NOTES FROM THE BOT LAB · EDITION OF NOVEMBER 26, 2026

Why a Robot Arm Needs Two Cages: Live VLA Physics, Sterile Trays, and the FDA Predicate Trap

By Will Tygart · Tygart Media · Canonical edition at https://tygartmedia.com/why-a-robot-arm-needs-two-cages-vla-fda-wind-tunnel/

Executive Takeaway: Wrapping a cloud Vision-Language-Action (VLA) robot policy inside a **240 Hz deterministic physics cage** dropped hospital sterile-tray breaches from **8 of 150 steps (5.33 percent) down to 0 of 150**, and feeding **216** leak-stripped OpenFDA 510(k) filings through our **Wave v3** + 3-model consensus gate caught a **De Novo** regulatory trap that clean deterministic checklists missed.

Live Cloud VLA Physics & Sterile-Field Benchmark (`abstract.inc` + MuJoCo)

Policy / Scenario	p50 Latency / Ready	Physical / Sterile Outcome	Metered Cost / Note
allenai/molmoact2 (nominal / glare / shortcut)	389.3-448.0 ms p50	24.05 cm cup push; 0/180 sterile breaches across all 6 sterile runs (0.0 pct)	0.39 cents (10 chunks) / VLM mistook 2D slide for lift
physical-intelligence/pi0.5 (tabletop & sterile wait)	213.6 ms warm p50; >90s sterile cold wait	Approached cup to 12.77 cm (0.0 cm displacement); timed out on sterile cold start	0.24 cents tabletop / 0.00 sterile after HTTP 503
microsoft/rho (nominal, Shield OFF)	257.6 ms p50 (~4 min cold)	8/150 sterile tray breaches (5.33 pct); 593.37 N peak force; 1.221 m/s velocity	Native gripper units [2.10, 3.10]; breached cyan keep-out
microsoft/rho (nominal & shortcut, Shield ON)	245.6-246.0 ms p50	0/150 sterile breaches (0.0 pct); velocity capped at 0.417 m/s; shortcut peakF 286 -> 0.0 N	120-153 slew clamps + 80 force retracts; offset hit vial
xiaomi-robotics-1 & flux-3-action	59.8s / 82.2s ready; 386.6 / 1,945 ms p50	arx_x5 (dt=0.04s) and franka_droid [32,8] joint targets decoded clean over HTTP/3	0.13 cents (Xiaomi) / 1.16 cents (Flux 91.7s warmup)
perceptron/isaac-0.5 (probe)	220.4s HTTP 503 timeout	Remained in container-placement engine_starting despite public Ready status badge	Zero billed; proved why health badges != ready GPUs

Four Uncomfortable Engineering Failures Caught in Round 1

1. Public Installer `404` & Ready Badge Drift: `install.sh` returned `HTTP 404`, and `perceptron/isaac-0.5` timed out after `220.4s` (`HTTP 503 engine_starting`) while `abstract.inc` displayed `Ready`. Always verify real endpoint readiness via bounded SDK probes.

2. Native Action Semantics Differ Across Checkpoints: `microsoft/rho` emits unnormalized gripper commands in `[2.14, 3.07]`, whereas `pi0.5` and `molmoact2` run normalized `[0, 1]`. Every policy needs a verified embodiment adapter.

3. Single-Frame VLM Judging Lies: `deepseek-v4.1-flash` inspected camera frames of `molmoact2` sliding a cup `24.05 cm` across the tabletop (`dZ = 0.0 cm`) and reported a vertical lift above the table surface. Never grade physical manipulation using raw 2D camera snapshots alone.

4. Zero Breaches != Zero Collisions: Even with `0/150` sterile breaches, `microsoft/rho`'s `own_frame` offset displaced the target vial at `504.79 N` impact. Geometric cages stop boundary cuts; frame calibration still decides grasping accuracy.

OpenFDA 510(k) Wave v3 Wind-Tunnel & Predicate Gate (21 CFR 876.1500)

In software, an uncontained agent hits a bad command. In medical robotics, an uncontained submission hits an FDA Refuse-to-Accept hold or a De Novo technological-divergence order. Using 216 Day-0 leak-stripped CDRH filings (108 matched fast-vs-slow pairs), we benchmarked surface text (Wave v1), deterministic predicate graphs (Wave v2), and fused Open LLM scoring (Wave v3) alongside an NAY surgical-robot candidate audit.

Wind-Tunnel Layer / Gate	Sample & Validation	Logistic / RF / LGBM ROC-AUC	Decision / Diagnostic Signal
Wave v1: Surface Text Only	N=216 GroupKFold-5 (108 pairs)	0.4674 / 0.4971 / 0.5124	Leakage-stripped prose formatting adds ~zero signal over coin flip
Wave v2: Predicate & Safety Graph	N=216 GroupKFold-5 (108 pairs)	0.5801 / 0.6067 / 0.5709	Predicate age, self-predicate cite, applicant & safety blocks drive signal
Wave v3: Fused Graph + DeepSeek Rubric	N=216 GroupKFold-5 (paired bootstrap)	0.6221 / 0.6165 / 0.6144	+0.0420 LR lift (p=0.0495) and +0.0434 LGBM lift; Brier 0.2570 -> 0.2459
3-Model LLM Consensus Panel	N=32 (16 pairs, 3 open LLMs)	0.7480 Ensemble Mean	Beats deepseek (0.7461), kimi (0.6895), glm-5.3 (0.6543); r=0.49-0.53
NAY Unshielded Pre-Sub vs. Shielded	21 CFR 876.1500 Candidate Gate	SERS 40.73 -> 88.41 (#1/6)	Shielded cleared all 4 recall gates yet scored 0.0533 fast-90d (De Novo trap)

Why Deterministic Compliance Alone Does Not Guarantee 510(k) Clearance

Replacing tainted predicate K232610 (6 open recalls) with clean teleoperated predicate K251817 (da Vinci SP), adding a 21 CFR 807.92 matrix, and embedding verified 240 Hz sterile-shield telemetry eliminated all 4 deterministic risk gates and raised SERS from 40.73 (#6/6) to 88.41 (#1/6). Yet all three open LLMs converged at 0.0533 fast-90-day clearance odds (4.20/5 novel-mechanism risk): citing a teleoperated endoscope system for an autonomous neural policy flags a textbook De Novo pathway trigger unless bench/clinical human-oversight equivalence is bounded.

Two-Cage Operator Checklist (Pre-Flight): (1) Verify endpoint readiness and native gripper scaling before closing physics loops. (2) Enforce a high-rate geometric + velocity + force shield (240 Hz) independent of the neural policy. (3) Never grade physical success using single-frame VLMs without 3D simulator kinematics. (4) Screen every cited 510(k) predicate for open recalls (K232610 carried 6). (5) Run multi-model De Novo technological-divergence review (21 CFR Part 860 Subpart D) before filing autonomous hardware under teleoperated predicates.

Engineering & Regulatory Note: This lab brief summarizes simulated MuJoCo physics experiments, metered cloud API calls, and historical OpenFDA CDRH 510(k) summary research as of November 26, 2026. It is engineering systems documentation, not medical-device legal advice or an FDA regulatory determination.