap: **human video as the data substrate and where it saturates** (TRI's 30k-hour negative, TouchScale, TACROSS, EgoLAP, Workhorse, ExoBridge, OmniHOI, IronMind) — currently split across *Paying at training time* and the WAM thread, and it should be promoted the moment a second group reproduces TRI's saturation finding on a different corpus.

10 Open question

Is RoboJEPA's capability ladder a property of robotic world models, or of the frozen encoder underneath it?

The ladder is the most actionable thing published this week: reaching at ~10^20 FLOPs, moving while holding at ~3×10^20, static-obstacle avoidance at ~10^21, and object dynamics through contact only at ~10^22. If those thresholds are properties of the *problem*, they are a budget calculator for everyone — you can read off what capability a given compute spend buys before you spend it. If they are properties of **V-JEPA 2.1's representation**, they are a budget calculator for exactly one encoder, and the field has just been handed a very precise measurement of something contingent.

The paper names the limit itself: the encoder is frozen, so the law describes a dynamics predictor on a fixed representation, not joint encoder-predictor scaling. What makes the limit load-bearing rather than pro forma is that this week and last supply direct evidence the representation moves the very quantities the ladder is measured in. W40's ReWAM found **raw DINO features beating Video-VAE latents by 4.9–9.2 points despite worse pixel reconstruction (SSIM 0.70 vs 0.91)** — so an encoder's reconstruction quality and its usefulness for action are not the same axis. LeWAM trains a **41M** JEPA from scratch and gets task-state probe R^2 0.80 while scoring **0.00 on action-independent distractors** where a reconstruction-based model scores 0.73 — so what an encoder chooses *not* to represent is as consequential as what it does, and object-pushing is precisely the capability where distractor-invariance should matter most. Juno gets real-robot shift robustness out of a **15M** from-scratch JEPA on 60 demos. And Long-WAM's result — history helps AR-pretrained backbones and does nothing for a bidirectional one — is the same lesson one level up: the initialization decided whether the mechanism worked at all.

INFERENCE \2192 the ordering of the ladder is suspiciously aligned with how much of the signal survives a representation trained for prediction rather than for control. End-effector pose is the most linearly decodable thing in a robot video; object dynamics through contact is the least. A different encoder could plausibly reorder the top two rungs or move the 10^22 threshold by an order of magnitude in either direction, and nothing in the paper rules that out.

What would settle it:

Rerun the 22M→2B ladder twice with everything else matched — identical data mixture, identical teacher-forcing plus rollout losses, identical rollout lengths and resolution schedule, identical nine-step held-out DROID and RoboCasa evaluation — on two frozen encoders: V-JEPA 2.1 as published, and a frozen DINOv3-L, the representation ReWAM showed beats video latents for action. Report the four capability thresholds for each. Three outcomes, all informative: if the thresholds land within a factor of two, the ladder is about world models and it becomes the field's planning instrument; if they shift by an order of magnitude but keep their order, the *order* is the finding and the FLOPs numbers are encoder-specific; if the order changes, then "capability emergence" in this setting is a statement about linear decodability in a frozen feature space and should be described that way.

The cost is the honest part of this question. It is not an inference-only sweep like last week's — it is a retrained ladder, and at 22M→2B that is real compute, though an order of magnitude below the 8B runs and well inside a single academic cluster's budget. The code, the data list and the fitting procedure are all released, so nothing has to be reinvented. And a cheaper partial version exists: fit the law on the 22M–300M rungs only, on both encoders, and compare the fitted exponents and error floors. That is a weekend of GPU time on a small cluster and it would already tell you whether the two curves are even the same shape.

ALSO THIS WEEK (4+ IN QUEUE, NOT COVERED ABOVE)

Nothing scoring 4 vanishes silently. All preprints; abstract-level unless the queue recorded a full read.

- NAVA-WAM — Meta AI + Physical Intelligence; pretrains the action policy itself on action-free video, skipping the latent-action and representation stages entirely. No code.
- PointWAM — POSTECH + KAIST + RLWRLD; forecasts scene *and hand* 3D point trajectories; human-video pretraining lifts DexJoCo 12.1 → 69.0. No code.
- InterMimicGen — UIUC + NVIDIA; tracker and reference data grow together, verified references expanding 150× over five rounds to 16,059 motions / 140.68 h / 157 objects across six robot configurations. But the single generalist tracker reaches 73.2/64.9/58.3% against per-object specialists at 93.8/75.4/84.9 — a large generalist penalty, worst on grasping — and training is 4 GPUs × 4,096 envs. No release stated.
- Vela — Fudan + CUHK; predicts B-spline control points plus a motion-dependent horizon instead of fixed-rate chunks. Pretraining is 64 H200s for 300K steps, so the reachable variant is π0.5-Vela at 42.0 LIBERO-X vs 39.3 for π0.5 against full Vela's 45.3 — most of the gain is in pretraining you cannot run. **Useful negative: post-hoc spline fitting on π0.5 outputs does not recover it (38.9–40.0).** No code.
- TUCO — Stanford + NTU + CMU + Berkeley; influence functions decomposed into target-return utility and rollout coverage to pick which *sim* demos help co-training. The +65.0/+35.0/+80.0 pp headline is against full-pool co-training, which suffers negative transfer; sim-to-sim margins over full-pool are 8.4/10.2/5.8 pp. Only **ten** recorded real rollouts per task, serving as both target demos and scoring rollouts. No code.
- QF3 — Berkeley + Amazon FAR; flow matching plus Q-maximization through a one-step denoised-action prediction, with the critic's velocity correction clipped per-coordinate around the replay flow target. No integration or differentiation through the ODE. The clip stays active throughout training (81–85% of dimensions inside the band at 500k steps), a 29-DoF Unitree G1 trains from scratch and deploys zero-shot, three seeds throughout, and velocity tracking trains on **a single A10G**. Website only, no code URL; all manipulation results are simulation.
- EgoLAP — Princeton + TRI + Physical Intelligence; humans and robots share *motion intent* expressed as structured language actions plus motion-level reasoning, both landing in the VLM's pretrained semantic space. 80.1% mean real task progress, 2.3× over alternative action representations; the matched ablation gives 16.4% sim / 44.5% real for language actions vs 1.6/35.2 for EgoFAST. ~10k hours on TPUs; reasoning labels are **Gemini-generated**, not human; website only.

- RoboRender — Stanford + NVIDIA + MSR (Fei-Fei Li, Jiajun Wu); generative video rendering of simulated trajectories beats domain randomization **3.6×** for zero-shot transfer, 71% real and 7.1× over raw sim. Project page only; high reproduction cost.
- WARP-VLA — Yonsei + KETI; a mixture-of-experts adapter maps perturbed wrist-camera features back to a pseudo-canonical anchor with **no calibration and no policy fine-tune**: π0.5 39.2 → 78.3 on perturbed LIBERO, GR00T-N1.7 41.3 → 66.5, LIBERO-Plus 28.8 → 56.5 with zero adapter retraining, and real xArm6 36.7 → 48.7 with canonical performance preserved. Directly on-filter for a rig whose wrist camera gets re-mounted between sessions. Release stated; **single seed**, one episode per config on LIBERO, and training is 8× RTX PRO 6000 for 30 h — not one GPU.
- MuJoCo 3.15.0 — covered in §4; listed here because the queue scored it 4 and the `dampratio` breaking change deserves to appear twice.

Rate this issue 1-5 and tell me what was noise.