Independent verification for physical AI



STRESS-TEST. BREAK. REPORT.

We stress-test robot policies in simulation and audit generated worlds frame by frame — returning a physics verdict with every failure shown.

Evidence figures from Lab artifacts. Financial figures from plan assumptions.

Physical AI is funded at scale, but the hard tasks remain unsolved.

ROBOTICS VC (2025)

$27.6B

$27.6B across 1,009 deals in 2025 (PitchBook Q4 2025 preview).

RLBENCH SIM SUCCESS

89.4%

EquAct 2D/100 setting (arXiv 2505.21351, 2025).

BEHAVIOR-1K SIM SUCCESS

12.4%

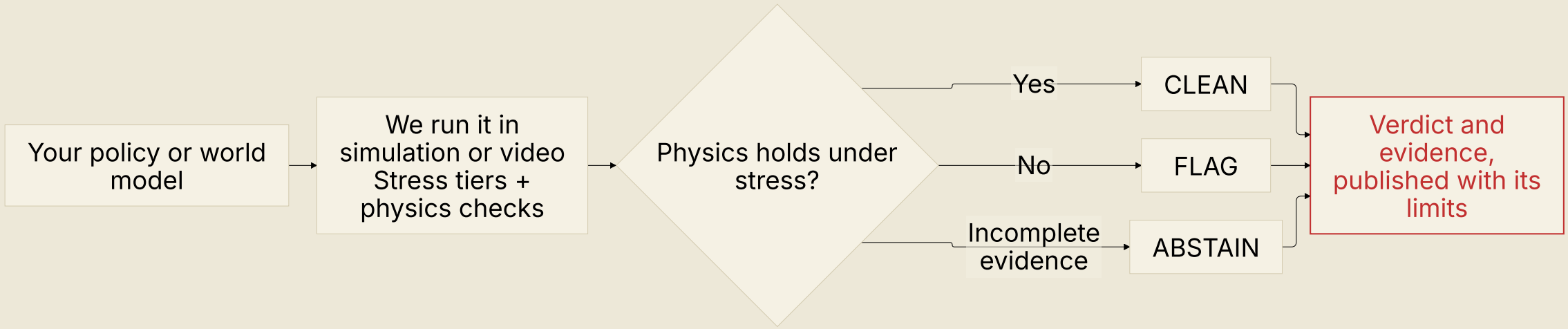
BEHAVIOR Challenge winner, full-task success on BEHAVIOR-1K in simulation (AI Index 2026, Ch. 2).

Note: Sources: PitchBook Q4 2025 Robotics & Physical AI VC Trends; EquAct (arXiv 2505.21351, 2025); AI Index 2026 Ch. 2 for BEHAVIOR-1K.

PROOF Capital is real ($27.6B), difficulty is real (89.4% RLBench 2D/100, 12.4% BEHAVIOR-1K sim success), checking is still done by builders.

TAKEAWAY The gap between funding and verified capability is the opportunity.

Your policy goes in. A verdict with evidence comes out.



PROOF Three-state gate (FLAG/CLEAN/ABSTAIN) with per-clip measured support plane.

TAKEAWAY The checker refuses on thin evidence and does not assume the floor.

The harness applies and measures the perturbation correctly. A learned-policy run needs a partner's checkpoint and has not happened.

TASK	NOMINAL	MILD	MODERATE	SEVERE
Stack	0.900	0.840	0.840	0.820
Door (non-monotonic)	0.800	0.760	0.660	0.760

Note: Signal: Stack and Door. Lift and PickPlaceCan are insensitive controls — in the run tables, not on this slide.

PROOF Scripted OSC_POSE controller, four tasks, four tiers, 50 episodes per cell. Lift and PickPlaceCan hold at 1.000 across every tier. Stack moves 0.900 to 0.820. Door moves 0.800 to 0.760 with the moderate tier at 0.660, which is non-monotonic and is reported as such.

TAKEAWAY What this establishes: the ladder applies, the scoring measures it, and the harness reproduces. What it does not establish: policy robustness. That run needs a learned checkpoint, and it runs on a partner's own. The relative gate passes a weak policy at 0.8x its own weak baseline; an absolute floor is not yet defined — the gate tier is the only threshold.

On real robot footage the gate flags, and shows its working.

DATASET	TRACKS	FLAG RATE (CLEAN)	95% CI	TARGET	VERDICT
towel	67	0.045	[0.0153, 0.1236]	0.020	FAIL
candy	83	0.012	[0.0021, 0.0651]	0.020	FAIL
ziploc	57	0.193	[0.1113, 0.3134]	0.020	FAIL

Note: Run verdict: FAIL under SAP v2's Wilson 95% upper bound ≤ 0.02 rule. Towel FAIL (0.045 [0.0153, 0.1236], upper 0.124), Candy FAIL (0.012 [0.0021, 0.0651], upper 0.065), Ziploc FAIL (0.193 [0.1113, 0.3134], upper 0.313). The contact penetration threshold (0.0080 m) sits below the measured tracking noise floor (0.0160 m); this is the attributed cause of the false positives. Under the frozen rule a PASS at a 0.020 target needs at least 189 clean tracks with zero flags (280 with one); each dataset here has fewer (67, 83, 57), so none could pass.

PROOF towel 3/67 flagged, candy 1/83 flagged, ziploc 11/57 flagged; towel detection 1.000 on 134 injected violations from 67 clips.

TAKEAWAY The gate flags clean footage on all three datasets under the frozen rule.

ARTIFACT PROVENANCE

p2e1-real-footage-gate →

Generative world-model cohort

COGVIDEOX FLAGGED	COSMOS FLAGGED	WAN FLAGGED	REAL-CONTROL SPECIFICITY
0.889	**1.000**	**0.889**	**0.000**
8 of 9 held-out clips flagged	9 of 9 held-out clips flagged	8 of 9 held-out clips flagged	15 real controls, 15 flagged

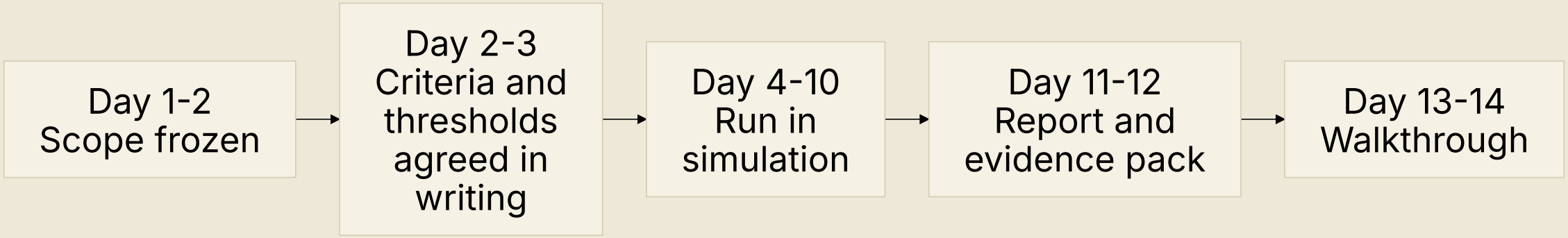
PROOF P2-E2 published: 3 models × 9 clips + 15 real controls. Real-control specificity 0.000 (15/15 flagged). Verdict FAIL. Superseded checker (binary FLAG/CLEAN, no ABSTAIN). Re-scoring this cohort under SAP v2 is outstanding; until then its flag rates carry no information (real-control specificity 0.000, 15/15 flagged).

TAKEAWAY The generative cohort is measured and published. The gate failed on real controls; re-scoring under SAP v2 is outstanding.

ARTIFACT PROVENANCE

p2e2-generative-cohort →

Two weeks. Criterion agreed in advance. One private report.



PROOF Objection: unreleased policy is exposure. Answer: criterion agreed first, report private, method public.

TAKEAWAY You commission one independent check of a claim you framed, before committing hardware.

Three ways to buy: an audit, a pilot, a run. Prices indicative; design-partner validation in progress.

OFFER	SCOPE	PRICE
Design-partner pilot	Two weeks: scope and criterion frozen, the run, one private report	Set on the published ladder and agreed on a call
Fixed-scope audit	Foundation: 6 clips, 5 days, short report • Standard: 6 clips, 5 days, full report	Foundation $1,900 • Standard $3,500
Per engagement	Policy stress or world-model physics run, criterion frozen before run	Fixed fee, quoted per engagement

PROOF Gross margin 78–84% (one-off), 70% (continuous) — plan assumption; variable costs not published.

TAKEAWAY Margin is a plan assumption, not a measured figure.

A thin verification layer under a large physical-AI build-out.

TAM

$39M

Global robotics + world-model QA budget holders: 1,080 orgs (PitchBook robotics cap table + AI Index 2026 lab survey) × $36K ASP (blended audit+pilot+continuous)

SAM

$7.7M

Reachable ICP (sim-first, NA/EU): 430 orgs (filtered by HQ, public cloud spend >$1M, robotics/AI team >5) × $18K ASP (single engagement)

SOM (YR 3)

$0.8M

5-person capacity at 6-7 engagements/yr × $25K avg (design-partner pilot → standard scope)

PROOF TAM: 1080 companies × $36,000 = $39,000,000. SAM: 430 reachable ICP × $18,000 = $7,740,000. SOM: 25–35 design partners × $25,000 = $600,000–900,000. Steady burn $78,000/mo = $936,000/yr; SOM ($600,000–900,000) does not cover burn in year 3.

TAKEAWAY Break-even on SOM in year 3 is not supported; TAM/SAM represent upside if independent verification scales.

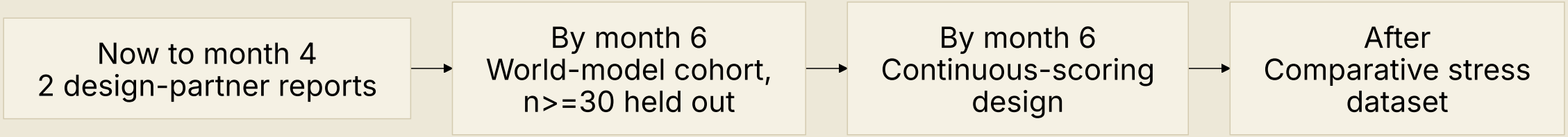
Robocurve publishes benchmarks. Titus consults. Prismatical is a personal site. Haga referees.

PLAYER	WHAT THEY SELL	WHAT HAGA DOES THAT THEY DON'T
Robocurve (robocurve.org, $10M seed, Initialized Capital)	Open-source benchmarks and public result pages: CIs on charts, no verdicts, no price.	Freezes the criterion before the run, refuses on thin evidence, keeps the run private.
Titus (titus.io)	Evaluation as a service for industrial operators; pre-release site, no public product.	Ships a product with published artifacts — failures included.
Prismatical (prismatical.com)	Solo thesis: sim policy evaluation plus synthetic data; personal site.	Pre-revenue, self-funded, published runs with failures.

PROOF Robocurve $10M seed verified (Initialized Capital, Dealroom). Titus and Prismatical assessed from public sites. Haga: pre-revenue, self-funded, published runs.

TAKEAWAY Criterion-first, private, and willing to refuse to certify.

2 design-partner reports by month 4. World-model depth by month 6.



PROOF Plan spends $997,000 of $1,000,000 raised over 14 months. Remaining $3,000 unspent.

TAKEAWAY Round buys the milestones above and stops. Last months need larger close or smaller scope.

One founder, self-funded, every published run carries its failures.

APPLIED ML/AI ENG

7 years

Fraud detection, multi-agent pipelines, retrieval, evaluation harness to production. Cut inference latency 2.1s→180ms (91.4%); fraud FPR 14%→3.5% at 1.2M tx/day; built 1,200+ case eval harness. BS SE (Hons.), GIFT University 2019.

BUILT

4 tasks · 3 world models

P2-E2 real-control specificity 0.000 — gate fails on real footage; generative rates reported, not certified.

FUNDED BY

Self-funded

First commit 2026-07-13 (haga-core). No outside capital.

- IP assignment signed 2026-07-17. Equity & IP with founder (no legal entity formed; assignment to company on incorporation).
- Seeking hardware/robotics co-founder: equity only (revisable after funding/incorporation).
- First hire month 4: ML/robotics-eval engineer.

Note: Planned in North America (Delaware C-Corp default, Canada alternative). Depends on lead investor requirements. No final decision made yet. Incorporation triggered on first committed investment (SAFE execution).

PROOF Team budget 14 months: $503,500. Headcount: 3 (founder + 2 contractors) → 4 (month 4 hire) → 5 (month 9 co-founder).

TAKEAWAY Money buys named people; headcount rises on published dates.

Raising $1.0M — burn, runway, and where it goes.

AVG BURN (MO 1–12)
$70,083
Mo 1 $53,500 · steady $78,000 (mo 10)

RUNWAY (MONTHS)
14
Recomputed from burn schedule.

TEAM FUNDED
3 → 5 people
Founder + 2 contractors · hire mo 4 · co-founder mo 9

DESIGN-PARTNER PILOT
Priced (indicative)
Set on published ladder, agreed on call.

WHERE THE MONEY GOES

Bucket	Share	Amount
Team (Founder + Hire + Contractors)	50.4%	$503,500
Compute (GPU)	19.3%	$193,000
Product	9.4%	$93,800
GTM	12.1%	$121,300
Legal + Buffer	8.5%	$85,400

WHAT THE MONEY BUYS

#	Milestone	When
1	2 live design-partner evaluation reports	≤4 mo
2	World-model scoring at depth (n≥30 held out)	≤6 mo
3	Continuous-scoring product design	≤6 mo

Note: Full 18-mo plan costs $1,309,000; last 4 months need larger close or smaller scope.

PROOF Plan spends $997,000 of $1,000,000 raised over 14 months. Remaining $3,000 unspent.

TAKEAWAY The allocation above is what the plan commits over the funded window; the rest is headroom.

To earn the right to raise a Series A.

INSTRUMENT

Post-money SAFE (YC standard)

No discount, no MFN

VALUATION CAP

$8,000,000

Standard YC side letter only; offered to lead investor or single check >= $250K

PRIMARY CONTACT

Mushood Hanif · Founder

mushood@hagalabs.com hagalabs.com linkedin.com/in/mushood-hanif