This Week in Robotics

THE LINE

Touch stopped being an input this week. Two results, pointing from opposite ends at the same conclusion: the value of contact data is in the training signal, not in the machinery you build to consume it.

Spark-0 (preprint, Xspark AI + THU + MMLab@HKU + HKUST-GZ + PKU) uses 1,000 hours of human full-hand tactile data purely as *train-only supervision* for the visual tokens of a Qwen3-VL-4B backbone — glove readings become image-space contact maps, a projector head predicts them, and the head is deleted at inference. On the Ego Humanoid benchmark, where the robot has no tactile sensor at all and none is used at test time, mean success goes from 52.45% to 71.72% against a vision-only pretraining variant on identical RGB data and identical compute.

ForceBench (preprint, PKU-PI Lab + Berkeley + NUS + ePyBot) takes the other half. It puts one no-force baseline and six force-integration designs on a single shared π0.5 backbone with the same 1,600 demonstrations and the same training protocol, and runs 16,000 seeded simulation episodes plus 800 real trials on a Franka FR3 with a wrist F/T sensor. The design that wins is the one with **zero new parameters**: normalise the six-axis wrench, discretise it, and append it to the state string in the prompt. 35.31% against 23.25% for no force in sim; 63.00% against 40.00% on hardware. Every architecture built specifically to fuse force — token fusion, cross-attention, future-force prediction, temporal windows — either ties it or loses to it, and both temporal designs score *below the no-force baseline* on the assembly task.

Put together: getting contact into the model is worth 10–20 points. How you get it in is worth roughly nothing, except that elaborate interfaces can cost you on precision assembly. inference: the last two years of force-fusion architecture papers were measuring the presence of the signal and attributing it to the mechanism. This is the deep read (§6), and it is the most interview- defensible result of the week.

Second story, smaller but concrete: **the artifacts that actually shipped this week are the mid-sized ones.** RoboParty's RP1 put a full humanoid hardware stack on GitHub under CERN-OHL-W-2.0 with STEP models, Gerbers, placement data, harness drawings and per-configuration BOMs — files, not a pledge. LTM-pi shipped training and inference code under Apache-2.0. RAX shipped code, per-embodiment skill libraries and twelve complete real-robot session logs. Meanwhile the week's headline dataset, HumanTouch at 1,000 hours, is published at **100 hours** (§2).

Window: 2026-10-11 → 2026-10-18. Queue: 5 part files, ~300 scored lines, 19 at 4.

Coverage caveats carried from the harvester, unchanged in kind from W41 and worth reading before you treat any absence below as a negative. arXiv's `/new` listing truncates its served body to ~45 entries while its header reports the untruncated count, so on the two days a cs.RO listing was reachable, recall was 45/45 of the served body but 45/255 and 45/89 of what arXiv claims to announce. cs.LG and cs.CV were never enumerated on any of the seven sweeps. The 10-17 sweep found that 32 of 45 announced cs.RO entries in the 2610.207xx–2610.215xx band were missing from the ledger — a band two earlier sweeps had already claimed as covered on the strength of a high alphaXiv dedup rate. **Read a high dedup rate as a sample, not a census.** The weekend produced no arXiv announcement at all, so the flat band on 10-17 and 10-18 is a calendar artefact.

1 Policy learning & VLAs

LTM — memory as one token per transition

LTM: Latent Transition Based In-Context Memory for VLAs · preprint · Beihang + BAAI · code: github.com/buaa-colalab/LTM-pi, Apache-2.0, verified present (training, inference, cache generation, a RoboMME evaluator; weights and datasets linked externally, not bundled).

Architecturally new:
History is not stored as frames, tokens, or a summarised language trace. A *frozen* latent action encoder (CD-LAM or DreamDojo) plus a lightweight adapter maps each consecutive observation pair to a single memory token representing the transition between them. One token per transition, causal KV-caching so encoding overlaps execution, a delayed-transition prefix with lag *k*, and sparse initial-frame anchors for video-conditioned tasks. No memory-specific annotations, no external reasoning model, no intermediate language. Instantiated on π0.5 as LTM-π.

What the eval covers:
Third-party benchmarks, which is the part that matters — RoboMME (16 tasks) and LIBERO-Mem, not a benchmark the authors built. RoboMME 58.54% against 17.93% for memory-free π0.5; LIBERO-Mem 80.5 against 42.5 for the best baseline; real-robot counting 11.25 → 82.50%. A 200-token memory budget still reaches 84.00% against 87.50% uncompressed, and a random-fixed-representation ablation isolates transition *content* from bare temporal position.

Soft:
SimpleMemVLA at 1× data scores 88.25%, which beats LTM-π at 10× data (87.50%). The authors report this. It means the headline is "a cheap memory mechanism that scales with data," not state of the art. Only the Full and *k*=5 rows use multiple seeds; every other ablation is single-seed. Multi-object tracking through sequential interactions remains at 25.0%.

Reproducible with one arm and one GPU?
Inference yes — 66.71 ms on an H100, and the authors state it runs on a single RTX 4090. Training no: 8×H100, batch 128, 40k–120k steps.

LoopWAM — learning from your own bad rollouts

LoopWAM · preprint · abstract read by the harvester, not verified independently.

Architecturally new:
Source-dependent noise sampling. One dual-stream DiT (Wan2.2-TI2V-5B base) trains on a mixture of expert demonstrations and its own deployment rollouts of mixed quality, with the quality level *steered through the instruction* rather than filtered out of the data. That is the interesting bit — failures are kept and labelled rather than discarded.

What the eval covers:
RoboChallenge Table30-V2, **a third-party benchmark**, 30 tasks and 4 embodiments. 23.67% → 43.33% over three improvement rounds; shirt folding 13.2 → 55.56%.

Soft:
No seeds, no code statement, compute unstated. Pairs directly with Back on Track (Stanford, world-model planning for recovery from deployment failures the policy cannot predict) and Value as Memory — three groups attacking the deployment-feedback loop in one week.

SmolWAM — the one that actually fits your budget

SmolWAM · preprint · ML Collective · code and weights released.

Distils Cosmos Policy from 28 to 12 blocks, single-frame latents, one rectified-flow step, trained in **3.6 A100-hours**. This is the only policy-learning item this week that a reader with one GPU can retrain end to end today. The caveats are honest and limiting: LIBERO only, no real robot, and the "edge device" claim in the title is unmeasured — no latency number on edge hardware appears. Treat it as a distillation recipe with released weights, not an edge result.

Physical time as a conditioning variable — now a two-paper cluster

Chronos Policy (Stanford REALab) conditions a diffusion transformer on capture and execution timestamps via time-aware RoPE plus sampling randomisation; 3×3 seeds, 450 episodes per cell, one arm, no code, platform unnamed, only 10 real trials per task. Physical Time as a Missing Dimension of Robot Policy Memory Dynamics is the follow-up and goes further: persistent memory relaxes in **closed form over measured elapsed wall-clock time** with per-channel learned time constants, implemented as AdaLN/AdaRMS modulation on both Diffusion Policy and π0.5, requiring no event-time labels. Real Galaxea A1, 5 tasks, 20 trials; LIBERO 40 tasks, 3 seeds; 4×H200 for 80 h; no code. Its weakness is that the only real-robot baseline is a reactive policy, so the comparison that matters — against a frame-count-conditioned memory — is missing.

INFERENCE \2192 this is the same move as LTM from a different direction. Both say the index into history should be something other than "number of frames ago," and both are cheap to bolt onto an existing head. If you only read one, read LTM, because it released code.

Backfill the W41 window missed

The 10-08/10-09 and 10-12 announcement cycles were never enumerated when W41 shipped. Four items from those cycles score 4 and are flagged here rather than left in the ledger:

- Dream to Remember (Freiburg + TUM) — continual learning with a *frozen* world model rehearsing prior tasks in imagination while a steering module (noise space, not weights) learns the new one. CALVIN eight-task 63.7 → 74.8 where LoRA collapses to 11.7; three real Franka tasks learned **entirely in imagination**; works even when the world model is too weak to support policy optimisation. No replay buffer, no new data. Code release not stated. Project page. This is the strongest unread paper in the ledger.
- Splat2Act (Georgia Tech + Waymo + TRI) — phone scan to 3DGS twin, privileged oracle generates 68,021 pickup demos, personalises a frozen-encoder π0.5 from 0.0% to 62.5% sim / 63.9% real on Stretch 3; 101 building-scale scans. Dataset promised *on acceptance*. OOD lab tables drop to 22–33%.
- StereoEngine (USC) — warps monocular demos to a right view and inpaints disocclusions, never training on real stereo; dataset stated released; removes the stereo rig and re-teleoperation cost outright.
- Reflex (UW + Stanford + USC + NUS + UNC + AI2) — three-stage decomposition for reactive humanoid catching; code at `github.com/ToberJ/reflex`; real G1 65% and 40%, sim 85.7 against privileged 87.9.

---

2 Manipulation, dexterity & sensing

Spark-0 — touch as supervision, not as a sensor

Spark-0 · preprint, 15 Oct · code github.com/XsparkAI/Spark-0, MIT, 7 stars, 4 commits · weights `hf.co/TianxingChen/Spark-0` (the `pretrain_spark_0` directory exists).

The mechanism. Two roles for touch, separated cleanly. (1) *Train-only spatial alignment*: 320 taxels per hand are mapped onto a reconstructed MANO mesh, diffused geodesically over its surface, rendered into the head camera, pooled onto a 12×12 token grid and used as a soft target for a focal+Dice head on layer-25 VLM tokens. A frozen Depth-Anything-3 LARGE teacher supplies dense geometry targets for the same tokens. Both heads are deleted at inference. (2) *Tactile pathway*: a 20-frame (333 ms) regional force history over six regions per hand plus a 3D taxel encoder produce 24 tactile tokens that condition the action expert only. **Raw touch never enters the VLM**, which is what lets one checkpoint serve both vision-only and tactile deployment.

The numbers that matter

Are the four-variant pretraining table, because it is the only part with a controlled comparison (identical RGB data, identical compute, identical downstream fine-tuning):

VARIANT	SPARKARENA-SIM	EGO HUMANOID	SPARKARENA-REAL
Scratch (no pretraining)	31.71	**58.34**	38.00
Visual PT (same data, touch discarded)	36.57	52.45	44.00
100 h touch	35.14	59.42	40.00
1000 h touch	36.86	71.72	52.00

Where it's soft, and this is load-bearing. The advertised "+19.27 points from touch" is measured against Visual PT, which is *worse than no pretraining at all* on the same benchmark (52.45 against 58.34). Against Scratch the gain is +13.38. Against Scratch, 100 hours of touch buys **+1.08**. So the scaling story rests on two points, one of which is a baseline that went backwards, with no 300 h rung in between. Contact-map supervision and the tactile pathway are never ablated separately, and the geometry teacher is in all three pretrained variants, so the paper cannot attribute the gain to contact specifically. Real evaluation is 10 trials per task across 5 tasks. Seed is 0, singular. Insert Test Tubes is 0% for every method including Spark-0.

Release reality, checked today. The paper says HumanTouch is 1,000 hours. The public dataset, `chuqiaoLyu/Xspark-HumanTouch`, is a **~100-hour v1 subset over 10 canonical tasks**, dated 2026-08-15, with a roadmap entry promising the 1,000-hour expansion "within a month" — which has now lapsed. The card states 360 effective points per hand against the paper's 320, and explicitly says the glove values are *relative response, not physical force*, which conflicts with the force-calibration tick in the paper's own comparison table. SparkArena is stated as released with assets, paired initial conditions and scoring protocols; I could not find it on the Hub or in the repo README. 4,940 downloads on the 100 h subset, so people are using it.

Reproducible?

No. 12 nodes × 8 GPUs, 400k pretraining steps at global batch 576, then 80k fine-tuning steps, on a vendor embodiment (Marvin M6S arms, WUJI Hand 2, Moxian gloves). What is transferable is the idea, not the artifact — see the steal list in §6 for the cheap version.

ForceBench — see §6

The deep read. Short version: on one backbone with everything else held fixed, prompt-discretised force ties or beats every purpose-built fusion architecture, and temporal force windows actively hurt on assembly.

Dexterity from Supervised Play — the play policy is a data collector

Dexterity from Supervised Play · preprint · Amazon FAR + UMD (Malik, Aloimonos, Sharma) · code and data promised, **not released**.

Minutes of goal-free human teleop train a diffusion-transformer BC "play policy," which is then run autonomously to collect hours of contact data; an action-conditioned world model (Diffusion Forcing) scores 1,024 MPC candidates proposed by that same play policy. No simulator, no reward, no task demonstrations. Real 22-joint Sharpa hand on a Tianji Marvin M6: in-hand reorientation 68.6 / 90.0 / 97.1% at 0.4 / 0.6 / 0.8 rad tolerance against play-alone at 20.9 / 38.3 / 54.1 and an SO(3) random walk at 6.3 / 11.4 / 17.5. **701K frames against a performance-matched PPO baseline at 344M frames — 490×.** Autonomous play from 0.54 h to 5.11 h cuts 50-step rotation MAE monotonically from 34.96° to 28.40°. A distilled self-play policy runs at 21.8 ms, 7× faster than the 134 ms replan, and perception plus planning fit on **one RTX 5090**.

The catch is unusual and the authors state it: the paper is an admitted draft whose conclusion says frozen quantitative results remain to be inserted, and the hardware claim is provisional. 128 seeded sim trials, 64 visual trials, no real-robot seed counts. Treat every number above as subject to revision.

FOMO-Dex — the compliance result with a negative inside it

Force-Modulated Dexterity · preprint · Amazon FAR + CMU · no code, project page only.

Force corrections are applied **only along task-geometry directions** rather than isotropically. Screwdriving goes from 4.2 to 34.7 turns per episode. The buried finding is better than the headline: imitating the *raw operator commands* and applying force modulation at inference gives a **0% lift**, while imitating the *force-corrected actions* gives 67%. The correction has to be in the data, not in the controller. That is a direct constraint on how you build a teleop pipeline.

Against it: bolt-and-nut regresses versus stiff control (20.3 against 23.7 turns), naive isotropic compliance beats the method on force exceedance in three of four tasks — so the win is task progress, not force regulation — the user study is n=4, and there are no seeds on the BC tables. The rig (7-DoF arm + 22-DoF hand + wrist F/T + fingertip tactile + joint torque + glove + Vive) is not a one-arm budget.

FOMI — force feedback from gears, not sensors

FOMI · preprint · HKUST.

A scissor mechanism with meshed sector gears gives the teleoperator passive force feedback using **no sensors and no actuators**. Four contact-rich tasks, 3 novices × 3 sessions, π0.5 trained on 25/50/100 demos with 22 rollouts. Release promised, not delivered. The honest gap: the forces are *felt* by the operator but **not recorded for the policy**, which, read against FOMO-Dex above, is exactly the thing that determines whether a force-feedback rig buys you anything downstream. Still the cheapest entry point in the teleop cluster if the files land.

The tactile data pile got much bigger, and mostly gated

- TacClaw-Infinity — a tactile UMI handheld rig plus curation pipeline yielding a claimed 10,000+ hours, 1M+ episodes, 1,000+ tasks across 50 sites and 20 organisations, with zero robot in the loop. 3-axis 120 Hz DM-Tac-X sensing. Release links stated, **not verified**. Only 3 evaluation tasks; annotations pass at 83% on a 5% sample; haptic study n=9. The scale is an order of magnitude beyond anything in this ledger and the evaluation is an order of magnitude thinner.
- EgoExplorer (SJTU + Shanghai AI Lab + ZJU) — 100 h wearable vision + whole-body motion + finger articulation + tactile. **Release verified**: CODE and 118,910 parquet rows with a LeRobot loader, gated behind manual approval, no licence named. Its Calibrate-Complete-Constrain pipeline cuts body-point error 71.9% against the strongest baseline, but the references are tracker-fitted, not ground truth, and the humanoid retargeting half is simulation-only and called preliminary by the authors.
- LightwheelAI EgoStandard / EgoDemo / EgoPro — 257/216/195 likes and 8,107/25,727/2,605 downloads. All three gated; cards unreadable from here, so no episode count, sensor spec or licence is verified. Methodological note from the harvester worth keeping: six sweeps sorted the Hub by *created* and saw only uncarded dumps, while the datasets people actually download sit under *trending*.

Also in contact-rich control

HiFun (HKUST-GZ + ETH) reaches 98.3% after roughly **one hour of real online training** with a frozen contact-aware hand skill under a coordination policy, residual RL with a fingertip-force reward and a per-task kinesthetic prior. Six tasks on Franka FR3 + Sharpa + XHand, 20 trials per task, no seeds, authors' own tasks, no code — but one arm, one hour, and a plausible small-scale target if the recipe is ever written down. CQDC-Warp (PKU + Galbot + THU) is GPU differentiable quasi-dynamic contact simulation on NVIDIA Warp for joint motion-and-contact gradient planning; plausibly one-GPU if code appears, and none is stated.

---

3 Platforms & embodiment

Demonstrated, with numbers and a real robot:

- Whole-Body Dexterity (preprint, Amazon + UCSD + Michigan) — a humanoid uses arms and torso as contact surfaces to brace and slide bulky objects rather than treating the hand as the only end effector. RL experts on any-to-any repositioning distilled into one policy taking proprioception and object geometry only; **zero-shot to a real Unitree G1**; 90% against 18% for direct retargeting and 4% for a generalist whole-body controller. Trial counts are not stated and there is no code link on the page, which caps how much weight the 90% carries.
- Contact-UAN (preprint, MIT CSAIL) — a residual actuator network trained with **no torque labels at all**, using proprioception-only contact-consistent reset inference, from 6.5 minutes of real G1 walking. 20% lower torque and 43% lower force-tracking error; mjlab/MuJoCo + BeyondMimic, PPO at 4,096 envs. One robot, one lab, no trial counts, no seeds, no code. If it holds, it removes a torque-sensing prerequisite from actuator identification.
- Compliant GMT (preprint, MIT) — residual reference adaptation makes a general motion-tracking controller compliant in full-body response, as a module rather than a retrain. Abstract-unverified.

Vendor-reported, no independent measurement:

- Gemini 4 Argon for Robotics (vendor blog, dated 10-13) — one hosted model spanning robot control, real2sim from camera feeds, human video *and hand-drawn sketches*, hindsight labelling, dense value and progress estimation, and sim-data generation. The one quantitative claim: generated simulation data lifted real success from 56% to 75% and cut completion from 35 s to 24 s. No trial counts, no weights, no on-device model, paid-tier access only. The ERQA 87% claim carries no comparison figures. Architecturally this is the most ambitious thing in the section and evidentially it is a blog post.
- NEURA Neuron AI (trade press) — multisensory model adding touch, audio and memory, with an eleven-minute autonomous laundry demo *promised* for GTC Berlin on 20–22 Oct. The demo had not run at time of writing, so there is no success rate. Laundry is now the industry demo task; note LoopWAM's shirt folding above and Dyna's towels below.
- Rhoda AI (trade press) — tactile prediction through an inverse-dynamics model driving blindfolded search. Video only, sped up 6× and 12×, no success rate, no trial count, trained on 2 balls and tested on 7, sensor unnamed with no spec, no paper, no code.
- Unitree UniUni (press release) — multi-robot choreography with LiDAR stage mapping. The article itself calls the demo promotional rather than measured.
- Plim Robotics (trade press, company-sourced) — Devanthro rebrands; wheeled home robot with a step-in teleop mode showing the operator's face on the visor. Company-listed specs only (1.65 m, 80 kg, 7-DoF arms, 3 kg payload, Jetson Thor, ROS 2, 2 h runtime). No independent test, no price, no date. The teleop framing is the only reason it is here.

Video is the only evidence:

UniUni, Rhoda, the NEURA laundry demo, and the Pudu D7/D5-W spec announcement. Treat all four as unverified.

---

4 Open source & tooling

Ranked by what reduces build friction *this month*.

1. **RoboParty RP1 hardware** — the week's one unambiguous artifact. Verified present: STEP models and PDF drawings, CNC/3D-print/rubber part files, PCB BOMs with Gerbers, placement data and fabrication requirements, harness BOM and assembly drawings, and per-configuration BOMs for the 5-DoF and 7-DoF MotionHead. **CERN-OHL-W-2.0** for RoboParty-owned hardware, CC BY-SA 4.0 for docs, GPL-3.0 on the umbrella repo (143 stars). MimicLite policy checkpoints at 24 and 28 DoF are listed alongside. Caveats: the ROMOMO joint-module documentation excludes software and firmware, there is no retail price or shipping schedule, and the comparison plot in the announcement is a Unitree G1, not an RP1. Against five weeks of this ledger recording cheap tactile designs that are "two PDFs," this is files on disk under a named licence.
2. **RAX** (Harvard + NVIDIA + Berkeley + MIT + Stanford; Huang, Zhang, …, Abbeel, Torralba, Ermon, Song, Du) — a frozen coding agent learns manipulation by online exploration, with **no demos, no teleop and no policy training**, compiling verified skills that carry preconditions, checks, failure modes and verifying runs. Code and per-embodiment skill libraries

released for Franka FR3 and Rainbow RB-Y1; no weights, because nothing is trained; no licence named on the page. The under-advertised asset is the session log dataset (CC-BY-4.0): 12 complete real-robot sessions with every tool call and error kept verbatim, 60 Hz Franka state, 10 Hz RB-Y1 state **including wrist forces**, RGB plus uint16-mm depth with intrinsics and poses, faces blurred and 13 sessions pruned, both disclosed. The peg-insertion session alone has 5,785 tool calls with **2,241 errors**, tallied by joint-limit, self-collision and force-torque-guard refusal. That is a contact-rich *failure* corpus, which is the scarce kind. The claims are soft: the +37.4% over baselines is on the authors' own harness, the 241× and 103× figures compare agent wall-clock against a compiled program rather than against any policy, and no arXiv ID exists yet.

3. **LTM-pi** — Apache-2.0, training and inference code, cache generation, normalisation-statistics computation, a RoboMME batch evaluator and a WebSocket serving protocol. Weights not bundled; the README links π0.5 base params, a DreamDojo LAM 400k checkpoint and a precomputed memory cache. 0 stars, 7 commits — nobody has touched it yet.
4. **Physical Intelligence API beta** (vendor blog, 10-13; the post itself returned a client error, contents secondhand) — hosted fine-tuning and inference on PI infrastructure with an upload-train-test-collect-around-failures loop. Application-gated, onboarded in batches, **models served are not named**, no pricing, no hardware compatibility list. openpi remains the separate open route. First PI post since 2026-04-16.
5. **Spark-0** — MIT-licensed training and serving code on a vendored StarVLA, with a `pretrain_spark_0` checkpoint on the Hub. Useful as a reference implementation of the contact-map supervision head even if you never touch the hardware.

Negative confirmations, which are also information:

- **LeRobot is still v0.6.1 (2026-08-03)** — fourth consecutive sweep, now eleven weeks post-deal with no release. v0.7.0 remains an open roadmap issue.
- MuJoCo 3.15.0 (10-05), Isaac Lab v3.0.0-EA (09-16, still early access; 3.0 GA only *targeted* for end of October), ManiSkill v3.0.1 (Apr), RoboCasa v1.0 (Feb), Genesis 1.4.4 (10-11, via PyPI because the GitHub releases page keeps returning client errors). **Nothing in window.**
- ROS 2: no release, no new REP. Three third-party LLM-to-ROS bridges were active on Discourse — AgenticROS (most active topic, 18 replies / 1,425 views), ros2_lingua, and DialoStack, whose interesting design choice is keeping the LLM **out of the control flow**. Also ros2_cuda_ipc for zero-copy GPU buffer sharing between ROS 2 processes, and Sub-Millisecond Reflexes (TUM + MBZUAI), which reworks the ROS 2 control path so behaviour adapts without stop-to-switch controller transitions. All third party, none benchmarked.

5 Money & moves

Boston Dynamics appoints Rohit Prasad CEO (trade press, dated 10-06, effective 10-07 — ten days old and missed by four sweeps). Twelve years at Amazon on Alexa and Nova, fourteen before that at Raytheon BBN; he replaces an interim CEO who was the CFO, after Robert Playter stepped down in February. **INFERENCE** \2192 a company whose identity is hardware excellence just hired a foundation-model executive rather than a roboticist. Read with the Atlas build-time admission below, the bet is that the remaining hard problem is the model and the manufacturing line, not the machine.

SafeWorld raises USD 12.2M seed (trade press, dated 10-08). Co-led by Shine Capital and a16z Speedrun, with BoxGroup, the CMU Endowment, Innovation Endeavors and SV Angel; co-founder Ding Zhao directs CMU's Safe AI Lab, so this is a CMU RI-adjacent spinout. The product builds scenarios from incident reports, standards and robot logs, and replays them against thousands of reactive human-motion variations on every software update. **INFERENCE** \2192 this is regression testing for policies, sold as safety. Someone has decided the bottleneck on deployment is not capability but the inability to prove a policy update did not break something — which is the same gap the benchmark-validity thread has been tracking from the research side. No paper, no code, no fidelity claim.

Agility Robotics, three items in one week. SEC S-4 declared effective with a Nasdaq merger vote set for 2 November; Nyobolt named for a nine-minute Digit 5 recharge, early access 2027; and a reported USD 300M order sourced only to a newsletter headline with no customer, unit count or schedule. **INFERENCE** \2192 against the ~USD 150,000 Digit 5 BOM carried from W41, a 1,000-robot order is roughly USD 150M of hardware cost against USD 300M of revenue, before any of the integration

labour that Dyna's chief scientist spent a whole talk describing. The charging deal is the tell — duty cycle, not capability, is the variable being optimised for public-market scrutiny.

Shanghai targets 8,000 humanoid robots in factories by 2028 (trade press). A dated municipal number, which is rarer than it sounds. **INFERENCE** \2192 a government target is a demand-side commitment, not a shipment figure, but it sets the denominator Chinese vendors will be measured against and it explains the volume of Chinese loco-manipulation data papers in this ledger (λ0, KODA, EgoMatrix, Spark-0, TacClaw-Infinity).

Boston Dynamics: an Atlas takes a day to build, target one hour (trade press). Stated goal, no timeline, not achieved. **INFERENCE** \2192 naming prototype-shop throughput as the constraint on iteration speed is a research-velocity statement disguised as a manufacturing one. You cannot run a data-collection fleet on a day-per-unit build.

Google DeepMind is building a Cambridge-area robotics lab (trade press, 10-12, single-sourced to Aaron Saunders, the former Boston Dynamics CTO). No headcount, no timeline, no charter, not on a DeepMind surface. **INFERENCE** \2192 combined with Gemini 4 Argon shipping as a hosted robotics model, DeepMind is building a physical-robotics site under a hardware executive while selling the model layer as an API. That is the Figure-and-OpenAI shape — own the embodiment, rent out the brain — appearing at the one company that previously did neither.

---

6 Deep read of the week

ForceBench: Benchmarking Force Integration for Robotic Manipulation · preprint, 14 Oct · PKU-PI Lab (Peking University) + UC Berkeley + NUS + ePyBot Intelligence · Han, Jin, Li, Lou, Liu, Xue, He, Yao, Zhan, Wu, Pang · project page forcebench-platform.github.io

Chosen over Spark-0 deliberately, for the same reason Mulligan was chosen over RoboJEPA last week. Spark-0 is the bigger claim; ForceBench is the one whose result you can act on tomorrow, on one arm, and defend under follow-up questioning. It is also the first controlled comparison this ledger has seen in the force cluster after seven weeks of single-lab force papers, every one of which proposed an architecture and none of which was comparable to any other.

Setup

Task set. Eight simulated manipulation tasks across three deliberately separated categories — *physical-property inference* (Weight Sort, CoM Sort: distinguishing objects by mass or centre-of-mass that look identical), *resistance-dependent interaction* (Faucet Direction, Faucet Turn, Stove Press, Stove Turn, Drawer Pull: stopping or switching on a felt resistance transition), and *contact-rich assembly* (Nut Thread). Parameterised physical mechanisms — mass, centre of mass, resistance profile, initial pose — are instantiated from seeds, so a task instance is reproducible.

Embodiment. Simulation: one arm with a TCP-pose action space, control at 15 Hz over 60 Hz physics, 224×224 external and wrist RGB, 7-D proprioception, and a six-axis wrench read as the parent-on-child reaction at the wrist-hand joint in joint frame, **without gravity compensation**. Real: Franka Research 3, KWR75B six-axis F/T sensor, Robotiq 2F-85, external plus wrist cameras.

Data. 1,600 successful demonstrations, 200 per task, 1.0033M control steps, generated by task-specific scripted experts that use privileged simulator state through the shared controller; only trajectories meeting the evaluation criteria are kept, and privileged state never reaches the policy. Real: 200 demonstrations per task.

What is held out. Training instances use seeds 0–199; evaluation uses seeds 300–499, shared across all methods. Real evaluation uses the same 20 initial conditions for every method. A separate policy is trained per task. Total: 10 systems × 8 tasks × 200 episodes = **16,000 simulated episodes**, plus 8 systems × 5 tasks × 20 trials = **800 real trials**.

Method

There is no new model here; the contribution is the matrix. One π0.5 backbone, one demonstration set, one training protocol (10,000 steps, batch 32 in sim and 64 on hardware), and seven configurations that differ only in how the wrench enters:

- **No Force** — baseline; images, language, proprioception.
- **Token Fusion** — an MLP maps the wrench to six continuous tokens placed before the action tokens in the action-expert suffix.
- **Prompt Fusion** — the wrench is appended to the robot state, normalised, and discretised into numerical tokens *in the language prompt*. No new parameters.
- **Cross-Attention** — force tokens enter a cross-attention layer between the action-expert output and the action projection.
- **Token Fusion + Pred** — Token Fusion plus a ten-step future-force prediction head, training only, auxiliary weight 3×10^{-4}.
- **Temporal Prompt** — reads the last 16 samples (~1 s), appends the current sample plus per-axis mean and standard deviation to the state; state grows from 7 to 25 numbers; **no parameters added**.
- **Dual-Path Fusion** — Temporal Prompt plus a second pathway feeding 25 masked window statistics through a two-layer MLP into four continuous tokens, with the window zeroed on 30% of training samples so the policy can still act from the prompt alone.

External references, not matched: π0, ForceVLA, ForceFlow.

Results

The number that matters is not the winner's score, it is the ordering. Simulation overall mean success: **Prompt Fusion 35.31%**, Dual-Path Fusion 35.13%, Token Fusion + Pred 26.25%, Token Fusion 23.56%, **No Force 23.25%**, Cross-Attention below the baseline. Hardware: **Prompt Fusion 63.00% and Dual-Path Fusion 63.00% against No Force 40.00%**, with ForceVLA at 33.00%.

Three readings:

1. **The cheapest interface wins.** Discretising the wrench into the state string adds no parameters, no encoder and no attention, and it is the only design that improves on every one of the five real tasks. Everything with an architecture attached either matched it or lost.
2. **Category means hide inverted trade-offs.** Dual-Path leads physical-property inference at 85.75% against the baseline's 49.75% — and scores **16.50% on assembly against the baseline's 31.50%**. On real Block Insert, Dual-Path drops to 20% where No Force gets 50%. Temporal windows help when the informative signal is a sustained property and hurt when it is a transient geometric constraint. Two methods tied at 63.00% overall have opposite failure profiles.
3. **Success rate and intermediate progress come apart.** Prompt Fusion and Dual-Path both reach the Drawer Pull opening range with an intact latch in 81 episodes, and complete the task 46 and 9 times respectively. On Weight Sort, Prompt Fusion's gain is almost entirely a decision effect — wrong destinations fall from 86 episodes to 17 while incomplete placements barely move (6 against 7). Adding predictive supervision to Token Fusion does the opposite: incomplete placements fall 84 → 14 and wrong destinations rise 15 → 30.

The baseline worth naming: **π0 with no force at all reaches 36.00% on Nut Thread, the best assembly score of any of the ten systems.** The assembly category's winner is a model that cannot feel.

Where it's soft

- **The authors disclaim their own headline, and you should quote them doing it.** Appendix C states that "highest observed" denotes the ordering of recorded point estimates, not an established significant advantage. Prompt Fusion and Dual-Path differ by **three episodes out of 1,600** — 0.19 pp. No paired significance tests, no training-seed variance, one seed per configuration.
- **One task in the assembly category.** Every assembly claim above, including the inverted trade-off that makes the paper interesting, rests on Nut Thread in sim and Block Insert on hardware. The authors say this should not be generalised; it is also the finding most likely to be cited as if it were.

- **20 real trials per task-method pair.** One trial is 5 pp. The 63.00 against 40.00 gap is 23 trials out of 100 and is probably real; the Prompt-versus-Dual-Path tie is not a measurement.
- **The designs are not clean single variables.** The paper admits the input designs jointly vary representation, normalisation, capacity and insertion location, and that the temporal systems are "separate references, not a controlled history-length ablation." So "prompt beats tokens" could be about discretisation, about normalisation, or about where in the stack the signal lands. Only the predictive-supervision comparison is genuinely one-variable.
- **Compute asymmetry against the external references.** ForceFlow trains for 100,000 steps against everyone else's 10,000 and executes 32 of 64 predicted steps against everyone else's 50 of 50 — and scores 110 successes out of 1,600, with 0.00% on assembly. That is not evidence about ForceFlow; it is evidence that the harness is not neutral to execution horizon. The paper says as much.
- **The wrench is a joint reaction without gravity compensation in sim and a raw sensor-frame reading on hardware.** The authors explicitly warn these are not isolated contact-force estimates and should not be compared in magnitude across the sim-real boundary. So the benchmark compares *interfaces to a force-like signal*, not force.
- **No release verified.** Project page only; I found no repository or dataset.

Steal list

1. **Try the free one first.** Before building any force encoder, append the normalised, discretised six-axis wrench to your state string and retrain. It costs one line in the observation builder and it is the configuration that beat everything here. If it does not move your task, the architecture probably will not either.
2. **Split your eval by what the force is telling you.** Property inference, resistance transition, and geometric constraint are three different regimes and a method can win one while regressing on another. An aggregate success rate across a mixed task suite will hide exactly the trade-off you need to see.
3. **Record intermediate-state outcomes, not just success.** Destination choice, mechanism progress, and final assembly geometry logged separately are what turned an ambiguous 63-versus-63 tie into a legible result. This costs nothing at collection time and is irrecoverable afterwards.
4. **Temporal windows are not free.** A 16-sample window of per-axis mean and standard deviation is the cheapest possible history and it still cost 15 points on assembly. If your task is insertion, start from the current sample.

Cost to reproduce a scaled-down version

A two-condition replication — No Force against Prompt Fusion — on two real tasks, one property- inference and one insertion:

- **Hardware:** any 6/7-DoF arm, a parallel gripper, a wrist six-axis F/T sensor, one external and one wrist camera. The F/T sensor is the only non-obvious cost; a KWR75B-class unit is low-four-figures new. INFERENCE \2192 a joint-torque-derived wrench on a torque-sensing arm is the cheap substitute, and given the paper's own caveat that it never had a calibrated contact-force estimate anyway, that substitution is less lossy here than it would normally be.
- **Episodes:** 200 teleop demonstrations per task, so ~400 total. At a conservative 1 minute per demonstration including resets, that is roughly 7 hours of collection, realistically a day and a half with setup.
- **Training:** π0.5 fine-tuning, 10,000 steps at an effective batch of 64. INFERENCE \2192 on a single 80 GB card with gradient accumulation, ~40–80 GPU-hours per policy; four policies (two configurations × two tasks) puts the experiment at **roughly 160–320 GPU-hours**, or under a week on one rented H100. On a 24 GB consumer card you are LoRA-only and the comparison stops being matched to the paper.
- **Evaluation:** 20 paired initial conditions × 2 methods × 2 tasks = 80 real rollouts. Add seeds — the thing this paper lacks — and it is 240 rollouts for three training seeds, which is two days of operator time and would make your version *stronger evidence than the original*.

That last point is the opportunity. A three-seed, two-condition replication of the single cleanest claim here is a weekend of compute and two days of robot time, and it would be the first seeded result in the force-integration literature.

7 Relevant to what you're building

`CONTEXT.md`'s `## Active build` section is still an unfilled placeholder — **sixth week running**, and the seventh consecutive harvester sweep to record it. Per the standing rule, I am not manufacturing workstreams to connect items to. Section skipped.

The cost is no longer abstract. Seven sweeps have produced roughly 300 scored queue lines, and **nothing has scored 5 in the entire W42 cycle**, because the score-5 clause requires an item to be load-bearing for a named project. This week specifically, the three items that would most likely have hit a real build — ForceBench's prompt-fusion result, FOMO-Dex's finding that force corrections must be in the data rather than the controller, and the RP1 hardware drop — are ranked on generic interest rather than on whether they change anything you are holding a soldering iron over.

8 Market signal

Boards checked 2026-10-18.

- **OpenAI: 19 robotics roles**, identical to the 10-11 count. Titles enumerated in full this week: control systems, electrical, actuator technician, field engineer, inference, safety, motors and actuators manufacturing, motor design, dynamometer test infrastructure, PCB layout, procurement, prototyping lab, robotics software, EHS, distributed data systems, ML systems and training architecture, commodity management, and two TPMs. **Fifth consecutive week with zero titles containing policy, VLA, imitation, robot learning or manipulation.** Five weeks is not a snapshot.
- **Figure: header reports 100 roles**, against 96 on 10-11 and 97 on 10-04 — the first clear increase in three weeks. Pagination still defeats a clean Helix enumeration; page 1 surfaced one Helix role matching the keywords (*Helix AI Engineer, Video Pretraining*), and the *Reinforcement Learning Engineer, Whole Body Control* role remains filed under Controls, not Helix. **I make no claim about whether the Helix count moved.**
- **Physical Intelligence: 12 roles**, same total and same composition as 10-11 — ML Infra (Data Systems), ML Infra (Platform), Controls, Forward Deployed Robotics, Build & Release, Robot Build Technician, Robot Operator, Manufacturing, NPI TPM, Robotics Software, Business Operations, Mechatronics Intern. **Still zero policy-research titles.** Reading the bodies rather than the titles this week: the Controls role is the closest match, listing dynamics-model building and naming "robot learning" and "learned policies" as desirable; the Robotics Software Engineer posting states explicitly that the role does **not** involve designing ML models. Two weeks of a deployment org, now with the API beta (§4) to explain it.
- **New surfaces this week:** DeepMind's Cambridge-area physical-robotics lab under Aaron Saunders (§5) and SafeWorld post-seed. Neither has a public board yet. Boston Dynamics under a new CEO is the third.

Comp datapoints, primary postings:

- **Skild AI, Deep Learning Manipulation Engineer: USD 100,000–300,000 base**, San Mateo, fully in office. Requirements are notably generic — a degree or equivalent experience, Python plus one DL framework, familiarity with RL and/or imitation learning, distributed training experience. New datapoint this week and the only published band attached to a title with "manipulation" in it anywhere in this ledger.
- **OpenAI, ML Engineer, Distributed Data Systems – Robotics: USD 380K–500K + equity**, SF. **Unchanged for five weeks.**
- **Figure, Helix AI Engineer, Modeling: USD 200,000–400,000 base**, San Jose, five days in office. Unchanged.
- **Physical Intelligence** publishes exactly one rate, again: **Robot Operator at USD 25/hour**. Compensation display is off on every engineering posting.

INFERENCE \2192 the Skild band is the interesting one, and not because it is high. A 3× spread from 100K to 300K on a single manipulation title means the band is a level ladder printed as one posting — the company is hiring anywhere from a new graduate to a staff engineer against the same requirements list, which is what a team does when it does not yet know what seniority the problem needs. Set against OpenAI's 380–500K for a *data pipeline* role, the ordering from W40 and W41 is unchanged and now

five weeks stable: **the highest published number in applied robotics is still attached to data infrastructure, not to policy research.**

Skills recurring in the work that got attention this week:

1. **The controlled matrix, and the discipline to disclaim it.** ForceBench's Appendix C — stating in plain language that "highest observed" is a point-estimate ordering, that no paired tests were run, and that training-seed variance is not quantified — is the single most imitable page published this week. Spark-0's four-variant table (Scratch / Visual PT / 100H / 1000H on shared RGB and shared compute) is the same instinct. Being able to say exactly what your own table does not establish is now the thing that separates a credible ablation from a leaderboard entry.
2. **Deleting the sensor at inference.** Spark-0's train-only contact supervision, FOMO-Dex's training-only correction, ForceBench's training-only future-force head, LTM's frozen latent-action encoder. Four papers this week in which the expensive signal is paid for once, during training, and never appears at deployment. Being able to reason about which signals can be amortised this way is a distinct competence from knowing how to fuse them.
3. **Reporting the number on hardware the reader owns.** LTM-π at 66.71 ms on an H100 and stated to run on a single 4090; Dexterity from Supervised Play fitting perception and planning on one RTX 5090 and distilling to 21.8 ms; SmolWAM at 3.6 A100-hours. Still rare, still differentiating.
4. **Treating failure data as an asset.** RAX shipping 2,241 logged errors with refusal reasons; LoopWAM training on mixed-quality deployment rollouts with quality steered through the instruction; Beyond Successful Robot Demonstrations modelling failure dynamics rather than discarding unsuccessful episodes.

Table stakes, updated. Released code at submission — cleared this week by LTM-π (Apache-2.0), RAX, Spark-0 (MIT), RoboParty RP1 (CERN-OHL-W-2.0), EgoExplorer (gated), Reflex; **not** cleared by ForceBench, FOMO-Dex, FOMI, Dexterity from Supervised Play, Whole-Body Dexterity, LoopWAM, Chronos Policy, Physical Time, Contact-UAN. End-to-end latency rather than head latency. An ID/OOD split. Knowing which benchmarks are compromised. New entry: **evaluating on a benchmark you did not build** — LTM used RoboMME and LIBERO-Mem, LoopWAM used RoboChallenge; Spark-0, ForceBench, Evo-DexBench, WorldMemBench and Orbia all score on their own.

Still differentiating. Seed bands on real hardware, which remain so rare that no paper this week has them. A paired test on real trials. Publishing evaluation rollouts and not only training data (Mulligan, W41, is still the only instance in this ledger). An explicit statistical-scope section that names what the numbers do not support. And shipping fabrication files under a named licence, which until RP1 this week had not happened at all in six weeks of tracking.

9 Threads

`THREADS.md` updated: no new threads, no closures, 12 open and at cap. Two candidates are queued behind the cap and recorded there with promotion conditions.

THREAD	STATUS	WHAT WOULD COUNT AS THE NEXT REAL UPDATE
World-action models as unified policy + simulator	**ADVANCING**	RoboHarness read in full — it claims a simple harness beats both VLAs and WAMs, and if it holds it reframes the entire W41–W42 cluster. Still the W40 question: a matched t_denoise sweep on a WAM *with* embodied pretraining.
Planning on a learned model	**ADVANCING**	Diffusion steering on real hardware. Back on Track, VeriVLA and PILOT all landed this week and all are abstract-unverified.
Where the episode starts	**ADVANCING**	Dexterity from Supervised Play's autonomous-play ladder (0.54 h → 5.11 h, monotonic MAE reduction) is the week's entry; DexMimic and CoDA extend reset/environment allocation. Next: a third-party replication of Mulligan's Sobol-plus-farthest-point sampler alone.

THREAD	STATUS	WHAT WOULD COUNT AS THE NEXT REAL UPDATE
Force as a first-class action	**advancing — bar cleared**	ForceBench is the controlled comparison this thread has wanted for seven weeks. Next: the same matrix with **training seeds** and more than one assembly task; and the still-unrun falsifier — a wrench-action policy trained on position-teleop data matching one trained on force-feedback teleop data. FOMO-Dex's 0%-versus-67% result is evidence *against* that falsifier succeeding.
Contact sensing without dedicated tactile hardware	**advancing — partially met**	Spark-0 is the closest anything has come to the seven-week-old bar. It is not a clean hit: the policy deploys with no sensor, but its training labels *were* collected with one. The clean version is now one experiment away — see §10.
Paying at training time, not run time	**ADVANCING**	Spark-0, TacClaw-Infinity, EgoExplorer, LightwheelAI, λ0, KODA and EgoJEPA all pay at training time this week. Next: whether any cheap-tactile release lands **with a licence**, and the five-week-unrun matched comparison of one-step action heads on one backbone.
Benchmark validity in physical AI	**ADVANCING**	ForceBench's statistical-scope appendix is the week's model. LIBERO-RIG (cameras bumped, calibration drift, controller-rate changes) is the most build-relevant new benchmark. Still unmet since W38: a major-lab table with seed bands **and** paired evaluation together.
Online correction as the last mile	**ADVANCING**	LoopWAM, Back on Track and Value as Memory all attack deployment feedback. Next: a third-party rerun of EXPO-FT, now six weeks out with no independent run by anyone.
Low-cost open hardware as training substrate	**advancing — first artifact in six weeks**	RP1 shipped STEP, Gerbers, harness and BOMs under CERN-OHL-W-2.0. The bar is unchanged and now actually reachable: an independent group reporting success rates on released low-cost hardware, with code. TACROSS at USD 10.86 remains the highest-value target the moment its files exist.
Claimed open releases vs actual artifacts	**narrowing**	Delivered and verified: LTM-π, RP1 hardware, RAX code and logs, Spark-0 code and checkpoint, EgoExplorer. Pledged and not delivered: HumanTouch at 1,000 h (100 h public, expansion lapsed), FOMI, Dexterity from Supervised Play, ForceBench, TacClaw-Infinity (links stated, unverified). Carried: TACROSS, SoTa, TouchScale, Praxis-1, StreamPI. Next: whether HumanTouch's 1,000 h actually lands.
Who owns the open robot-learning stack	**ADVANCING**	PI opened a gated hosted API; DeepMind shipped a closed frontier robotics model across the whole lifecycle; LeRobot is eleven weeks silent; RoboParty answered with open hardware. Next: the first post-deal LeRobot release, and whether any of the five platform-holder commitments is written into a foundation or steering body.
Simulation fidelity for transmission and contact	**STALLED — kill date is W43, next week**	Nothing in window: no MuJoCo release, Genesis unreachable for a third sweep. CQDC-Warp and DexNinja are the only new entries and neither states a release. If no sim-to-real result crediting the discrete integrator, IPC flex contact or routed-cable modelling appears by next week, narrow to the reproducibility finding and kill the contact-fidelity half.

Queued behind the cap, with promotion conditions:

- *When to replan, and when to think* — this is now the largest uncatalogued cluster in the ledger: Perturbation Informed Execution, Policy-Aligned Predictive Verification, When to Ask for Help (which proposes a measurement standard, not just a method), Fast-VLA, FastVLA, 1Step-WAM, VeriVLA, and RACE from W41 — eight papers in nine days on the same question. **Promote the moment the simulation-fidelity thread is killed at W43**, which is the slot it is waiting for.
- *Human video as the data substrate, and where it saturates* — carried from W41 unchanged. Promote when a second group reproduces TRI's 30,000-hour saturation finding on a different corpus. This week added volume (λ0, KODA, EgoMatrix,

EgoCPT, EgoJEPA, Human Demonstrations as Post-Training Data) but no replication.

10 Open question

Is Spark-0's cross-embodiment gain a contact-supervision effect, or a data-matching effect?

The claim is that touch supervises visual representations and that the benefit survives into deployments with no tactile sensor. The evidence is a 19.27-point jump on Ego Humanoid from Visual PT to 1000H on identical RGB data at identical compute. Three things make that reading unstable.

First, **the baseline moved the wrong way.** Visual PT scores 52.45 on Ego Humanoid against Scratch's 58.34 — pretraining on this mixture, with touch stripped, actively *hurt* transfer to the unseen embodiment. Measured against no pretraining at all, the full tactile model gains 13.38, not 19.27. When a comparison's denominator is below the do-nothing control, the gap is partly a story about the denominator.

Second, **100 hours buys almost nothing.** 100H reaches 59.42 against Scratch's 58.34: +1.08. The entire effect appears between 100 and 1,000 hours, with no rung in between, which is a scaling claim drawn through two points.

Third, **the two roles of touch are never separated.** Contact-map supervision and the tactile input pathway are present together in both the 100H and 1000H variants, and the frozen geometry teacher is in all three pretrained variants. The paper's own report concedes this. So "contact supervision improves visual representations" is one explanation; "data containing touch happens to be better-matched, better-curated manipulation data, and the contact head is along for the ride" is another, and nothing in the table distinguishes them. The paper gestures at a third — that contact supervision aids transfer to an *unseen visual domain*, which is consistent with the effect being large on Ego Humanoid and ~zero on SparkArena-Sim (+0.29), where the target embodiment already appears in pretraining.

Why this matters beyond one paper: ForceBench, in the same week, found that the *interface* through which contact reaches a policy is worth approximately nothing, while the *presence* of the signal is worth 12–23 points. Spark-0 claims the opposite shape — that a specific supervision channel, applied to a specific layer, is where the value lives. Both cannot be the general case. If Spark-0's gain is a data effect, the two results agree and the field's instruction is simple: get contact data, do not engineer around it. If it is genuinely a representation effect, then there is a channel that matters and ForceBench measured the wrong axis.

What would settle it. Four pretraining runs on the fixed 1,000-hour mixture, everything else matched, varying two binary factors independently: contact-map loss on/off, and tactile input pathway on/off. Evaluate all four on Ego Humanoid with no touch at test time. The (off, off) cell is the critical one — it is the full HumanTouch data with both tactile mechanisms removed, which is *not* Visual PT, because Visual PT discards the touch-bearing episodes' advantage along with the signal. If (off, off) lands near 71.72, the gain is data. If it lands near 52.45 and (on, off) recovers it, the contact head is doing the work and Spark-0's framing is right. Add a 300-hour rung to any one column and the scaling claim stops resting on two points.

The cost is the honest part, and it is bad: this is four 400k-step pretraining runs at 96 GPUs, which is roughly four times the compute the paper already spent and well outside anything this reader can run. But a scaled-down version is tractable and would be informative: the same 2×2 at 100 hours of HumanTouch — the subset that **is** public — on a smaller backbone, reporting only whether the (off, off) and (on, off) cells separate. If 100 hours cannot resolve them, that is itself the answer to why 100H bought one point.

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

Nothing scoring 4 vanishes silently. All preprints unless noted; abstract-level unless the queue recorded a full read. The first three are **backfill** from the 10-08/10-12 announcement cycles the W41 window never enumerated, not new work this week.

- DreamTrue — CAS-IA + Alibaba; renders action trajectories into image-space conditions with geometric calibration, then counterfactual post-training against a video reward model. Defect rate 48.12 → 6.25; AgiBot World Challenge 2026 winner.

Code at `github.com/brave-eai/DreamTrue`.

- LeWAM (JEPA variant) — Brown, Balestriero group; joint next-latent prediction and action flow matching under an isotropic Gaussian latent constraint, one model serving both direct policy and planning. Contact-rich 28.6 → 89.7, long-horizon 10.9 → 46.2, 32.3× faster planning. Distinct record from the 2610.12407 LeWAM in the W41 ledger. No code stated.
- CASTER — Virginia Tech; recasts synthetic-video-to-trajectory as constrained optimisation needing a semantic basin of attraction, and shows prior SOTA is this formalism with simplifications removed. +34.1 pp average over the best baseline; CASTER-scenes dataset public.
- MagCilia — NUS + CityU HK; compact magnetociliary tactile sensor reading normal and tangential force plus surface texture. Cost and fabrication not verified, no release stated. Flagged because it is the only standalone sensor-hardware paper in two weeks.

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