This Week in Robotics

Window: 2026-08-24 → 2026-08-30. Issue one.

THE LINE

A 0.68M-parameter predictive-coding policy from Sony CSL beat parameter-matched Transformer and LSTM baselines by 4–7× on LIBERO, which is a direct attack on the premise that language-conditioned control needs VLA-scale parameters — while capital moved the opposite way, with SoftBank reportedly negotiating control of 1X at ~$6B.

1 Policy learning & VLAs

PredVLA

PREPRINT / Sony CSL Tokyo

Architecturally new: observations never enter the recurrent dynamics. A hierarchical multi-timescale RNN (PV-RNN lineage) predicts visual features and proprioception; observations act only through gradient descent on latent free variables over a sliding window at test time — error regression, not an input pathway. Setting the inference step count to zero gives an *exact* open-loop ablation on the same weights. Front end is frozen and off-the-shelf (ResNet18 + MiniLM + fixed PCA), so 675,732 trainable parameters are the entire learned controller. Eval: LIBERO only, sim only, one embodiment, separate training per suite. No code. Unreproduced single-lab claim. Full treatment in §6.

StreamPI

PREPRINT / HKU + ACE Robotics

Adds temporal context to a single-frame VLA (π0.5) with no added parameters — instruction-anchored streaming over a T-frame window rather than a separate memory module. The “zero extra parameters” claim means temporal aggregation reuses the existing attention budget rather than adding a recurrent head. LIBERO 98.3% at T=5 (+2.8 pp Goal, +2.6 pp Long over π0.5 — near noise on a saturated benchmark); the real-robot side is the case, at +26.7 to +36.6 pp across four memory-dependent and precision tasks, e.g. cup insertion 92% vs. 60% over 25 trials. Small-n and single-lab, but that gap is outside plausible seed variance. REPO has real code (openpi/LeRobot based); README as of 2026-08-30 says weights still pending. Not a one-GPU reproduction until they land — π0.5 finetuning is not a from-scratch budget.

Riemann-1.0

PREPRINT

One causal autoregressive model as both policy and action-conditioned simulator over multi-view video, state and action; claimed 200K+ hours spanning egocentric human video, handheld-gripper demos and heterogeneous robot data. Reported: RoboTwin2.0 94.3%, LIBERO 99.0%, RoboCasa-365 62.6%, real-world 85.0% success / 94.4% progress, ~15% over the strongest open baseline. The queue tagged this “open-sourced”; I could not verify it — the abs page names no repository and no matching HF repo exists. **Treat the release claim as unconfirmed.** A fleet-data result either way.

CLAP

PREPRINT / PROJECT

Cross-embodiment action-conditioned video world model: latent actions learned from unlabeled video, grounded by end-effector poses, spanning DROID, Bridge, bimanual YAM, a G1 humanoid and EgoDex. The latent-action space is the shared interface and EE pose the only embodiment-specific grounding — which is why novel-embodiment adaptation is claimed within one gradient step. SVD backbone, OXE + EgoDex, 100K steps. CODE and CHECKPOINTS linked. The number that matters: under 12 GB VRAM, runs on a 3060.

2 Manipulation, dexterity & sensing

VISTA

PREPRINT

Reads contact off the *visual deformation* of a compliant gripper — a 3D displacement field from ordinary cameras — fed back as incremental gripper corrections. The point isn't accuracy vs. a tactile array; it's deleting the tactile hardware and its wiring. Three tasks (cross-scale grasping, cap unscrewing, calligraphy), beating 3D Diffusion Policy and a tactile baseline. Three tasks is thin and no code is stated — project materials are a Google Sites page. **INFERENCE** `\2192` cheapest path to a contact signal for a small build, if your gripper deforms visibly and repeatably.

TacForcing

PREPRINT

Puts tactile feedback inside streaming action generation rather than in a reactive layer wrapped around it. Same architectural instinct as StreamPI, applied to touch.

Lie group-constrained MeanFlow grasping

PREPRINT

Flow matching on $SO(3)\times\mathbb{R}^3$ with an algebraic semigroup-consistency condition: ≤5 network evaluations, millisecond grasps, up to 39× over diffusion/flow baselines on ACRONYM, claimed direct real-world transfer. No code. The transferable idea — semigroup constraint as a few-step sampling trick on a manifold — is not grasp-specific.

Also (all preprints): a tendon-driven five-fingered hand with dual-modality tactile per finger, PRISM GPU sampling-MPC with QP projection for bimanual contact, and on-robot juggling learned in minutes — the last a reminder that online adaptation on hardware can beat closing the sim2real gap.

3 Platforms & embodiment

Thin week. **Microduck** (PRESS): $399, 25 cm, 15-DOF biped from Pollen/Hugging Face, RK3566, ToF lidar, two IMUs; SDK, sim, and RL training stack open on GitHub, hardware files not. Ships before Christmas 2026. Capability shown is video only (waddling, beak pickup, fall recovery, roller skating) — demo-video evidence, staged-vs-demonstrated unknown.

SOLO (perceptive humanoid locomotion) and Poppy LQR walking on low-cost open hardware are the only platform preprints with method behind them. No humanoid company published a technical result in window.

4 Open source & tooling

Friction-reducing this month, in order: **CLAP**'s 12 GB world model (above), which makes learned dynamics testable on hardware you own. **CRESSim-Neo** (preprint), GPU-batched deformable/surgical sim with PyTorch hooks. **Sharpa's tactile deformation encoder** — Apache-2.0 ConvNeXtV2 weights from a sensor vendor; small, but pretrained tactile features from someone with real data are rare. **Muse-Robotics-1**, 121.7M-param MIT-licensed VLA, no paper, no card, unverified.

Nothing in window from LeRobot, MuJoCo/MJX, Isaac Lab, ManiSkill, RoboCasa or ROS 2 core. Genesis' release page has 404'd for three consecutive sweeps — status unknown, not "no release."

The eval item worth your time: statistical audit of physical-AI benchmark redundancy (preprint): 51 models × 12 benchmarks; collapsing two substitute pairs moves 22 of 51 models by three or more places, and four benchmarks retain 78.5% of the ranking utility of all twelve. Read as: most benchmark tables you see are reporting the same axis several times.

5 Money & moves

GENERALIST	1X	NSF CENTRE	LOCUS / NEXERA
$3B	**~$6B**	**$30M**	**n/d**
+50% in 10 weeks	SoftBank majority-stake talks — unconfirmed	over ten years, UT Austin	terms undisclosed
$200M extension led by 8VC			

Sources per item below — press and trade press.

Generalist → $3B

PRESS

$200M extension led by 8VC, up from $2B in June — 50% in ten weeks, on Gen 1.5's claim of learning tasks from 3–12 second video demos. **INFERENCE** \2192 capital is pricing *demonstration efficiency*, not task coverage; the pitch that moved is "fewer episodes per skill."

SoftBank ↔ 1X, ~$6B, majority stake

PRESS / reported by The Information; SoftBank declined comment, 1X did not respond — unconfirmed

INFERENCE \2192 a *majority* stake is not a growth round. If it closes, it reads as a strategic owner absorbing a humanoid platform rather than a market marking one up.

Locus acquires Nexera

TRADE PRESS

NeuraGrasp — soft end-effector combining suction and pinch, six product generations — goes onto Locus' Array mobile manipulator. **INFERENCE** \2192 the logistics incumbents have concluded manipulation is bought, not built, and what they are buying is end-effector hardware, not policy.

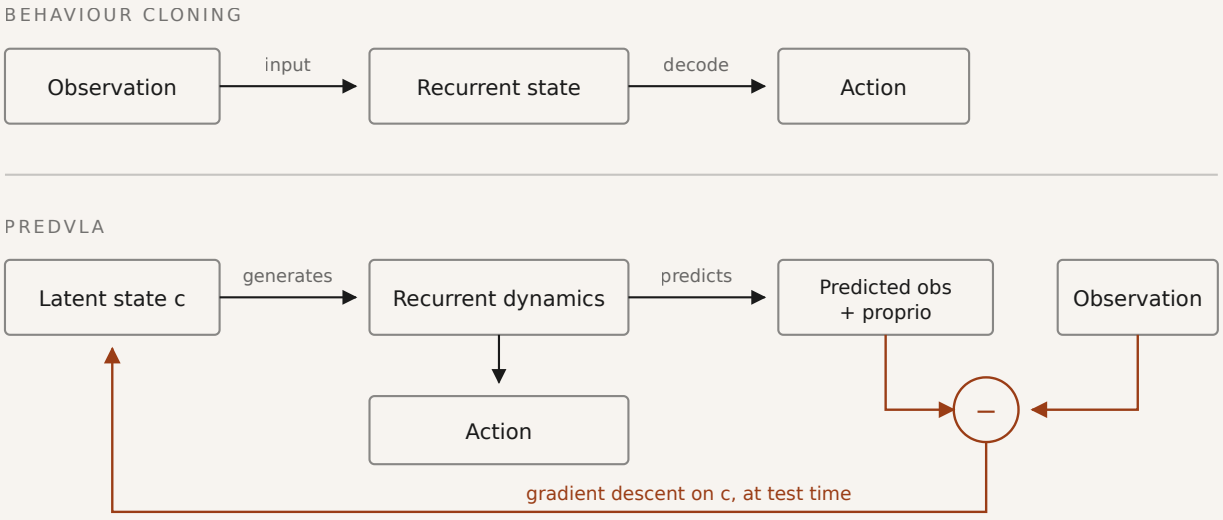
EXL/iMerit folds a physical-AI annotation shop into an analytics firm; **NSF** put $30M over ten years into a UT Austin center on robots and people learning together. **INFERENCE** \2192 the data-labeling layer is consolidating into enterprise services while long-horizon HRI retreats to federal money.

6 Deep read: PredVLA

arXiv:2608.26673 — preprint, Sawada & Kasahara, Sony Computer Science Laboratories.

Setup. LIBERO, all four suites, official demos, Franka in sim, 14 seeds, each suite trained separately. Held out: nothing beyond LIBERO's own splits — no cross-embodiment, no hardware.

Method. Inherited: multiple-timescale RNN and PV-RNN's prior/posterior free-variable formulation; frozen ResNet18/MiniLM front ends; Gaussian-mixture action head. New: (a) running it as a *language-conditioned* policy under the modern VLA protocol, which predictive-coding sensorimotor work had not done; (b) the strict rule that observations touch the model only through free-energy minimization over latents, which is what buys the exact open-loop ablation; (c) two lateral pathways — a low-dim visual→action bottleneck, and an efference copy of the model's own predicted action and proprioception fed back to the visual branch, so no measurement enters forward dynamics.



The architectural rule the paper is built on. In behaviour cloning the observation is an input to the recurrent state; in PredVLA the dynamics generate a prediction and the observation reaches the latent state only as a gradient on prediction error. Setting the number of gradient steps to zero severs the orange path entirely, which is what makes an exact open-loop ablation possible on unchanged weights.

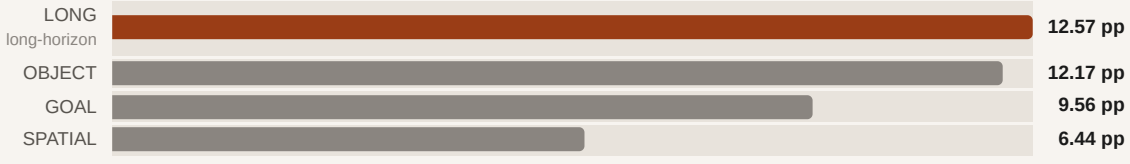
The number that matters. Not the 75.35% four-suite mean — the controlled comparison. At matched parameters and identical front end, demos, action head and protocol: BC-Transformer 19.73%, BC-RNN 10.26%. Action-chunking Transformers with *more* parameters do worse (16.70% at chunk 8, 3.44% at chunk 16). Chunking collapses at this budget.

LIBERO four-suite mean success rate, at matched parameter budgets

All five share a frozen front end, the same demonstrations, action head and evaluation protocol, so the gap is attributable to the controller. The two action-chunking variants carry more parameters than PredVLA and still place last — but these are the authors' own baselines, with no reported hyperparameter search at this scale.

arXiv:2608.26673, table 1 and table 2 — preprint, unreproduced

Percentage points lost when online error regression is switched off



Setting the inference step count to zero gives exact open-loop control on identical weights — the ablation the architecture makes possible. The cost of losing closed-loop correction roughly doubles from the easiest suite to the long-horizon one. These are the paper's matched-seed deltas, not differences of the headline 14-seed means, which are computed on a different seed set.

arXiv:2608.26673 — preprint, unreproduced

Where it’s soft.

- Sim only, one embodiment, one benchmark; 40.57% on LONG is not a long-horizon result.

- Per-suite training means no multi-task claim survives.
- The baselines are the authors' own parameter-matched implementations, not published systems, with no reported hyperparameter search. A 3.4% chunk-16 Transformer is a suspiciously broken baseline — plausibly a tuning artifact of the sub-1M regime rather than evidence against chunking.
- The comparison is against small BC policies, not π0.5 or OpenVLA. The paper does not claim to beat those.
- Compute asymmetry runs *in the authors' favor*, which is the honest direction.

Steal list.

1. The exact open-loop switch: if the observation pathway is separable, you can measure what closed-loop correction is actually doing on identical weights. Closing it cost 6–13 pp here. Most policies cannot answer this at all.
2. Frozen encoder + fixed PCA basis fitted on your own demos — cheap, deterministic, and it makes the learned part small enough to iterate on in minutes.
3. Efference copy as the feedback path: feed back the model's *predicted* action and proprioception, not measurements, keeping forward dynamics generative.
4. Per-channel corruption sweeps. Their sharpest finding is asymmetric — 25% proprioceptive corruption cost 49.7 pp on OBJECT, 50% visual corruption under 5 pp. Worth running on anything you deploy.

Cost to reproduce. 0.68M trainable parameters, frozen front ends, public LIBERO demos: estimate single-digit GPU-hours per suite on one consumer card, no data collection, no pretraining. Test-time error regression is the only unusual cost, and 46 ms/step includes it. INFERENCE \2192 a weekend, with the risk sitting entirely in whether the baselines were fairly tuned.

7 What this changes for a small build

Assume one arm, one GPU, no fleet. Three items move something this week.

PredVLA is the only paper here you could fully reproduce. 0.68M trainable parameters, frozen off-the-shelf front ends, public LIBERO demos, single-digit GPU-hours per suite. If you want a policy you can actually read end to end and instrument, this is the architecture to clone — with the caveat from §6 that the baselines may be undertuned.

CLAP makes world models testable without a cluster. Under 12 GB VRAM on a 3060 is the difference between "read the paper" and "run the thing." If a learned dynamics model was previously out of budget, it isn't this week.

VISTA removes a hardware dependency. If a tactile array was on your build list purely to get a contact signal into the loop, a compliant gripper plus the cameras you already have may be enough. No code released, so this is a design cue rather than something to pull, and it needs a gripper whose deformation is visible and repeatable.

Nothing this week bears on data-collection rigs or teleop cost.

8 Market signal

Where headcount is visibly going: Generalist ($600M raised in total), 1X if the SoftBank deal closes, Locus (absorbing the Nexera manipulation team), and a new NSF-funded centre at UT Austin. Worth noting that this week's manipulation-specific demand came from a *logistics* buyer, not a humanoid one.

Three technical clusters recur across the work that got attention. Streaming and asynchronous inference (StreamPI, FlashVLA, few-step flow sampling) appears three separate times — the field has moved from "does the policy work" to "does it run at control rate." Contact and force as first-class policy inputs rather than safety wrappers (VISTA, TacForcing) is the second. Cross-embodiment latent-action interfaces (CLAP, camera-centric pretraining) is the third.

Table stakes, in the sense that they no longer distinguish anything: LIBERO ≥95%, action chunking, flow or diffusion action heads, a repo linked at submission. Still scarce: controlled parameter-matched ablations (almost nobody runs them, and they are

why PredVLA is legible), real-robot evaluations with stated trial counts, and any evidence the authors know which benchmarks are redundant.

No credible compensation or levelling datapoints surfaced this week. The aggregator pages that rank for “robotics engineer salary 2026” are SEO content with no stated methodology; reporting nothing beats laundering them.

9 Threads

First issue — board initialized.

THREAD	STATUS	WHAT WOULD COUNT AS THE NEXT REAL UPDATE
Does VLA scale buy language-conditioned control?	**ADVANCING** (PredVLA)	A sub-10M policy on *real* hardware, multi-task, vs. a published baseline rather than an in-house one
Contact sensing without tactile hardware	**ADVANCING** (VISTA, TacForcing)	An independent lab reproducing visual-deformation sensing on a different gripper, or released specs
World-action models as unified policy + simulator	**ADVANCING** (Riemann-1.0, CLAP, WALL-SS)	Someone using a released world model for policy improvement, not just video prediction
Claimed open releases vs. actual artifacts	**CONTESTED** (StreamPI pending, Riemann unverified, Muse no card)	Weights appearing, and someone running them
Humanoid capital vs. demonstrated capability	**ADVANCING** (SoftBank/1X, unconfirmed)	The 1X deal confirming or dying; a humanoid company publishing a method, not a video
Manipulation acquired rather than built	**ADVANCING** (Locus/Nexera)	A second logistics incumbent buying an end-effector or policy team this quarter
Benchmark validity in physical AI	**NEW** (redundancy audit)	A major lab adopting a reduced benchmark set, or LIBERO leaving headline tables
Low-cost open hardware as training substrate	**NEW** (Microduck, Poppy LQR)	Microduck shipping and a third party training a policy on it
Inference latency as deployment bottleneck	**ADVANCING** (StreamPI, FlashVLA, MeanFlow)	Latency reported alongside success rate as a default

ALSO SEEN

GaussVLA — Mamba backbone consuming 3D Gaussian tokens. Architecture note, no eval that isolates the Gaussians.

GaussianDream++ — compact Gaussian world-state tokens inside the VLA. Overlaps CLAP without the release.

Zero-WAM — in-context imitation from one human video. Interesting claim, no code, unreproduced.

R3 — RL post-training for free-form language reasoning that steers a low-level policy.

RA-VLA — retrieval for test-time adaptation. Retrieval corpus not characterised.

One Policy, Many Embodiments — camera-centric action geometry for cross-embodiment pretraining. Same territory as CLAP.

LM-X — progress, event and uncertainty heads on a VLA. Interpretability claim without a decision-relevant eval.

Memory Anchors — continual learning without full retraining. Small task suite.

TemporalFlow-VLA — surface temporal flow as execution history. Close to StreamPI, weaker eval.

GRAFT — cached vision-language prefix makes online VLA adaptation cheap. Worth revisiting if code lands.

FLARE — failure detection and recovery wrapped around long-horizon VLA execution.

WALL-SS — next-scale autoregression for longer world-model rollouts.

4DSynth — language or a photo into editable 4D scenes with physics.

Figure: Index — crowdsourced smartphone data pipeline, $1B committed, nothing released.

Jetson Orin Nano 2 — 78 TOPS, 8 GB, 2× inference at 40% less power. Ships H1 2027, so it changes nothing yet.

> **Anthropic Model Hardware Standard** — driver primitives over MCP for agents controlling lab arms. Research preview, nothing released.

10 Open question

Is PredVLA’s 4–7× margin evidence for predictive coding, or evidence that Transformers and action chunking simply don’t function below one million parameters? The paper can’t distinguish these — every baseline is its own implementation in a regime nobody optimizes for. Settled by either a properly tuned sub-1M Transformer from an independent group, or PredVLA holding its margin at 10M–100M parameters where Transformer baselines are known-good. The second is cheap, and the authors didn’t run it.

Sources labeled inline. All arXiv items are preprints, unreviewed. Single-lab claims flagged where relevant.