BETTER FIXED-STATE FIDELITY NEED NOT IMPROVE CLOSED-LOOP PERFORMANCE IN QUANTIZED NEURAL POLICIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Does closer action agreement with a full-precision policy imply better closed-loop performance after quantization? We hold trained source policies fixed, vary calibration data or compiler settings, and execute compiled policies on Edge TPU and Hailo accelerators in simulated control tasks with paired initial conditions. In a complete, prespecified eight-rule Hailo FetchReach family, one artifact has a 16% smaller fixed-state mean action-error norm than another yet incurs a 34-fold larger estimated mean return loss relative to FP32. An early–middle contrast selected from that family and locked before replication favors early-state calibration in seven of eight independently trained FetchReach policies, and outcome-informed audits retain large calibration effects under alternative compiler settings. Within the original family, a full-budget, AdaRound-enabled compiler intervention reduces action root-mean-square error (RMSE) on calibration-disjoint states for all eight rules, but mean return decreases for five; retrospective family-adjusted intervals place all five changes below zero but within the prespecified 0.5-return practical band. In a separate prospective benchmark of eight fresh FetchReach policies and 128 Hailo artifacts, a secondary analysis finds lower RMSE on a calibration-disjoint panel in 53 of 64 policy–rule comparisons, 22 of them with lower observed mean return; these are descriptive contrasts clustered within policies. Post-outcome episode resampling retains 16 of the 53 RMSE improvements with adjusted RMSE-change intervals below zero; the count varies across bootstrap seeds. Changing action coordinates or evaluation states changes which comparisons reverse, and reversals also occur when both executables are scored on the same visited states. We provide an execution-bound qualification protocol and provenance-recording harness that compare paired task outcomes against task-specific tolerances. The results distinguish fidelity-based candidate selection from task-tolerance approval and support evaluating the measured executable on its intended execution path rather than fixed-state fidelity alone.

1 INTRODUCTION

Resource constraints on embedded devices motivate policy distillation or low-bit execution (Rusu et al., 2016; Kresse & Lampert, 2026). Post-training quantization (PTQ) adapts a trained policy for low-bit execution. Calibration-based toolchains use selected policy inputs, called *calibration data*, to set quantization parameters and may optimize weight rounding (Nagel et al., 2020). "INT8" specifies the numeric format, but not calibration data, compiler settings, or the executable.

That omission matters under feedback: action differences can propagate across an episode. In our complete, pre-specified eight-rule Hailo FetchReach family, calibration states vary while the source policy, task, paired starts, compiler path, runtime, and device remain fixed; terminal success ranges from 96.6% to 21.4%. Across these artifacts, fixed-state error magnitude does not consistently order task performance. For the same calibration rules, a change in Hailo compiler settings improves fixed-state action fidelity without consistently improving mean task return; the observed return decreases remain within the practical tolerance (Figure 1).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

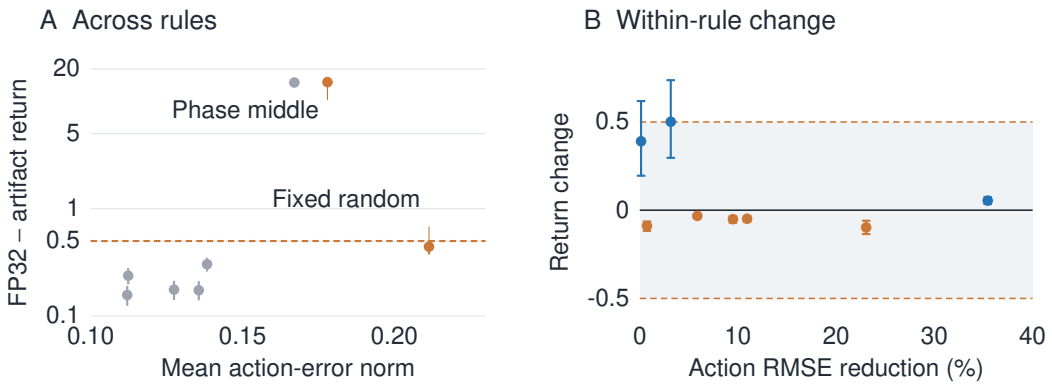

Figure 1: **Better fixed-state fidelity need not imply better task performance. A:** Complete Hailo family with fixed source policy and execution path; logarithmic return-loss axis and marginal paired 95% CIs. Highlighted rules reverse the scalar ordering; fixed-random is not established as tolerance-compatible. **B:** Full-budget Hailo DFC level 3 (AdaRound enabled) minus level 0; RMSE improves for 8/8 rules, but 5 mean returns decrease. Retrospective Bonferroni family 95% CIs bound all decreases within the ± 0.5 practical band. A's dashed line marks task tolerance; B's shading marks practical change, not qualification against FP32. The panels use distinct fidelity measurements and task contrasts: A does not isolate the effect of error magnitude, nor B the effect of AdaRound (Appendices B.1.1 and B.2; construction in Appendix D.4).

Working terms. A *compiled artifact* is a device-specific executable produced from a source policy. Its *execution path* comprises the compiler configuration, runtime, hardware, and runner used to build and execute it. Together they form the qualification unit. Qualification evaluates task degradation relative to the FP32 reference against an application-specified tolerance under specified task conditions; the assessment may remain indeterminate. *Return* is cumulative episode reward; *terminal success* indicates whether the goal condition holds at the final step. Section 3 specifies the protocol.

Ranking, selection and approval are different questions. A counterexample to fidelity–return ordering need not make minimum-error selection a poor strategy. Our selection audit does not establish that RMSE selection passes over a confirmed-compatible alternative for a degraded artifact, and some contrasts favour RMSE selection (Section 4.5). Our results challenge fidelity-only qualification, not the use of fidelity for candidate screening or selection. What remains unvalidated is using fidelity alone to approve the declared task tolerance on an execution path.

From checkpoint to deployed policy. Calibration and compiler changes can alter trajectories and returns through feedback even when the source policy is fixed. Checkpoint performance alone therefore does not establish the deployed policy's task performance.

We therefore recommend paired task tests of the exact executable on its intended execution path against the FP32 reference. Beyond action mean-squared error (MSE), we test whether a deliberately strong dynamics-aware local proxy combining measured action errors with a fixed local linearization can replace task evidence. This proxy serves as a falsification control rather than a proposed method and does not establish a transferable qualification rule.

The paper makes two contributions:

1. controlled evidence of a fidelity–task mismatch. Fixed-state error magnitude can fail to order task performance, and fidelity improvements need not improve task return. Complete-family audits and policy replications establish repeatable construction sensitivity within FetchReach. Across complete tested calibration families and compiler interventions, we separate the direction of fidelity–task disagreement from its practical magnitude;
2. an execution-bound qualification protocol in which construction checks precede paired task evaluation against an application-specific tolerance. A released provenance record links

each tested executable to its execution path as reproducibility support for the task-audit protocol. This contribution is an evaluation procedure, not an established improvement in statistical decision accuracy over the tested baselines.

We do not estimate how often material artifact differences occur in deployed populations or their typical magnitude. The compiled policies run on real devices in simulated environments; physical sim-to-real transfer remains untested (Section 5).

2 Related Work

Post-training quantization and calibration. Quantization-aware training accounts for quantization during learning to support integer-only inference (Jacob et al., 2018). Post-training methods instead quantize an already trained model, using techniques such as clipping, learned rounding, and block reconstruction to preserve predictive accuracy (Banner et al., 2019; Nagel et al., 2020; Li et al., 2021). Other work examines how the amount and distribution of calibration data affect quantization performance (Hubara et al., 2021; Yu et al., 2021; Shang et al., 2024).

Quantized policies and embodied control. Quantization has also been studied for policies that act through repeated interaction with an environment. QuaRL uses quantization to improve RL efficiency (Krishnan et al., 2022), while quantization-aware training enables continuous-control policies to run on integer FPGA hardware (Kresse & Lampert, 2026). For vision–language–action models, DA-PTQ addresses errors that accumulate along trajectories (Xu et al., 2026), and QuantVLA calibrates attention and output scales (Zhang et al., 2026). QuantWAMs uses closed-loop rollouts on reachable states to refine quantization choices (Zhou et al., 2026). These studies already address policy performance during interaction, going beyond prediction error on fixed inputs. We instead examine executables produced by vendor compilers while holding the trained source policy fixed and varying calibration or compiler settings. There, calibration choices alone, with policy, runtime and device fixed, move task return beyond its tolerance, and fixed-panel action agreement does not order those changes.

ML compilation and deployment. TVM compiles neural networks into code optimized for different hardware targets, while Ansor searches for efficient implementations of tensor computations (Chen et al., 2018; Zheng et al., 2020). Our study concerns the task behavior of the resulting executable: in the vendor quantization pipelines we test, calibration data and compiler settings can change the policy's state-to-action mapping. We therefore record these build inputs alongside the executable and evaluate it on the intended runtime and device.

Quantized feedback. Classical quantized-control research develops stability and performance conditions for specified dynamics, quantizers and communication constraints (Delchamps, 1990; Wong & Brockett, 1999; Brockett & Liberzon, 2000; Elia & Mitter, 2001; Fu & Xie, 2005; Liberzon & Nešić, 2007; Nair et al., 2007; Nešić & Liberzon, 2009; Jiang & Liu, 2013). Our audits instead measure task degradation of a compiled policy relative to its full-precision reference. Errors arise from the compiled executable rather than an assumed additive-noise process (Widrow & Kollar, 2008) or deliberately injected bit flips (Rakin et al., 2019).

Quantization sensitivity and directional alignment. HAWQ and HAWQ-V2 use the curvature of the neural-network loss to assign higher numerical precision to more sensitive layers (Dong et al., 2019; 2020). Our auxiliary analysis instead concerns how action errors propagate through a feedback system. The comparison pairs a sensitivity summary averaged over error directions with a local score that accounts for measured-error directions, their timing, and how their effects combine over time. We ask whether this additional information supports task qualification.

Verification and monitoring. Reachability methods bound the states a modeled system can visit to check safety properties (Ivanov et al., 2019; Xiang & Johnson, 2018; Tran et al., 2020; Wang et al., 2024). Runtime monitors assess safety-related conditions during execution (Xiang, 2022); observed violations can also guide repairs to a policy and its certificate (Yu et al., 2025). Studies of open/closed-loop gaps and causal confusion highlight a related evaluation problem: performance on fixed data can differ from performance during interaction (Lechner et al., 2022; de Haan et al.,

2019). Appendix G.1 relates the auxiliary diagnostic to perturbation geometry, policy regularization, and calibrated monitoring. Here, qualification rests on paired task outcomes under specified task conditions and tolerances; it certifies nothing about physical safety.

3 ARTIFACT-LEVEL QUALIFICATION PROTOCOL

3.1 QUALIFICATION UNIT AND PROVENANCE

What is being qualified. Qualification evidence is specific to the tested executable, runtime, and accelerator; it does not automatically transfer with the executable to another runtime or accelerator. The record links source policy, export, calibration data, compiler configuration, and runner by digests and settings, including requested/effective optimization levels, build-host accelerator availability, runtime version, and accelerator identity. The separate runtime host rechecks the executable's digest before accepting measurements.

Why retain the executable. A digest identifies the measured file; it says nothing about reproducing those bytes or that behavior. Matched level-0 rebuilds replayed 12,288 historical episode rows exactly, but a locked 200-build audit found two or three SDK-emulator action fields in four of 20 cells. One rebuild difference appeared on both emulator and accelerator. The two GPU-visible rebuilt early–middle return contrasts differ by 0.00461 [0.00251, 0.00685],[1] about 0.03% of their 14.67 effect. We retain the executable and a shared-state action fingerprint; panel agreement does not establish global equivalence (Appendices I.2 and D.3.2).

At a shared state s, the action error is $e_b(s) = \pi_b(s) - \pi_{\text{fp}}(s)$: the difference between the action produced by compiled artifact b and that of the FP32 reference for the same input. Evaluating this difference across states gives an error field, which the diagnostics below summarize by magnitude or analyze by direction. These fixed-state comparisons do not themselves determine task qualification.

Executable reference harness. The released reference implementation separates executable checks from task evaluation. In Stage 0, toolchain-specific adapters provide graph and precision checks and action outputs on a fixed state panel. The harness records build and execution identities and checks that outputs are finite and repeatable within a construction-specific tolerance. Device-qualified records require a target-runtime action fingerprint linked to preprocessing and action transformations; emulator fingerprints are recorded separately, and the arrays alone do not verify their execution origin. Only a Stage-0 pass permits Stage 1 to analyze task outcomes paired by identical starts, compute a bootstrap interval, and return a machine-readable decision against the application tolerance. Passing Stage 0 establishes eligibility for the task audit, not acceptable task performance. For example, one Hailo checkpoint yields sharply different artifact task outcomes (Figure 1). The record identifies the tested executable and runtime; rebuilding or changing that path does not inherit its task-loss decision.

Primary protocol and proxy stress test. Artifact status is determined by paired full-episode task outcomes. The local proxy is a diagnostic, not a second qualification protocol; it does not alter the task-based qualification decisions.

3.2 PAIRED TASK AUDIT AND DECISION RULE

Paired comparison and statistical unit. For a fixed source policy, each reset seed defines a pair of episodes: one using FP32 and one using the compiled artifact, with the same initial conditions, evaluation horizon, and outcome measure. Bootstrap intervals resample these episode pairs together. In replications across independently trained policies, the policy is the top-level statistical unit. Fresh evaluation seed blocks are fixed before outcomes are observed; robustness reruns identify the blocks they reuse. Pairing reduces differences in starting difficulty, but it does not establish absolute safety. The local-window analyses use separate statistical units, specified in Appendix G.1.

Task-audit decision rule. Qualification compares the uncertainty interval for task degradation with a tolerance fixed before the audit; the tolerance is task-specific, not a universal deployment

[1] Unless stated otherwise, bracketed intervals denote 95% confidence intervals (CIs).

standard. For Ant, 250 return units was fixed using a separate pilot-informed design, just above its conservative Holm-adjusted minimum detectable effect of 227–231 units. This design-based threshold equals 4.5% of the FP32 mean return and has no direct interpretation as a single physical quantity. Detectability is not an application loss budget: the Ant classification is conditional on this design threshold, which no external deployment requirement validates. For the 50-step Fetch audits, a tolerance of 0.5 return units corresponds to about 1 cm greater goal distance averaged over the episode. These classifications do not rank relative losses: Ant's 667-unit loss is 12.0% of its FP32 return and is materially degraded, whereas the FetchReach EdgeTPU loss of 0.435 is about 62% of its block-specific FP32 return magnitude ($|-0.701|$) but remains indeterminate.

For task T and artifact b, mean paired degradation $D_{b,T}$ is FP32 minus artifact for higher-is-better outcomes, reversed otherwise; positive values indicate worse performance. Before outcomes, the task distribution, horizon, primary outcome, tolerance δ_T, family of comparisons, confidence procedure, and resampling rule are fixed. For interval $[L_{b,T}, U_{b,T}]$, an artifact is *tolerance-compatible* when $U_{b,T} < \delta_T$, *materially degraded* when $L_{b,T} > \delta_T$, and *indeterminate* otherwise. Touching or spanning the tolerance is indeterminate. Within-family performance ranges use corresponding labels: *compatible with limited heterogeneity*, *materially heterogeneous*, and *indeterminate heterogeneity*.

The original backend audits used two Holm-adjusted tests (Holm, 1979) and reported marginal intervals. Appendix C.6 adds a post-outcome studentized-max analysis for the comparison family. It may weaken, but cannot strengthen, the practical recommendation; the registered result remains unchanged. These three analysis roles are recorded separately. Statuses apply only to the specified task, reset distribution, horizon, and outcome, distinguishing statistical differences from practically important losses without certifying numerical equivalence or physical safety. Physical sensor, actuator, energy, and thermal behavior are not evaluated.

3.3 Dynamics-aware proxy challenge

Closed-loop amplification energy (CLAE) is the weighted squared state deviation predicted from measured post-clipping errors by a 20-step local linear model; effects can reinforce or cancel. Comparators are MSE and a trace score retaining error size and average gain, but not direction or cross-time interactions. Nonlinear rollouts separately redirect errors toward lower-gain directions; norms match before re-clipping, which can change applied norms (Appendix G.1.2). CLAE is not a calibrated failure probability or qualification rule. Definitions and checks are in Appendices G.1 and G.5.7. Table 1 summarizes evidence roles and scope.

Table 1: Evidence overview. Design status records when outcomes could inform the design. Analysis units identify comparisons or resampling; they do not assert independence. Rows have distinct designs and are not one multiple-testing family.

Evidence	Design status	Analysis unit	Finding and scope
Ant, FetchPush, FetchReach	Pre-outcome; Ant pilot-informed	Reset seed	Artifact-specific task effects
16-artifact family	Pre-specified complete family	Reset seed	Within-path calibration heterogeneity
Eight-policy + held-out mixture	Prospective replication and held-out follow-up	Policy	Within-task replication and response shape
Row-order task audit	Locked contrast + breadth tests	Reset seed	Task consequence; breadth misses
Timing of action replacement	Development-informed; locally locked	Policy	Equal replacement counts; no norm-matched causal claim
Decision replay + new policies	Retrospective / pre-outcome	Reset seed / policy	Decision trade-offs; no interval superiority or error control
Level/budget + build controls	Outcome-informed reruns; locked rebuild/OpenVINO	Seed / build / conversion	Path robustness; OpenVINO has no device/task arm
Dynamics-aware intervention	Locked mechanism stress test	Episode	Local directional sensitivity
Score selection + Ant cohort	Pre-outcome / held-out	Episode / policy	Selector criteria not met; Ant materiality gates not met

4 Task-level artifact evidence

This section asks how artifact construction and fixed-state fidelity relate to closed-loop task performance.

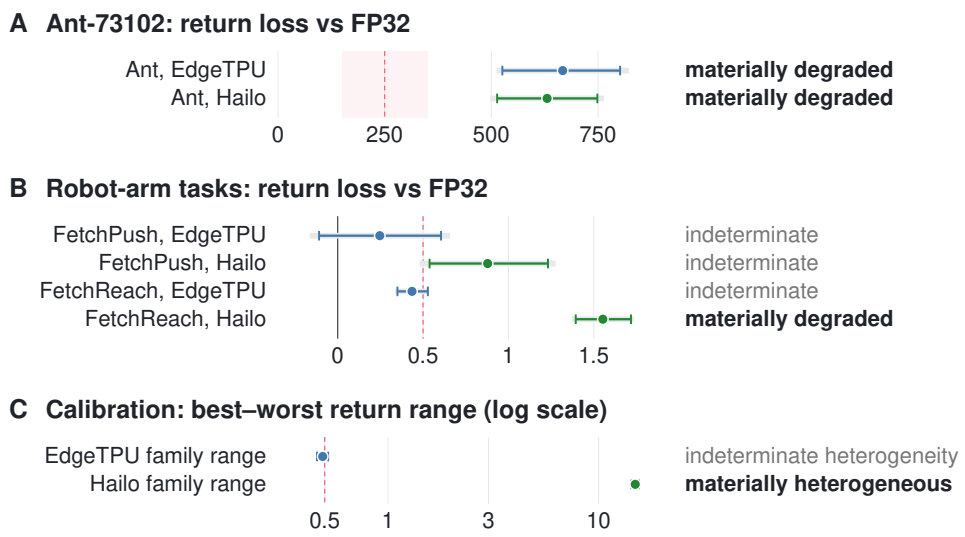

Figure 2: The same nominal precision does not imply the same task performance. **A–B:** individual INT8 artifacts versus FP32. **C:** calibration alone varies within each policy–backend pair: eight pre-specified artifacts, 1,024 shared seeds, and a separately judged best–worst return range, not FP32 loss. Hailo artifacts use the automatically selected level-0 path. Dots: estimates; coloured bars: locked marginal 95% intervals; dashed lines: precommitted tolerances (A: 250; B–C: 0.5). Grey bands and statuses: post-outcome, 20,000-replicate studentized-max audit (Appendix C.6), which cannot strengthen a status. The pink band in A spans the 150–350 sensitivity grid.

Hailo build-path scope. Unless stated otherwise, the original Hailo calibration family, cross-policy and mixture tests, FetchPush audit, and row-order audit used Hailo Dataflow Compiler (DFC) 3.33.1 on a CPU-only build host, where the compiler automatically selected optimization level 0. Each level selects a different optimization procedure; the levels do not accumulate. After observing the original results, matched reruns tested the complete family and eight-policy contrast at explicitly requested levels 0–3. A separate GPU-visible level-3 task audit tested one selected contrast; FetchPush and the mixture tests remain limited to level 0. These checks assess whether previously observed effects persist across build settings, not whether they replicate in new tasks or policies. Compiler-level, calibration-budget, and GPU task-audit details appear in Appendix D.

Finding. In these audits, both Ant artifacts lose about 12% return; Hailo FetchPush is indeterminate against the task tolerance after family adjustment; FetchReach return loss is larger on Hailo than EdgeTPU. No universal backend ordering follows. Separately, Hailo FetchReach's within-path calibration spread motivates the policy, mixture, and compiler controls below.

4.1 Cross-backend task audits

Paired artifact–FP32 audits report return (locked marginal 95% CIs) and robot-arm terminal success: return summarizes the whole path; success records goal attainment at the final step.

On 320 paired Ant episodes and 512 per robot-arm task, Ant return losses are 667 [526, 802] for EdgeTPU and 631 [514, 749] for Hailo—mainly through reduced forward progress. Both Ant lower bounds after family adjustment exceed 350, so the classifications remain unchanged over the illustrative tolerance range of 150–350 (Figure 2A). On FetchPush, Hailo loses 0.878 [0.538, 1.232] return and terminal success falls from 73.2% to 59.0%. Its family-adjusted return interval crosses the task tolerance, leaving the status indeterminate; EdgeTPU's return interval includes zero. On FetchReach, return losses are 0.435 [0.350, 0.528] for EdgeTPU and 1.553 [1.395, 1.718] for Hailo, with terminal success of 88.5% and 65.4%, respectively. Figure 2 also shows post-outcome family-adjusted bounds and statuses. Appendices C.3–C.5 give the full analyses.

4.2 COMPLETE CALIBRATION-ARTIFACT FAMILY

Figure 2C audits all eight pre-existing artifacts per backend on 1,024 fresh paired seeds; none was selected from task outcomes. Phase slices model truncated logs, replay-buffer slices, or fixed-duration capture windows; their prevalence is not estimated. Both backend families show differences across their eight artifacts (paired Friedman tests, Holm-adjusted $p < 0.001$). The FP32 mean return is -0.663 on this block and -0.701 on the earlier 512-seed block; the baselines are block-specific. EdgeTPU's best-to-worst mean-return range is 0.489 [0.460, 0.519]; its CI straddles the fixed 0.5 threshold, so practical heterogeneity is indeterminate. Hailo's range is 14.926 [14.318, 15.549], and terminal success is 96.6% for its best-return artifact and 21.4% for its worst. Recomputation and family adjustment preserve both statuses (Appendix C.6). Calibration can materially change task outcomes for a fixed policy and backend (Appendix B.1.1).

Outcome-informed complete-family CPU reruns retain large calibration spreads under alternative compiler settings and full calibration budgets. The GPU-visible task bridge is restricted to selected pairs, not the complete family; it also retains material contrasts. These checks address dependence on the original build configuration, not independent policy replication (Appendix D.1).

Fidelity improvement does not imply task improvement. Across the eight Hailo artifacts, error size correlates with degradation (Spearman $\rho = 0.833$), but does not consistently order task performance. Fixed-random has the largest mean error norm (0.2114) yet loses only 0.443 return, whereas phase-middle has a 16% lower norm (0.1779) and loses 15.083 return—a 34-fold larger point-estimate task effect.

In a separate within-family comparison at full budget, switching from level 0 to AdaRound-enabled level 3 lowers action RMSE on 800 calibration-disjoint states in all eight rules but improves mean paired task return in only three (Figure 1B). Phase-middle and phase-late show partial recovery averaging 0.446 [0.299, 0.593]; the other six lose 0.044 [0.033, 0.055] on average, although one improves. All five decreases have paired CIs excluding zero and eight-rule Bonferroni family intervals entirely within $(-0.5, 0)$; the largest upper bound on deterioration is 0.135 return. These post-outcome bounds concern L3 versus L0, not qualification against FP32. This is fidelity–task non-monotonicity; it shows neither material compiler harm nor an isolated causal effect of AdaRound (Appendix B.2).

The five artifact pairs selected before task outcomes for similar fixed-state scalar errors have absolute mean-return differences of 0.081–0.169. All paired CIs exclude zero but lie within the $[-0.5, +0.5]$ practical band, showing detectable but not material task differences (Appendix B.1.1).

4.3 POLICY REPLICATION AND FIDELITY SENSITIVITIES

The calibration-contrast replication cohort comprises eight independently trained policies testing a Hailo early–middle contrast fixed before evaluation. Seven of eight return differences are positive; the policy-level mean is 8.595 [4.422, 12.511] (policy-bootstrap CI). Exact sign-flip $p = 0.015625$; sign-only sensitivity $p = 0.0703$ (Figure 3, Appendix C.1). This supports recurrence within FetchReach, not across arbitrary tasks; policy-population prevalence remains unestimated. Post-outcome reruns at explicitly requested level 1 retain all eight positive contrasts, without adding independent policy replication (Appendix D.2).

Calibration mixtures reveal a step response. At fixed row budget, the two policies with large prior early–middle contrasts improve sharply as early-state fraction λ rises from 0 to 0.25, then plateau; two with small contrasts vary within the 0.5 band. The original monotonicity criterion is not met. A separately locked follow-up tests the other four policies (Figure 3B). All four steps—mean return over $\lambda \geq 0.25$ minus return at $\lambda = 0$—exceed 0.5. This is a fresh mixture-response test, not prediction on task-unseen policies (Appendix C.2).

A locked FetchPush behavior-cloning policy shows a Hailo early–middle contrast of 1.708 [1.259, 2.151]; EdgeTPU's CI includes zero. This does not establish cross-task generality (Appendix C.3).

Fresh-policy fidelity–return comparison. A separate prospective task-decision cohort was locked with 8 new FetchReach policies and 128 Hailo artifacts (8 calibration rules, levels 0/3).

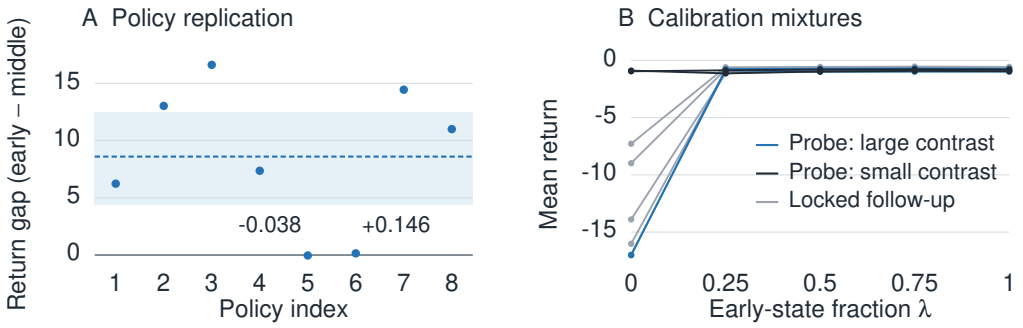

Figure 3: Calibration response on Hailo at level 0. **A:** Early–middle return contrasts, 256 starts/policy. Dashed line: policy mean (8.595); shading: its 95% CI. Indices 1–8 map to seeds 9510–9517. **B:** Mixture means, 256 fresh starts/policy; $\lambda = 0$: middle-only, $\lambda = 1$: early-only. Probe groups reflect prior early–middle contrasts; follow-up was locked separately on four known policies. Lines connect sampled mixtures; the response between them was not measured (Appendix C.2).

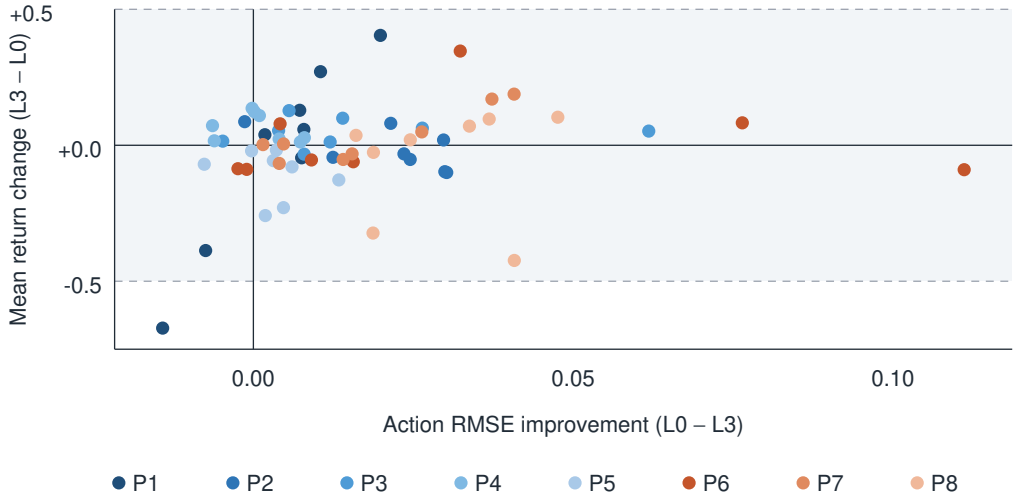

Figure 4: Fresh-policy fidelity–return changes: all 64 comparisons, not only the improved ones. Horizontal: fixed-panel RMSE improvement, the reverse sign of the ΔRMSE column in Table 14; vertical: mean return change. The lower-right quadrant holds the reversals the text counts. Colour marks the eight policies P1–P8, within which comparisons are dependent. These are descriptive point estimates from the sealed record, with no rule or threshold reselected for the plot. Shading is the ± 0.5 practical scale of the compiler contrast, not a qualification against FP32; Appendix B.3 gives the uncertainty and fidelity-definition sensitivities.

Secondary analysis finds lower action RMSE on calibration-unused states at level 3 versus 0 in 53/64 policy–rule comparisons; 22 of these have lower mean holdout return. These are point-sign reversals clustered within 8 policies; Figure 4 plots every comparison (Appendix B.3). For 15/22 reversals, existing within-policy return intervals exclude zero. Post-outcome episode-level resampling of that panel retains 16 of the 53 improvements, a count that moves with the resampling seed.

Restricting the action coordinates, or measuring on the states each artifact visits, moves which comparisons reverse: 23 of 64 are flagged by at least one of the four fidelity definitions and 10 by all four. Running both sealed executables on both distributions, the level-3 executable is more faithful

on the common level-0 states in 48 of 64 comparisons and 17 of those lose mean return. These sensitivities reuse the same policy cohort (Appendix B.3). Their reversal counts are not successive validation stages of one reversal set. The fixed fidelity panel and task holdout are distinct episode blocks; the combined results are not a joint simultaneous inference.

Under the original fixed-panel metric, no RMSE-improved comparison among the 53 has a within-policy return interval lying entirely below -0.5 (Tables 12 and 14).

4.4 Construction controls and limits of explanation

Calibration identity must describe the inputs that actually affect the build, which can differ from the array supplied to the compiler. On the tested CPU level-0 path, full-array builds agree with their same-order effective-prefix controls on measured actions and task outcomes. Reordering the supplied states can change the resulting controller, but the prospective task-effect breadth criterion fails. These controls narrow the construction evidence; they do not establish a general calibration preference or isolate the cause of the fidelity–return mismatch (Appendix D.2, Figure 5).

What the mechanism tests establish. Action-replacement and matched-magnitude interventions do not isolate timing or error magnitude as the cause of the large calibration effects (Appendix F). The tested dynamics-aware local proxy also fails to establish a transferable qualification rule: prospective alert prioritization, calibration selection and unseen-policy trust gating miss their registered criteria (Appendix G.4). These negative results leave paired task evidence as the basis of the reported decisions.

4.5 Task decisions under a fixed audit budget

The prospective task-decision cohort uses family-adjusted paired CI (CI rule). For mean FP32-minus-artifact loss $\widehat{D}$, fixed margin approves if $\widehat{D} < \delta - \tau$, rejects if $\widehat{D} > \delta + \tau$, and otherwise defers; primary $\tau = 0.3$ was fixed before outcomes. At the registered budget of 32 shared reset seeds per policy, no tested rule has an observed conflict with a resolved holdout status. Unresolved holdout cases remain, and CI superiority is not established (Appendix E.1). Post-outcome reuse of the original first 8 audit seeds/policy reduces conflicts versus cutoff for both CI and fixed margin; reductions are not unique to CI (Appendix E.3). In this retrospective cohort analysis, RMSE selection skips no compatible alternative for a degraded artifact; within-level differences from audit selection remain unresolved. Selection is not qualification (Appendix E.4).

5 Discussion and Limitations

Repeatability and mechanism. The audits show repeatable FetchReach calibration sensitivity; they establish neither prevalence, nor a generally best calibration rule, nor cross-task generality. At tested norms, timing contrasts are small; differing scales, visited states and intervention backgrounds preclude attributing earlier large correction effects to timing or magnitude alone.

Limits of transfer. The tested coverage and local-error scores establish no portable artifact selector. In each Ant cohort, all six return ranges exceed permutation-null medians, but the median materiality gate fails. Neither a shared calibration preference nor equivalence follows (Appendices C.8 and C.9). Policies run on real accelerators in simulated environments; physical sensor/actuator dynamics, energy/thermal behavior and robot safety remain untested.

6 Conclusion

Fixed-state error magnitude need not rank artifacts by task return, even where returns differ materially. Fixed-state fidelity gains need not improve task return. Family CIs place the original compiler intervention's decreases within tolerance; fresh-policy reversals remain descriptive. In the tested FetchReach cohort, changing action coordinates or evaluation states changes which comparisons reverse, but reversals persist. Qualification separates candidate selection from task-tolerance decisions bound to the measured executable, execution path, and audited distribution.

REPRODUCIBILITY STATEMENT

The release bundle records locks, provenance, statistical units, nulls, and sensitivities (Appendix I). Tracked bundles replay the 2,560/3,840-row compiler audits, the 24,576-row matched-level audit, and the 49,152-row level-ladder and 32,768-row calibration-budget audits, pinning arrays/executables or compressed outcomes plus lock and analyzer hashes. Verifiers recompute each decision; Walker2d-5100 re-execution reproduces every label and rate exactly. The release also retains the 8,192-row prospective coverage-ladder outcome and its locked analyzer (Appendix I.2).

The supplementary material is a single anonymous archive. It carries the separate auxiliary technical record cited throughout the appendices (`paper/iclr_auxiliary_controls.pdf`), the measurements and locked environment behind the headline results, and a verifier that checks the released bytes against their manifest and then runs `scripts/replay_iclr_sensitivity_release.py`. Without an accelerator, that replay recomputes the within-rule return and fidelity statistics, the fresh-policy paired contrasts from stored target actions and reward traces, and every registered audit and holdout decision from hash-bound episode summaries. Of the four fidelity sensitivities it recomputes the panel bootstrap in full and the other three summaries from their per-comparison rows; those three measurements need the policy runtime, the accelerator or a pinned simulator source and are not re-measured. Every step halts where a recomputed value disagrees with its sealed record. This is analysis replay: it neither re-executes anything on a device nor rebuilds an executable, and a rebuild need not reproduce the same bytes or action field. No independent reproduction by an outside reader is claimed.

The rebuild audit and OpenVINO construction probe reported here also provide source measurements for a related unpublished manuscript on provenance-field discrimination and execution-evidence reuse (Anonymous Authors, 2026). These measurements are shared, not independent replications; the present paper studies fixed-state fidelity and closed-loop task performance.

AI USE STATEMENT

Generative AI tools assisted with code drafting and implementation, figure layout, manuscript drafting and editing, and feedback on research methodology, experimental design, and result interpretation. They generated no measurements. The authors ran and inspected all code and retain full responsibility for the methodology, analysis, claims, and final manuscript.

REFERENCES

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/7090-hindsight-experience-replay.

Anonymous Authors. What should a quantized model's provenance record contain? An empirical audit of construction and behaviour. Unpublished manuscript, 2026.

Erik Aurell, Guido Boffetta, Andrea Crisanti, Giovanni Paladin, and Angelo Vulpiani. Predictability in the large: An extension of the concept of lyapunov exponent. *Journal of Physics A: Mathematical and General*, 30(1):1–26, 1997. doi: 10.1088/0305-4470/30/1/003.

Ron Banner, Yury Nahshan, and Daniel Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/c0a62e133894cdce435bcb4a5df1db2d-Abstract.html.

Roger W. Brockett. *Finite Dimensional Linear Systems*. SIAM, 2015. doi: 10.1137/1.9781611973884.

Roger W. Brockett and Daniel Liberzon. Quantized feedback stabilization of linear systems. *IEEE Transactions on Automatic Control*, 45(7):1279–1289, 2000. doi: 10.1109/9.867021.

Tianqi Chen, Thierry Moreau, Ziheng Jiang, Lianmin Zheng, Eddie Yan, Haichen Shen, Meghan Cowan, Leyuan Wang, Yuwei Hu, Luis Ceze, Carlos Guestrin, and Arvind Krishnamurthy. TVM:

An automated end-to-end optimizing compiler for deep learning. In *13th USENIX Symposium on Operating Systems Design and Implementation*, pp. 578–594. USENIX Association, 2018. URL https://www.usenix.org/conference/osdi18/presentation/chen.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *Advances in Neural Information Processing Systems*, 2019. arXiv:1905.11979.

David F. Delchamps. Stabilizing a linear system with quantized state feedback. *IEEE Transactions on Automatic Control*, 35(8):916–924, 1990. doi: 10.1109/9.58500.

Zhen Dong, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. HAWQ: Hessian AWare quantization of neural networks with mixed-precision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 293–302, 2019. doi: 10.1109/ICCV.2019.00038.

Zhen Dong, Zhewei Yao, Daiyaan Arfeen, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. HAWQ-V2: Hessian aware trace-weighted quantization of neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18518–18529, 2020.

Nicola Elia and Sanjoy K. Mitter. Stabilization of linear systems with limited information. *IEEE Transactions on Automatic Control*, 46(9):1384–1400, 2001. doi: 10.1109/9.948466.

Farama Foundation. Gymnasium MuJoCo environment documentation: Walker2D and Hopper. https://gymnasium.farama.org/environments/mujoco/walker2d/ and https://gymnasium.farama.org/environments/mujoco/hopper/, n.d. Accessed 2026-08-03.

Minyue Fu and Lihua Xie. The sector bound approach to quantized feedback control. *IEEE Transactions on Automatic Control*, 50(11):1698–1711, 2005. doi: 10.1109/TAC.2005.858689.

Google. TensorFlow Lite Python API. https://ai.google.dev/edge/api/tflite/python/tf/lite, 2026. Accessed 2026-08-16.

Google Coral. Edge TPU compiler. https://coral.ai/docs/edgetpu/compiler/, 2020a. Accessed 2026-08-03.

Google Coral. TensorFlow models on the edge TPU. https://coral.ai/docs/edgetpu/models-intro/, 2020b. Accessed 2026-08-03.

Google Coral. PyCoral API. https://github.com/google-coral/pycoral, n.d. Version 2.0.0; accessed 2026-08-16.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1861–1870, 2018.

Hailo. Hailo-8 AI accelerator. https://hailo.ai/products/ai-accelerators/hailo-8-ai-accelerator/, n.d. Accessed 2026-08-03.

Jesse Hemerik and Jelle Goeman. Exact testing with random permutations. *TEST*, 27:811–825, 2018. doi: 10.1007/s11749-017-0571-1.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*, 2017.

Itay Hubara, Yury Nahshan, Yair Hanani, Ron Banner, and Daniel Soudry. Accurate post training quantization with small calibration sets. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4466–4475. PMLR, 2021.

Radoslav Ivanov, James Weimer, Rajeev Alur, George J. Pappas, and Insup Lee. Verisig: Verifying safety properties of hybrid systems with neural network controllers. In *Proceedings of the 22nd ACM International Conference on Hybrid Systems: Computation and Control*, pp. 169–178, 2019. doi: 10.1145/3302504.3311806.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2704–2713, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Jacob_Quantization_and_Training_CVPR_2018_paper.html.

Zhong-Ping Jiang and Teng-Fei Liu. Quantized nonlinear control—a survey. *Acta Automatica Sinica*, 39(11):1820–1830, 2013. doi: 10.1016/S1874-1029(13)60079-8.

Fabian Kresse and Christoph H. Lampert. Learning quantized continuous controllers for integer hardware. In *Proceedings of The 8th Annual Learning for Dynamics and Control Conference*, volume 331 of *Proceedings of Machine Learning Research*, pp. 1857–1874. PMLR, 2026. URL https://proceedings.mlr.press/v331/kresse26a.html.

Kartik Krishna, Steven L. Brunton, and Zhuoyuan Song. Finite time lyapunov exponent analysis of model predictive control and reinforcement learning. *IEEE Access*, 11:118916–118930, 2023. doi: 10.1109/ACCESS.2023.3326424.

Srivatsan Krishnan, Maximilian Lam, Sharad Chitlangia, Zishen Wan, Gabriel Barth-Maron, Aleksandra Faust, and Vijay Janapa Reddi. QuaRL: Quantization for fast and environmentally sustainable reinforcement learning. *Transactions on Machine Learning Research*, 2022. arXiv:1910.01055.

Arsenii Kuznetsov, Pavel Shvechikov, Alexander Grishin, and Dmitry Vetrov. Controlling overestimation bias with truncated mixture of continuous distributional quantile critics. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5556–5566. PMLR, 2020. URL https://proceedings.mlr.press/v119/kuznetsov20a.html.

Mathias Lechner, Ramin Hasani, Alexander Amini, Tsun-Hsuan Wang, Thomas A. Henzinger, and Daniela Rus. Are all vision models created equal? a study of the open-loop to closed-loop causality gap. *arXiv preprint arXiv:2210.04303*, 2022.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. BRECQ: Pushing the limit of post-training quantization by block reconstruction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=POWv6hDd9XH.

Daniel Liberzon and Dragan Nešić. Input-to-state stabilization of linear systems with quantized state measurements. *IEEE Transactions on Automatic Control*, 52(5):767–781, 2007. doi: 10.1109/TAC.2007.895850.

Vivian Lin, Ramneet Kaur, Yahan Yang, Souradeep Dutta, Yiannis Kantaros, Anirban Roy, Susmit Jha, Oleg Sokolsky, and Insup Lee. Safety monitoring for learning-enabled cyber-physical systems in out-of-distribution scenarios. In *Proceedings of the ACM/IEEE 16th International Conference on Cyber-Physical Systems*, pp. 5:1–5:11, 2025. doi: 10.1145/3716550.3722022.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard, and Stefano Soatto. Robustness of classifiers to universal perturbations: A geometric perspective. In *International Conference on Learning Representations*, 2018. arXiv:1705.09554.

Siddharth Mysore, Bassel Mabsout, Renato Mancuso, and Kate Saenko. Regularizing action policies for smooth control with reinforcement learning. In *IEEE International Conference on Robotics and Automation*, pp. 1810–1816, 2021. doi: 10.1109/ICRA48506.2021.9561138.

Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7197–7206. PMLR, 2020. URL https://proceedings.mlr.press/v119/nagel20a.html.

Girish N. Nair, Fabio Fagnani, Sandro Zampieri, and Robin J. Evans. Feedback control under data rate constraints: An overview. *Proceedings of the IEEE*, 95(1):108–137, 2007. doi: 10.1109/JPROC.2006.887294.

Dragan Nešić and Daniel Liberzon. A unified framework for design and analysis of networked and quantized control systems. *IEEE Transactions on Automatic Control*, 54(4):732–747, 2009. doi: 10.1109/TAC.2009.2014930.

NVIDIA Corporation. NVIDIA TensorRT 10.3.0 Developer Guide. https://docs.nvidia.com/deeplearning/tensorrt/archives/tensorrt-1030/pdf/TensorRT-Developer-Guide.pdf, 2024. Accessed 2026-08-16.

Anay Pattanaik, Zhenyi Tang, Shuijing Liu, Gautham Bommannan, and Girish Chowdhary. Robust deep reinforcement learning with adversarial attacks. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2040–2042, 2018. arXiv:1712.03632.

Matthias Plappert, Marcin Andrychowicz, Alex Ray, Bob McGrew, Bowen Baker, Glenn Powell, Jonas Schneider, Josh Tobin, Maciek Chociej, Peter Welinder, Vikash Kumar, and Wojciech Zaremba. Multi-goal reinforcement learning: Challenging robotics environments and request for research. *arXiv preprint arXiv:1802.09464*, 2018.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021.

Adnan Siraj Rakin, Zhezhi He, and Deliang Fan. Bit-flip attack: Crushing neural network with progressive bit search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1211–1220, 2019. doi: 10.1109/ICCV.2019.00130.

Andrei A. Rusu, Sergio Gomez Colmenarejo, Çağlar Gülçehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. In *International Conference on Learning Representations*, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Yuzhang Shang, Gaowen Liu, Ramana Rao Kompella, and Yan Yan. Enhancing post-training quantization calibration through contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15921–15930, 2024.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024.

Hoang-Dung Tran, Xiaodong Yang, Diego Manzanas Lopez, Patrick Musau, Luan Viet Nguyen, Weiming Xiang, Stanley Bak, and Taylor T. Johnson. NNV: The neural network verification tool for deep neural networks and learning-enabled cyber-physical systems. In *Computer Aided Verification*, pp. 3–17. Springer, 2020. doi: 10.1007/978-3-030-53288-8_1.

Yixuan Wang, Weichao Zhou, Jiameng Fan, Zhilu Wang, Jiajun Li, Xin Chen, Chao Huang, Wenchao Li, and Qi Zhu. POLAR-Express: Efficient and precise formal reachability analysis of neural-network controlled systems. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 43(3):994–1007, 2024. doi: 10.1109/TCAD.2023.3331215.

Bernard Widrow and Istvan Kollar. *Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications*. Cambridge University Press, 2008. doi: 10.1017/CBO9780511754661.

Wing Shing Wong and Roger W. Brockett. Systems with finite communication bandwidth constraints. II. stabilization with limited information feedback. *IEEE Transactions on Automatic Control*, 44(5):1049–1053, 1999. doi: 10.1109/9.763226.

Weiming Xiang. Runtime safety monitoring of neural-network-enabled dynamical systems. *IEEE Transactions on Cybernetics*, 52(9):9587–9596, 2022. doi: 10.1109/TCYB.2021.3053575.

Weiming Xiang and Taylor T. Johnson. Reachability analysis and safety verification for neural network control systems. *arXiv preprint arXiv:1805.09944*, 2018.

Siyuan Xu, Tianshi Wang, Fengling Li, Lei Zhu, and Heng Tao Shen. DA-PTQ: Drift-aware post-training quantization for efficient vision-language-action models. *arXiv preprint arXiv:2604.11572*, 2026. URL https://arxiv.org/abs/2604.11572.

Emily Yu, Đorđe Žikelić, and Thomas A. Henzinger. Neural control and certificate repair via runtime monitoring. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(25):26409–26417, 2025. doi: 10.1609/aaai.v39i25.34840.

Haichao Yu, Linjie Yang, and Humphrey Shi. Is in-domain data really needed? a pilot study on cross-domain calibration for network quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3043–3052, 2021.

Jingxuan Zhang, Yunta Hsieh, Zhongwei Wan, Haokun Lin, Xin Wang, Ziqi Wang, Yingtie Lei, and Mi Zhang. QuantVLA: Scale-calibrated post-training quantization for vision-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39539–39549, June 2026.

Lianmin Zheng, Chengfan Jia, Minmin Sun, Zhao Wu, Cody Hao Yu, Ameer Haj-Ali, Yida Wang, Jun Yang, Danyang Zhuo, Koushik Sen, Joseph E. Gonzalez, and Ion Stoica. Ansor: Generating high-performance tensor programs for deep learning. In *14th USENIX Symposium on Operating Systems Design and Implementation*, pp. 863–879. USENIX Association, 2020. URL https://www.usenix.org/conference/osdi20/presentation/zheng.

Jiacheng Zhou, Jinfan Lv, Ruixuan Li, Longtai Zhang, Yan Wang, Wenqiang Zhang, and Lizhe Qi. QuantWAMs: Calibrating at the right granularity for world action models. *arXiv preprint arXiv:2607.28405*, 2026. URL https://arxiv.org/abs/2607.28405.

Kemin Zhou, John C. Doyle, and Keith Glover. *Robust and Optimal Control*. Prentice Hall, Upper Saddle River, NJ, 1996. ISBN 9780134565675.

Appendix Reading Guide and Claim–Evidence Map

For the main empirical claim, begin with the across-artifact and within-rule fidelity–return comparisons. The fresh-policy benchmark separates secondary fidelity evidence from qualification-decision results. Compiler controls, local-proxy tests, and reproduction records follow; experimental design status and registered outcomes are preserved throughout. A missed gate remains a miss even when a secondary contrast is positive. Topic blocks are not statistical families. Table 2 links each reader question to its evidence.

Table 2: Claim–evidence map. Historical experiment identifiers and current locations are linked in Table 80.

Main claim / reader question	Appendix	Key evidence	Scope
Does fidelity improvement imply task improvement?	B	Complete family, within-rule intervals, fresh-policy contrasts	Statistical and practical claims differ by experiment
Do effects recur across policies and builds?	C–D	Policy replication, mixtures, compiler/GPU/budget controls	Replication versus robustness
What do qualification decisions change?	E	Prospective benchmark, matched-deferral replay, selection audit	CI superiority and material selection failure not established
What do timing and local dynamics explain?	F–G	Timing interventions and proxy stress tests	Constructed interventions and local scope
How can the experiments be reproduced?	A, I	Contracts, manifests, result index, re-execution	Reproduction level is stated explicitly

A Shared Experimental and Statistical Contracts

A.1 Statistical Units and Selection Contracts

Table 3 distinguishes inferential units from workload. Each retained experiment reports its own sample sizes, interval family, selection rules, and analysis status in its result section; the full experiment-by-experiment workload ledger remains in the separate technical record. This summary does not assign a common inferential family to those audits.

Table 3: Statistical units and limits of pooling. More artifacts, windows, or replays do not automatically supply more independent policies.

Analysis	Unit and selection contract
Fixed-policy task audits and artifact families	Paired reset episode seed. Arms share starts and endpoints. Complete-family ranges account for selection of extrema; artifact rows sharing a policy are not independent policies. Multiplicity follows each audit's declared family.
Cross-policy contrasts	Independently trained policy seed is the top-level unit. Episode comparisons are aggregated within policy; known-contrast or capability-conditioned cohorts retain their selection limitations.
Compiler-path controls	Paired reset seeds within the declared cell or family for task effects; rebuilds and repeated executions are construction/runtime controls, not additional trained policies. Outcome-informed reruns are labelled separately from locked fresh-outcome comparisons.
Local mechanism interventions	Nominal episode within a policy/backend panel. Windows share trajectories; episode-cluster intervals preserve that dependence. Eligibility and horizon/margin strata constrain the claim. Saturation stress and descriptive secondary strata do not replace a registered primary.
Stored proxy rankings	State/window-level diagnostic rankings with their reported resampling scheme, not policy-population inference. Ranking an alert does not establish task prediction or calibrated qualification.
Decision comparisons	Paired task episodes support within-artifact decisions; artifacts share policies at cohort level. Retrospective resplits overlap and are not independent cohorts. Fresh-policy and prefix-sensitivity results retain their distinct status, including unresolved holdout cases.
Eligibility inventories	Policy-level descriptive counts with their original selection denominator. Targeted later additions do not enlarge the original pass-rate denominator or turn it into a representative policy sample.

Pairing controls baseline difficulty, not absolute safety. The registered audit, later conservative sensitivity, and practical recommendation remain separate, as specified in Section 3. Local score units and the outcome-dependent nature of retrospective conditioning are detailed in Appendices G.1.1 and G.5.6.

A.2 Policy, Task, and Panel Contracts

Training and policy class. The paper-facing Walker2d and Hopper panels use deterministic mean actions from Stable-Baselines3 PPO actor-critic policies in Gymnasium MuJoCo v5 environments (Raffin et al., 2021; Schulman et al., 2017; Towers et al., 2024; Todorov et al., 2012). FetchReach uses soft actor-critic policies (Haarnoja et al., 2018); FetchPush uses a 256–256 ReLU behavior-cloning student distilled from a TQC+HER teacher (Kuznetsov et al., 2020; Andrychowicz et al., 2017). Both use the multi-goal robotics environments (Plappert et al., 2018). The PPO and SAC policy and value networks are two-hidden-layer MLPs with 256 ReLU units per layer. Walker2d uses 17-dimensional observations and 6-dimensional actions; Hopper uses 11-dimensional observations and 3-dimensional actions. Each Walker2d/Hopper seed was trained for 200,000 environment steps with one environment, PPO rollout length 2,048, minibatch size 64, 10 PPO epochs per update, learning rate 3×10^{-4}, discount $\gamma = 0.99$, and clip range 0.2. These policies do not use VecNormalize; the external observation transform is the identity. Rollouts apply the raw deterministic mean action after the environment action-box clip. The exported actor graph itself does not contain obser-

vation normalization, action scaling, or clipping. Table 4 summarizes the policy classes, dimensions, and stored panels used by the paper-facing audits.

Table 4: Policy, panel, and calibration details for the paper-facing MuJoCo actors, Humanoid inventory policy, and GoalEnv task audits. Panel and calibration counts are stored observations unless a larger appendix expansion is explicitly named. For each of the eight FetchReach rules, TFLite receives 256 calibration rows and the historical DFC families receive a stride-4 64-row subset; the calibration-budget audit additionally tests DFC head-64 and full-256 inputs.

Role and seeds	Environment	Obs./act.	Policy/training	Normalization and panels
Main mechanism; seeds 5100, 5400, 5700, 5800	Walker2d-v5	17/6	PPO, 256–256 ReLU, 200k steps	Identity external transform; 2,048/2,048 state and calibration panels; 5700 expansion has 8,192 states
Saturation-stress main panel; seed 6100	Hopper-v5	11/3	PPO, 256–256 ReLU, 200k steps	Identity external transform; 2,048/2,048 state and calibration panels
Appendix inventory and precheck; seed 7100 S1	Humanoid-v5	348/17	PPO, 256–256 ReLU, 3M checkpoint plus 2M-step resume	External VecNormalize; 2,048/2,048 state and calibration panels
Full-episode task breadth; seed 9400	FetchPush Dense v4	28/4	BC, 256–256 ReLU, 25,600 TQC+HER teacher pairs, 1,000 epochs	Fixed input standardization; 4,096 calibration states; 128 capability seeds; the total-artifact and calibration-family audits use separate locked 512-seed blocks
Full-episode task replication; seed 9500	FetchReach Dense v4	13/4	SAC, 256–256 ReLU, 50k steps	Identity external transform; 2,048 stored states; 128 capability seeds; separately locked 512-seed blocks for the original, count, subset, TensorRT, and level controls; 1,024 for the complete family, 800 selection states, and 256 for breadth
Cross-policy artifact replication; seeds 9510–9517	FetchReach Dense v4	13/4	SAC, 256–256 ReLU, 50k steps	Identity external transform; disjoint 800-state calibration/measurement panels per policy; 128 capability and 256 outcome seeds per policy; CAL3 also audits seed 9510 on 256 fresh starts

Nominal state panels and calibration sets. We use three panel identifiers where multiple collections exist for one policy: **P-orig** is the original 2,048-state collection, **P-4096** is the larger 4,096-state Walker2d episode-count collection, and **P-exp** is the 8,192-state Walker2d-5700 precision expansion. Reported window counts are the eligible subsets of these collections, not the raw panel sizes. For each policy seed, nominal states are collected by resetting the MuJoCo environment with the registered panel seed and rolling out the deterministic FP32 actor until the requested state count is reached. The state panel stores the state identifier, nominal episode identifier, time index, MuJoCo simulator state, raw observation, normalized actor input, and FP32 reference action. The

mechanism analyses use common valid states across FP32, EdgeTPU, and Hailo backend dumps; horizon-valid windows additionally require the full nominal suffix for the chosen horizon. The default gate and calibration panels contain 2,048 states per actor. The Walker2d-5700 expansion panel contains 8,192 states from 41 nominal episodes and preserves the original 2,048-state prefix exactly. The post-training quantization calibration files are 2,048-observation NPZ artifacts saved with the actor bundle and converted to backend-specific calibration arrays as needed. The calibration axis audit in Appendix G.3.5 changes this calibration set while holding the policy, graph, backend, and evaluation states fixed.

Export and backend measurement. The exported FP32 reference actor is a fixed-batch ONNX graph with opset 17 and IR version 10. The graph contains only affine layers, ReLU activations, and the final linear action head; in the checked manifests the allowed operator set is `Gemm` and `Relu`. EdgeTPU artifacts are compiled from the integer TFLite export of the same actor graph and executed through the Coral PyCoral stack (Google, 2026; Google Coral, n.d.). Hailo artifacts are compiled from the corresponding quantized graph with the Hailo dataflow compiler. Each backend runner evaluates the stored actor inputs and writes one action vector per state to JSONL, together with policy hashes, model hashes, compiler version, runtime version, and repeat identifiers. A model hash pins the exact file that was measured, which is what the audits require; it does not imply that recompiling the same source reproduces that file. In the historical rebuild checks, both vendor compilers produced different artifact bytes, while integer TFLite exports and the checked Hailo quantization parameters and calibration statistics agreed. Independently compiled artifacts agreed exactly in the tested artifact–panel cells, not necessarily elsewhere (Appendix I.2); the later 200-build audit finds SDK-emulator field variation in 4 of 20 cells. The raw measured error field is

$$e_b(s) = \pi_b(s) - \pi_{\mathrm{fp32}}(s),$$

computed before the external action-box clip. Post-clip effective errors are also reported in the saturation, Humanoid, and E20 scalar-match audits.

A.3 Analysis roles and family-adjustment contract

The decision rule compares an interval against a tolerance, so the interval has to cover at the level the rule assumes. The main tables report marginal intervals per backend. Reading two of them together as one audit is a multiplicity the marginal level does not pay for. The studentized-max construction described here was added after outcomes as a conservative multiplicity audit; it was not part of the outcome locks and cannot support a stronger status. We recompute both backends from one shared bootstrap: paired FP32-minus-artifact differences are resampled by reset seed jointly across the two backends of an audit, and the family band uses the studentized maximum deviation over that family, with 20,000 replicates.

The analysis-role record (Appendix I.4) keeps `analysis_role`, `locked_before_outcome`, and `interval_scope` distinct for each analysis, alongside `registered_status`, `sensitivity_status`, and the practical recommendation. The registered entries use the original intervals and Holm tests, not this section's recomputed marginal intervals. The family band is a bootstrap sensitivity, not a finite-sample nominal-coverage guarantee. A future audit must fix its family and confidence procedure before outcomes; these historical results are not retrofitted to that prospective protocol.

B FIXED-STATE FIDELITY VERSUS CLOSED-LOOP RETURN

B.1 ACROSS-ARTIFACT FIDELITY ORDERING

B.1.1 COMPLETE FRESH FETCHREACH CALIBRATION-ARTIFACT FAMILY

E21 was locked after E20 as a separate fresh-seed follow-up. It includes all 16 pre-existing E20 artifacts rather than selecting another pair, plus FP32 and the two E18 references. Every condition uses the same 1,024 reset seeds ("103101–104124") and 50-step horizon. E18 reference outcomes and the E20 episode-head/phase-early outcomes were disclosed as known before lock; outcomes for the other 14 family artifacts and all E21 seeds were unopened. No artifact, seed, endpoint, threshold, or sample extension was allowed.

The primary analysis is a paired Friedman test across the eight fixed artifacts within each backend, with Holm adjustment across the two tests. The practical effect is the selection-aware bootstrap interval for the fixed-family best-minus-worst mean-return range. Table 5 reports the locked decisions.

Table 5: Complete fresh FetchReach calibration-family primary result. Higher return is better. The EdgeTPU range has indeterminate heterogeneity; the Hailo range is materially heterogeneous because its interval lies above the locked 0.5-return threshold. These statuses judge within-backend range materiality, not an individual artifact's degradation against FP32; the registered records retain their original labels.

Backend	Best artifact	Best return	Worst artifact	Worst return	Range [95% CI]	Holm p	Decision
EdgeTPU	Even episodes	-0.803	Phase late	-1.292	0.489 [0.460, 0.519]	< 0.001	Indeterminate heterogeneity
Hailo	Even episodes	-0.820	Phase middle	-15.747	14.926 [14.318, 15.549]	< 0.001	Materially heterogeneous

Table 6 shows that the large Hailo range is concentrated in phase-middle/late artifacts, whose maximum absolute effective action errors on the fixed-state panel are both 1.635. The descriptive scalar-error/degradation Spearman correlation is 0.286 ($p = 0.493$) for EdgeTPU and 0.833 ($p = 0.010$) for Hailo. The latter is descriptive because there are only eight artifacts and two Hailo outcomes were known before E21. It means the dramatic Hailo family result is not direction-only evidence; scalar fidelity is informative in that family.

Table 6: Complete calibration-family results by artifact. Return degradation is FP32 minus artifact return; positive is adverse. Mean error norm was fixed on the 800-state scalar-selection panel, before the complete-family task outcomes.

Backend	Calibration rule	Mean error norm	Mean return	Return degradation [95% CI]	Terminal success
EdgeTPU	Episode head	0.0969	-0.9895	0.3264 [0.2746, 0.3806]	0.9170
EdgeTPU	Episode tail	0.1533	-0.8052	0.1421 [0.1137, 0.1735]	0.9658
EdgeTPU	Even episodes	0.1556	-0.8028	0.1397 [0.1101, 0.1718]	0.9668
EdgeTPU	Odd episodes	0.1074	-0.9891	0.3260 [0.2748, 0.3807]	0.9160
EdgeTPU	Phase early	0.1600	-0.8069	0.1438 [0.1134, 0.1759]	0.9639
EdgeTPU	Phase middle	0.1904	-1.2918	0.6287 [0.5975, 0.6626]	0.9795
EdgeTPU	Phase late	0.1791	-1.2920	0.6289 [0.5967, 0.6607]	0.9785
EdgeTPU	Fixed random	0.1493	-0.8064	0.1433 [0.1136, 0.1751]	0.9678
Hailo	Episode head	0.1383	-0.9671	0.3040 [0.2602, 0.3487]	0.9268
Hailo	Episode tail	0.1274	-0.8395	0.1764 [0.1411, 0.2133]	0.9639
Hailo	Even episodes	0.1120	-0.8203	0.1572 [0.1247, 0.1910]	0.9658
Hailo	Odd episodes	0.1124	-0.9011	0.2380 [0.1978, 0.2800]	0.9463
Hailo	Phase early	0.1356	-0.8372	0.1741 [0.1401, 0.2119]	0.9570
Hailo	Phase middle	0.1779	-15.7465	15.0834 [14.4842, 15.7165]	0.2139
Hailo	Phase late	0.1670	-15.6283	14.9652 [14.3295, 15.6054]	0.2197
Hailo	Fixed random	0.2114	-1.1060	0.4429 [0.3753, 0.5120]	0.9238

The norm reversal is a descriptive contrast within this fixed family, not an isolated manipulation of scalar error. Phase-middle's marginal loss interval has lower bound 14.484, above the 0.5 tolerance. Fixed-random's interval $[0.375, 0.512]$ crosses that tolerance, so this is not a comparison between

an approved artifact and a rejected one. Neither the ratio of point losses nor the descriptive rank reversal supplies a simultaneous confidence interval for a selected pairwise contrast.

The EdgeTPU return-degradation values that tie at three decimals come from distinct artifacts and action functions. On the shared 800-state panel, episode-head and odd-episodes differ on 525 states (819 of 3,200 action components; RMSE 0.02712; maximum component difference 0.21875), while phase-middle and phase-late differ on 492 states (880 components; RMSE 0.02606; maximum difference 0.17969). The four-decimal task summaries above expose the smaller outcome differences as well.

All five scalar-matched pairs fixed before E20 outcomes were also evaluated on the new block. Their absolute mean-return differences range from 0.081 to 0.169, and every paired interval in Table 7 lies within $[-0.5, +0.5]$ despite all five Holm-adjusted tests being below 0.001.

Table 7: Supporting scalar-matched artifact pairs. Contrasts are artifact A minus artifact B return; all pairs were selected from fixed-state scalar and direction metrics before the scalar-matched-pair task outcomes.

Artifact A	Artifact B	Backend relation	A − B return [95% CI]	Holm p
EdgeTPU odd	Hailo even	Cross-backend	-0.169 [-0.214, -0.127]	< 0.001
EdgeTPU odd	Hailo odd	Cross-backend	-0.088 [-0.111, -0.066]	< 0.001
Hailo head	Hailo tail	Same-backend	-0.128 [-0.171, -0.086]	< 0.001
Hailo early	Hailo head	Same-backend	0.130 [0.090, 0.172]	< 0.001
Hailo even	Hailo odd	Same-backend	0.081 [0.049, 0.115]	< 0.001

The Hailo episode-head/phase-early contrast independently repeats the E20 ordering: phase-early minus head is 0.130 [0.090, 0.172], versus 0.156 [0.101, 0.215] in E20. Thus scalar matching does not imply identical task behavior, but the matched-pair effects observed here are modest. E21 supports artifact-specific qualification; it does not isolate direction, validate CLAE as a task selector, or estimate prevalence over calibration procedures.

Third-backend boundary. A separately locked TensorRT 10.3 explicit-Q/DQ family evaluated the same eight calibration rules on 512 fresh paired FetchReach starts. The original fixed-panel maximum-error guard had failed on one state; without rebuilding or replacing any engine, all eight exact engines then passed an independent, outcome-blind cross-policy holdout before task outcomes were authorized. The paired Friedman test detects heterogeneity ($p < 0.001$), but the selection-aware range is only 0.088 [0.082, 0.127], wholly below 0.5, and terminal success spans 93.4–95.9%. This third path bounds practical magnitude; its explicit-Q/DQ contract is not equivalent to either EdgeTPU or Hailo.

B.1.2 Fresh scalar-matched FetchReach artifact contrast

The two E18 reference artifacts differ substantially in scalar fidelity, so we did not reuse their opened outcomes for a matched comparison. Instead, E20 fixed eight calibration rules, compiled all 16 EdgeTPU/Hailo candidates, and measured them on an independent 800-state panel. Five of 120 possible pairs passed the prospectively locked two scalar gates (mean L2 and action RMSE) plus three eligibility filters (support, repeatability, and direction separation). The same-runtime-first rule selected Hailo episode-head versus Hailo phase-early before either task outcome was opened. Their post-clip mean-L2 and action-RMSE symmetric ratios are 1.020 and 1.035; their median statewise error-direction cosine is 0.203.

The locked outcome block used 512 fresh paired reset seeds, a 50-step cap, and phase-early minus episode-head dense return as its sole primary contrast. Table 8 reports the selected artifacts and paired results.

The paired return sign-flip test gives $p < 0.001$. Phase-early also ends 0.00498 m [0.00302, 0.00713] closer to the goal. The return difference is statistically clear but modest under the locked threshold. This one selected pair is a direct counterexample to scalar-fidelity sufficiency for task behavior, not a prevalence estimate. Because calibration changes more than the measured field's scalar summaries,

Table 8: Fresh same-runtime scalar-matched FetchReach audit. The artifact pair was selected without task-outcome access. Higher return and success are better. The return contrast is phase-early minus episode-head; its full interval is inside the locked $[-0.5, +0.5]$ practical band.

Hailo artifact	Mean error norm	Action RMSE	Mean return	Terminal success	Ever success
Episode-head	0.1383	0.07454	-0.9735	0.9316	0.9980
Phase-early	0.1356	0.07712	-0.8178	0.9766	1.0000
Paired phase-early − head	–	–	0.1557 [0.1012, 0.2151]	0.0449 [0.0254, 0.0664]	0.0020 [0.0000, 0.0059]

the result is a total-artifact contrast and does not isolate direction or validate CLAE as a task-outcome selector.

B.2 Within-rule compiler intervention

For a fixed source policy and calibration rule, CAL10 compares full-budget level-0 and AdaRound-enabled level-3 configurations, with all supplied rows used for quantizer statistics and level-3 training. Table 9 reports the paired task block and independent fixed-state panel. Compiler-budget construction details remain in Appendix D.4.

Table 9: Full-budget fidelity–return comparison by calibration rule. Returns are means on the same 1,024 paired starts; Δ is level 3 minus level 0, so positive is task improvement. Fixed-panel RMSE is against FP32 on the same 800 states and lower is better. The AdaRound-enabled level-3 compiler intervention lowers RMSE in all eight rules but improves mean return in only odd-episodes, phase-middle, and phase-late. The supporting Bonferroni family intervals for all negative changes lie inside the 0.5 practical band.

Rule	L0 return	L3 return	Return Δ	L0 RMSE	L3 RMSE
Episode head	-0.875	-0.972	-0.098	0.0963	0.0741
Episode tail	-0.791	-0.880	-0.089	0.0590	0.0586
Even episodes	-0.777	-0.826	-0.049	0.0569	0.0507
Odd episodes	-0.993	-0.938	$+0.054$	0.0922	0.0596
Phase early	-0.814	-0.846	-0.032	0.0515	0.0485
Phase middle	-15.746	-15.246	$+0.501$	0.1693	0.1640
Phase late	-15.628	-15.238	$+0.390$	0.1642	0.1640
Fixed random	-0.805	-0.856	-0.051	0.0537	0.0486
Family mean	-4.554	-4.475	$+0.079$	0.0929	0.0835

The full-budget level-3 compiler intervention, with AdaRound enabled, improves fixed-panel action RMSE against FP32 in all eight rules but improves mean paired task return in three (Table 9). Its family mean of $+0.079$ $[+0.041, +0.116]$ is carried by the two collapsed rules, which recover $+0.446$ $[+0.299, +0.593]$, while the benign six lose -0.044 $[-0.055, -0.033]$. A post-outcome paired sensitivity puts the level-3 range change at -0.550 $[-0.712, -0.389]$, a 3.7% reduction that leaves the spread about 28 times the practical threshold; it is a descriptive follow-up, not a locked equivalence result.

Rule-wise uncertainty and practical scale (supporting sensitivity). The original registered summaries above are unchanged. Jointly resampling the 1,024 paired episode seeds 5,000 times gives Table 10 and Figure 1B (family intervals). Five of eight point estimates are negative; five marginal 95% intervals are wholly negative, as are their eight-rule Bonferroni intervals in the released record. All five family intervals lie entirely within $(-0.5, 0)$; the largest deterioration upper bound is 0.135 return. This bounds the additional L3-versus-L0 deterioration, not either artifact's loss relative to FP32. The evidence is therefore fidelity–task non-monotonicity with small decreases, not material harm in the affected rules. The intervention changes the compiler configuration as a whole; it does not isolate AdaRound's algorithmic causal effect.

Fidelity-panel independence. The measurement panel is a separate stored FP32 trajectory collection (seed 19602) from the calibration collection (seed 9502), using the same policy. Their integer state IDs are source-local, so intersecting IDs across the two archives would be invalid. We instead verify collection and calibration-archive hashes and compare exact float32 observation rows.

Table 10: Post-outcome paired uncertainty audit. $\Delta J_r = J_{r,\mathrm{L3}} - J_{r,\mathrm{L0}}$; intervals are marginal paired percentile-bootstrap 95% intervals. The last column counts measurement states unused by that artifact's calibration and learned-stage training. Hashes, intervals, and stratum counts are retained in the calibration-budget uncertainty record (Appendix I.4).

Rule	ΔJ	Paired 95% CI	Unused panel rows
episode head	-0.098	$[-0.126, -0.071]$	800
episode tail	-0.089	$[-0.107, -0.071]$	800
even episodes	-0.049	$[-0.060, -0.038]$	800
odd episodes	$+0.054$	$[+0.041, +0.068]$	800
phase early	-0.032	$[-0.041, -0.024]$	800
phase middle	$+0.501$	$[+0.348, +0.669]$	800
phase late	$+0.390$	$[+0.247, +0.551]$	800
fixed random	-0.051	$[-0.065, -0.038]$	800

Each rule has zero overlaps: all 800 measurement states are absent from each rule's calibration and learned-stage training rows. Thus all eight RMSE improvements hold on calibration-unused states; the overlapping stratum is empty and its RMSE is unmeasured, not zero. This establishes panel separation rather than distributional generalization. This stored-data audit adds no rollout.

B.3 Secondary fidelity–return contrasts in the fresh-policy benchmark

This analysis uses fresh FetchReach policies and Hailo artifacts from the same prospective task-decision cohort and task holdout in Appendix E.1, not another policy replication. The cohort, calibration/fidelity split, and compiler configurations are defined there; rules are clustered within policies. The later timing intervention in Appendix F.1 reuses these policies on fresh episode blocks.

Secondary held-out contrasts reduce action RMSE on a panel disjoint from calibration in 53/64 policy–rule comparisons; 22/53 accompany lower mean return. These are point-sign reversals, not individually significant or material-harm counts. The 1 return-deterioration interval wholly beyond tolerance accompanies worse, not better, RMSE. It is not a material-harm example of improved fidelity. This compares full compiler configurations, not AdaRound in isolation. Figure 4 plots all 64 comparisons, so the reversal sizes are visible alongside their counts.

Table 11 summarizes the secondary fidelity–return comparison by policy.

Table 11: Secondary fidelity–return comparisons in the prospective task-decision cohort: level 3 minus level 0, with 8 calibration rules per policy. The last column counts negative task-holdout mean-return changes only among RMSE-improved rules. Rules are clustered within policies; counts do not establish individual significance or material harm.

Policy ID	RMSE-improved rules	Of these, lower mean return
seed202609201	6	1
seed202609202	7	5
seed202609203	7	1
seed202609204	5	0
seed202609205	6	6
seed202609206	6	3
seed202609207	8	3
seed202609208	8	3
Total	53	22

Interval audit of the secondary reversals. This post-outcome reporting audit intersects the fixed-panel RMSE sign with the already computed paired return intervals. It adds no hypothesis test or new confidence band. The original interval families cover rules within each policy, not all policies jointly or the subset selected for improved RMSE. The table is not an estimate of population reversal prevalence. Fidelity-panel sampling uncertainty is not included in this table's original interval counts; task intervals alone do not resolve RMSE improvement. The separate sensitivity in

Appendix B.3 addresses panel resampling. Table 12 reports the original policy families separately. Material interval means the entire return-change interval is below -0.5. Counting intervals here

Table 12: Existing uncertainty attached to fidelity improvements. All counts after the second column are restricted to RMSE-improved rules. Policy indices follow Table 11; rows are not independent artifact cohorts.

Policy	RMSE improves	Mean decreases	Interval below zero	Material interval
1	6	1	0	0
2	7	5	3	0
3	7	1	0	0
4	5	0	0	0
5	6	6	6	0
6	6	3	3	0
7	8	3	1	0
8	8	3	2	0
Total	53	22	15	0

neither changes registered decisions nor establishes material harm from a point-sign reversal.

Sensitivity to fidelity-panel episode sampling. A post-outcome sensitivity analysis resamples whole retained FP32 fidelity episodes within each policy, using common draws across rules and compiler levels and recomputing pooled action RMSE. Raw clipped device actions first reproduce every original panel RMSE. Of 53 observed improvements across 64 comparisons, 16 retain negative RMSE-change intervals under the approximate Bonferroni-adjusted percentile bootstrap over the full comparison family. For 5 comparisons, this also coincides with an existing negative within-policy return interval. That intersection is descriptive, not joint simultaneous inference: no combined error allocation over both axes was specified. Independent episode blocks do not preclude such an analysis. Many fixed-panel signs therefore do not establish panel-sampling robustness, which falsifies this plan's registered stability prediction; the original fixed-panel counts are unchanged. This analysis neither recollects an independent panel nor estimates population reversal prevalence; extreme adjusted percentile tails have limited Monte Carlo resolution. Original fixed-panel results are unchanged. We use 10,000 bootstrap replicates per policy, PCG64 with SeedSequence([20260921, zero-based policy index]), and NumPy's linear quantiles. Marginal intervals use quantiles 0.025 and 0.975; the nominal 95% family intervals use 0.000390625 and 0.999609375. Only 3.90625 draws are expected in each adjusted tail. Bonferroni adjustment does not make percentile-bootstrap coverage exact. More replicates reduce Monte Carlo error, not uncertainty from the limited number of source episodes. Whole episodes are sampled with replacement; we pool squared component errors before taking the square root, rather than average episode RMSEs. Original RMSE reproduction requires exact numerical equality (no tolerance).

Table 13: Complete policy accounting for the post-outcome panel sensitivity. Stable means the adjusted RMSE-change upper endpoint is below zero. Intersection additionally requires the existing return interval to be negative; it is not a joint confidence statement.

Policy	Episodes	Comparisons	Improved	Stable	Intersection
202609201	16	8	6	0	0
202609202	16	8	7	4	2
202609203	16	8	7	2	0
202609204	16	8	5	0	0
202609205	16	8	6	0	0
202609206	16	8	6	3	1
202609207	16	8	8	2	0
202609208	16	8	8	5	2

Table 14: All panel-bootstrap comparisons, including null and opposite-direction results. Changes are L3 minus L0. RMSE brackets are full-family adjusted; return brackets are the original within-policy family intervals. An asterisk identifies the descriptive intersection. Policy IDs map to Table 13.

Policy	Rule	ΔRMSE [interval]	ΔJ [interval]	Intersection
202609201	episode_head	0.00746 [-0.00975, 0.02913]	-0.3873 [-0.4529, -0.3258]	–
202609201	episode_tail	0.01418 [-0.02047, 0.03563]	-0.6720 [-0.7819, -0.5695]	–
202609201	even_episodes	-0.00724 [-0.03392, 0.01273]	0.1289 [0.0925, 0.1696]	–
202609201	odd_episodes	-0.01050 [-0.03742, 0.02462]	0.2705 [0.2179, 0.3248]	–
202609201	phase_early	-0.00180 [-0.01465, 0.00523]	0.0390 [0.0253, 0.0535]	–
202609201	phase_middle	-0.00789 [-0.02084, 0.00488]	0.0580 [-0.4215, 0.4960]	–
202609201	phase_late	-0.00754 [-0.02017, 0.00501]	-0.0464 [-0.5778, 0.4308]	–
202609201	fixed_random	-0.01985 [-0.03349, 0.00144]	0.4042 [0.3287, 0.4829]	–
202609202	episode_head	-0.02974 [-0.05923, 0.00803]	0.0196 [-0.0044, 0.0450]	–
202609202	episode_tail	0.00136 [-0.01319, 0.01269]	0.0871 [0.0705, 0.1030]	–
202609202	even_episodes	-0.01245 [-0.05598, 0.03403]	-0.0442 [-0.0786, -0.0051]	–
202609202	odd_episodes	-0.02452 [-0.04449, -0.00223]	-0.0522 [-0.0669, -0.0376]	*
202609202	phase_early	-0.02355 [-0.04541, -0.00138]	-0.0313 [-0.0455, -0.0154]	*
202609202	phase_middle	-0.02994 [-0.04498, -0.01350]	-0.0968 [-0.2973, 0.1180]	–
202609202	phase_late	-0.03023 [-0.04553, -0.01355]	-0.1001 [-0.3005, 0.1171]	–
202609202	fixed_random	-0.02148 [-0.04085, 0.00247]	0.0810 [0.0698, 0.0918]	–
202609203	episode_head	-0.00558 [-0.07152, 0.03860]	0.1276 [0.1010, 0.1529]	–
202609203	episode_tail	-0.00395 [-0.03536, 0.03124]	0.0540 [0.0361, 0.0718]	–
202609203	even_episodes	0.00480 [-0.02533, 0.04016]	0.0149 [-0.0011, 0.0307]	–
202609203	odd_episodes	-0.06182 [-0.08569, -0.02986]	0.0523 [0.0318, 0.0737]	–
202609203	phase_early	-0.02641 [-0.05041, 0.00314]	0.0633 [0.0471, 0.0797]	–
202609203	phase_middle	-0.01200 [-0.02347, -0.00186]	0.0124 [-0.3909, 0.3817]	–
202609203	phase_late	-0.00795 [-0.01680, 0.00035]	-0.0328 [-0.4027, 0.3024]	–
202609203	fixed_random	-0.01396 [-0.02989, 0.00703]	0.0998 [0.0898, 0.1096]	–
202609204	episode_head	0.00612 [-0.01658, 0.02818]	0.0163 [-0.0008, 0.0320]	–
202609204	episode_tail	0.00639 [-0.01917, 0.02340]	0.0722 [0.0378, 0.1125]	–
202609204	even_episodes	-0.00402 [-0.01565, 0.00920]	0.0233 [0.0142, 0.0322]	–
202609204	odd_episodes	-0.00094 [-0.04020, 0.04107]	0.1092 [0.0520, 0.1760]	–
202609204	phase_early	-0.00735 [-0.02167, 0.00847]	0.0120 [0.0052, 0.0189]	–
202609204	phase_middle	0.00020 [-0.00243, 0.00439]	0.1360 [0.0459, 0.2451]	–
202609204	phase_late	-0.00016 [-0.00268, 0.00410]	0.1230 [0.0360, 0.2324]	–
202609204	fixed_random	-0.00797 [-0.02105, 0.00907]	0.0280 [0.0178, 0.0394]	–
202609205	episode_head	-0.00360 [-0.01577, 0.00781]	-0.0173 [-0.0350, -0.0002]	–
202609205	episode_tail	0.00768 [-0.00883, 0.02744]	-0.0697 [-0.0902, -0.0508]	–
202609205	even_episodes	-0.01335 [-0.05122, 0.02487]	-0.1275 [-0.1768, -0.0830]	–
202609205	odd_episodes	0.00028 [-0.01103, 0.01163]	-0.0206 [-0.0299, -0.0112]	–
202609205	phase_early	-0.00604 [-0.01893, 0.01015]	-0.0792 [-0.1015, -0.0608]	–
202609205	phase_middle	-0.00470 [-0.01403, 0.00406]	-0.2296 [-0.4530, -0.0527]	–
202609205	phase_late	-0.00186 [-0.01116, 0.00736]	-0.2584 [-0.4736, -0.0862]	–
202609205	fixed_random	-0.00311 [-0.01367, 0.00850]	-0.0563 [-0.0795, -0.0358]	–
202609206	episode_head	-0.03235 [-0.05356, -0.01008]	0.3468 [0.2793, 0.4212]	–
202609206	episode_tail	-0.00417 [-0.02234, 0.01597]	0.0791 [0.0592, 0.1032]	–
202609206	even_episodes	-0.01562 [-0.03250, 0.00346]	-0.0607 [-0.0891, -0.0379]	–
202609206	odd_episodes	-0.00908 [-0.04128, 0.02666]	-0.0537 [-0.0690, -0.0370]	–
202609206	phase_early	-0.11112 [-0.13497, -0.07962]	-0.0897 [-0.1270, -0.0554]	*
202609206	phase_middle	0.00241 [-0.00275, 0.00765]	-0.0863 [-0.0932, -0.0805]	–
202609206	phase_late	0.00103 [-0.00373, 0.00541]	-0.0885 [-0.0973, -0.0814]	–
202609206	fixed_random	-0.07644 [-0.11760, -0.03177]	0.0828 [0.0545, 0.1165]	–
202609207	episode_head	-0.04077 [-0.05243, -0.02775]	0.1883 [0.1785, 0.1981]	–
202609207	episode_tail	-0.03730 [-0.06230, -0.01272]	0.1702 [0.1609, 0.1796]	–
202609207	even_episodes	-0.00472 [-0.01832, 0.01112]	0.0050 [-0.0028, 0.0129]	–
202609207	odd_episodes	-0.00151 [-0.02351, 0.01480]	0.0026 [-0.0074, 0.0131]	–
202609207	phase_early	-0.02633 [-0.05492, 0.00960]	0.0491 [0.0380, 0.0598]	–
202609207	phase_middle	-0.01404 [-0.03461, 0.00363]	-0.0522 [-0.3606, 0.2448]	–
202609207	phase_late	-0.01544 [-0.03618, 0.00247]	-0.0314 [-0.3550, 0.2748]	–
202609207	fixed_random	-0.00405 [-0.02754, 0.02474]	-0.0669 [-0.0745, -0.0592]	–
202609208	episode_head	-0.03687 [-0.05818, -0.01883]	0.0966 [0.0809, 0.1129]	–
202609208	episode_tail	-0.01877 [-0.04978, 0.01288]	-0.0262 [-0.0695, 0.0153]	–

Policy	Rule	ΔRMSE [interval]	ΔJ [interval]	Intersection
202609208	even_episodes	-0.03377 [-0.05858, -0.00494]	0.0707 [0.0553, 0.0862]	–
202609208	odd_episodes	-0.01604 [-0.04767, 0.00961]	0.0364 [0.0107, 0.0649]	–
202609208	phase_early	-0.02453 [-0.04791, 0.01049]	0.0199 [-0.0017, 0.0460]	–
202609208	phase_middle	-0.01871 [-0.03057, -0.00847]	-0.3227 [-0.6298, -0.0475]	*
202609208	phase_late	-0.04079 [-0.06077, -0.02720]	-0.4236 [-0.8110, -0.0974]	*
202609208	fixed_random	-0.04759 [-0.06653, -0.02514]	0.1033 [0.0849, 0.1227]	–

Table 14 provides every interval; Table 13 preserves the policy-level denominators.

How stable the retained counts are to the resampling seed. Each adjusted endpoint is a percentile with about 3.9 expected draws in its tail, so a registered amendment repeated the identical resampling on 20 independent streams of 10,000 replicates. The retained improvement count ranges from 15 to 18 and the return-interval intersection from 4 to 5; 4 comparisons change their retained label between streams. Both registered predictions of that amendment, that the counts and the labels would reproduce, are therefore falsified. Among the comparisons retained in the registered stream, the adjusted upper endpoint lies between 0.00138 and 0.07962 below zero, with a median of 0.01355, so a single retained label covers very different numerical margins. More replicates would shrink this variation, which is a property of the estimator and not of the evidence: the retained fidelity episodes carry their own uncertainty, and no number of streams reduces it.

Output coordinates versus executed position commands. A separate post-outcome sensitivity retains the original output metric and excludes the gripper component for an active-position comparison. The installed Gymnasium-Robotics version 1.4.2 matches the archived fresh-policy environment version; its FetchReach implementation blocks the gripper and scales position commands. A direct probe of the installed action-transform function confirms invariance to the gripper component. This is not a historical source-byte attestation. The original historical RMSE is raw output; some outputs exceed action bounds. Active, ignored and position-command metrics therefore use environment clipping separately, without replacing that historical metric. The fresh-policy original metric already uses clipping. Of 22 original fresh-policy point reversals, 22 retain their signs on active components. Thus the ignored coordinate alone does not explain those comparisons. This establishes neither a causal mechanism nor new significance. The earlier panel-bootstrap counts concern the original metric, not the active-component metric. The active-component metric adds the reversal 202609205 odd_episodes (original +0.00028, active -0.00064, mean return change -0.0206): its original-output change is positive while its active change is negative. It also removes the fidelity improvement of 202609204 odd_episodes (original -0.00094, active +0.00038, mean return change +0.1092); 202609207 odd_episodes (original -0.00151, active +0.00053, mean return change +0.0026), neither of which lost return. These are near-tie comparisons whose changes are small against their own RMSE levels, so the count difference is a relabelling of borderline cases rather than a new effect. Statements about improved fidelity without beyond-tolerance deterioration are stated for the metric that established them, and do not transfer between metrics by default. The full per-comparison coordinate values, source digests, transform probe, and preserved raw-output diagnostic are indexed in Appendix I.4. No new rollout or policy selection is performed. Table 15 compares each metric on the same complete families.

Table 15: Complete-family coordinate sensitivity: improving fidelity / lower-mean-return point reversals. Position scaling preserves active-component ordering. Counts are descriptive, not independent replications.

Family	Original output	Active position	Ignored gripper	Position command
Historical compiler	8 / 5	8 / 5	7 / 4	8 / 5
Fresh policies	53 / 22	52 / 23	48 / 22	52 / 23

Fidelity measured on the states each artifact visits. Every fidelity quantity above is measured on a fixed state panel. Under feedback the artifact and the reference need not visit the same states, so a post-outcome sensitivity recomputes the same level-3-minus-level-0 contrast on each artifact's own

holdout trajectories, pooling squared action components over all 25,600 recorded steps per arm. No rollout is added: the observation and the applied action were already retained at every step. Before any visited state is used, the same forward path reproduces each stored panel reference exactly; a mismatch halts the analysis. On visited states the contrast improves fidelity in 41 of 64 comparisons, of which 10 also have lower mean return, against 53 and 22 on the fixed panel. The two measures disagree on the sign of the fidelity change in 20 comparisons, so which artifact is the more faithful one depends on where fidelity is evaluated. Within-policy rank agreement with return is closer on visited states (median Spearman -0.49 versus -0.15 on the panel) but reaches -0.8 or below in only 3 of 8 policies. For policy 202609205 the panel records 6 fidelity improvements and the visited states record 0, so that policy's panel improvements do not survive the change of measurement distribution. The manuscript now reports four fidelity definitions on these same 64 comparisons. Restricting the metric to the active components on the same visited states gives 43 improvements and 12 reversals. Counting how many comparisons each definition flags as improving fidelity with lower mean return: the fixed panel 22, the panel restricted to active components 23, visited states 10, visited states restricted to active components 12. 23 distinct comparisons are flagged by at least one definition and 10 by all four. The definitions therefore describe overlapping but different sets, and the separate sensitivities are not a cumulative verification of one set of reversals. Table 17 lists every flagged comparison and the definitions that flag it. Table 16 gives the per-policy counts and correlations. This metric is endogenous and is not a selector: a degraded artifact visits different states, and obtaining the quantity requires executing the artifact on the task, which is the evidence the protocol already requires. Fidelity and return come from the same episodes, so these are not independent samples, and no causal mechanism is established. The registered predictions that reversals persist and that fidelity does not order return within every policy both hold; original fixed-panel results are unchanged.

Table 16: Fixed-panel versus visited-state fidelity on the same comparisons. Changes are level 3 minus level 0, so a negative value improves fidelity; reversals additionally have a lower mean return. Spearman values relate the fidelity change to the return change within each policy. Counts are descriptive, not independent replications.

Policy	Improvements		Reversals		Spearman	
	Panel	Visited	Panel	Visited	Panel	Visited
202609201	6	4	1	0	-0.88	-0.98
202609202	7	8	5	5	0.76	-0.57
202609203	7	5	1	0	-0.12	-0.29
202609204	5	6	0	0	0.33	-0.19
202609205	6	0	6	0	0.29	-0.40
202609206	6	6	3	3	-0.19	-0.26
202609207	8	6	3	1	-0.74	-0.86
202609208	8	6	3	1	-0.36	-0.88
Total	53	41	22	10	—	—

The same executables compared on one state distribution. Comparing each executable on the states it visits changes the action function and the evaluation states together. A registered follow-up separates them by running both sealed executables of every policy–rule pair on both sampled distributions, 204,800 single-row device inferences in total, on the same accelerator and runtime as the original audit, whose recorded environment is unchanged. Nothing is recompiled. Each executable re-run on its own sampled states reproduces the actions the task run applied there byte for byte, which is what makes the cross terms comparable; a mismatch would have halted the analysis. Write M_{ij} for the mean squared post-processing action error of the level-i executable on the states sampled from the level-j distribution. Each M_{ij} pools squared errors over all 4 action components of the 800 states held for that arm, with no component dropped and no position scaling: the paper's active-component and command metrics are separate definitions, reported in Table 15. The reference is the full-precision policy's action on those same states, clipped to the registered $[-1, 1]$ bounds, and the artifact's value is the post-clip action the runtime applied. The states are an outcome-blind sample locked before any cross inference: each arm's 25,600 recorded holdout steps are sorted on $\mathrm{sha256}(\mathrm{policy}|\mathrm{rule}|\mathrm{level}|\mathrm{episode_seed}|\mathrm{step})$ and the first 800 kept, the same states for either

Table 17: Every comparison flagged as improving fidelity with lower mean return, by fidelity definition. A bullet marks the definitions that flag it. Point signs only; no significance or materiality is implied.

Policy	Rule	Panel	Panel active	Visited	Visited active
202609201	phase_late	•	•	—	—
202609202	even_episodes	•	•	•	•
202609202	odd_episodes	•	•	•	•
202609202	phase_early	•	•	•	•
202609202	phase_middle	•	•	•	•
202609202	phase_late	•	•	•	•
202609203	phase_late	•	•	—	—
202609205	episode_head	•	•	—	•
202609205	even_episodes	•	•	—	—
202609205	odd_episodes	—	•	—	—
202609205	phase_early	•	•	—	—
202609205	phase_middle	•	•	—	—
202609205	phase_late	•	•	—	—
202609205	fixed_random	•	•	—	—
202609206	even_episodes	•	•	•	•
202609206	odd_episodes	•	•	•	•
202609206	phase_early	•	•	•	•
202609207	phase_middle	•	•	—	•
202609207	phase_late	•	•	•	•
202609207	fixed_random	•	•	—	—
202609208	episode_tail	•	•	•	•
202609208	phase_middle	•	•	—	—
202609208	phase_late	•	•	—	—

executable evaluated on that distribution. The own-distribution difference splits as

$$M_{33} - M_{00} = (M_{30} - M_{00}) + (M_{33} - M_{30}),$$

where the first term changes the executable at a fixed distribution, the second changes the distribution for a fixed executable. This is an identity of mean squared errors and carries no causal claim about task loss. On the common level-0 distribution the level-3 executable is more faithful in 48 of 64 comparisons, and 17 of those have lower mean return, so improved fidelity coincides with lost return even when both executables are scored on identical states. The distribution term is the larger of the two in absolute value in 29 comparisons, so where each executable is evaluated matters. Counting which term is larger says nothing about the size of the total, because the two carry opposite signs and partly cancel in 32 of the 64 comparisons. The relative sizes of the two components vary across comparisons, and this decomposition does not quantify either one's contribution to lost task return. Within policies, the common-distribution fidelity change and the return change have a median Spearman correlation of -0.60, reaching -0.8 or below in 1 of 8 policies and turning positive in 1. Both registered predictions, that such reversals persist and that the common distribution does not order return within every policy, hold. These counts come from that locked sample rather than the full trajectories, so they are not the counts of the full-trajectory comparison, and fidelity and return still come from the same episodes. Every term here requires executing an artifact, so none of them is a pre-deployment screen. These reversal counts describe different measurement definitions, not successive validation stages of one reversal set. The 22 of the headline comparison are point signs on the fixed panel, each executable scored where the panel states lie. The 10 shared comparisons are those every one of the four fidelity definitions flags. The 17 here are those improving on the common level-0 distribution. They answer different questions rather than progressively confirming the same comparisons; the interval and resampling analyses describe uncertainty in the fixed-panel count alone.

Table 18: Fidelity change on the common level-0 distribution, by policy. Improvements and reversals count the level-3 executable scoring lower squared error than the level-0 executable on the same sampled states; reversals additionally have lower mean return. Spearman relates that change to the return change within the policy. Point signs only.

Policy	Improvements	Reversals	Spearman
202609201	4	1	-0.81
202609202	6	3	-0.60
202609203	8	1	-0.24
202609204	6	0	-0.60
202609205	4	4	0.40
202609206	6	3	-0.21
202609207	6	2	-0.62
202609208	8	3	-0.60
Total	48	17	—

What these comparisons establish. Across artifacts, scalar fidelity does not consistently order task return. Within the original compiler-intervention family, five statistically resolved reversals remain within practical tolerance. Across fresh policies, secondary point-sign reversals recur; the interval audit retains their original within-policy scope, without establishing material harm from improved fidelity. These comparisons retain separate statistical families and analysis roles.

C Policy Replication and Cross-Task Scope

C.1 Cross-policy calibration-support replication

E23 fixed a contrast taken from the complete single-policy calibration-family audit and an outcome-blind error-tail diagnostic, then locked it before any E23 backend task outcome. Eight fixed 50k-step SAC policies first passed the same outcome-blind FP32 capability rule. Each policy used disjoint 800-state calibration and measurement panels, three fixed calibration rules (phase-early, phase-middle, and fixed-random) on both backends, and 256 shared fresh reset seeds. Policies, artifacts, runtime hashes, the 50-step horizon, the Hailo phase-early-minus-phase-middle primary, and the no-extension rule were committed before execution.

Before task outcomes, two exact action dumps for all 48 policy/artifact cells showed that phase-middle had higher post-clip error p99 and a larger predeclared early-state tail fraction in all eight policies on each backend. This repeated the E22 error-tail pattern without using task labels and did not gate or alter the outcome family. Table 19 reports the resulting policy-level task outcomes.

Table 19: Cross-policy calibration-contrast replication. The primary contrast is Hailo phase-early minus phase-middle mean return; higher is better. Terminal success is descriptive. Each row contains 256 paired reset seeds, but each independently trained policy contributes one value to inference.

Policy seed	Early return	Middle return	Return contrast	Early success	Middle success
9510	-0.571	-6.800	6.229	1.000	0.285
9511	-0.951	-13.975	13.024	0.977	0.316
9512	-0.698	-17.306	16.608	0.980	0.027
9513	-0.693	-8.050	7.358	0.996	0.539
9514	-0.897	-0.859	-0.038	0.984	0.992
9515	-0.765	-0.911	0.146	0.992	1.000
9516	-0.989	-15.427	14.438	0.988	0.160
9517	-0.848	-11.844	10.996	1.000	0.352

The mean policy contrast is 8.595, with policy-bootstrap 95% CI [4.422, 12.511] and policy median 9.177 above the locked 0.5 practical threshold. An exhaustive two-sided sign-flip test gives $p = 0.015625$ under policy-level sign exchangeability. Because that assumption is stronger than a test of signs alone, we also report the two-sided sign-test sensitivity: seven of eight contrasts are positive, giving $p = 0.0703$; six exceed 0.5. Pooled descriptive terminal success is 0.990 versus 0.459. The same contrast is positive for EdgeTPU in all eight policies but smaller (mean 0.622, median 0.342); Hailo phase-early versus fixed-random is only 0.031 on average. E23 therefore replicates one calibration-support consequence across policies. It does not establish that every policy is affected, isolate calibration support as the sole cause, validate CLAE, or estimate prevalence beyond this task, network class, and training recipe.

C.2 Calibration-support dose response and held-out replication

E23 showed that one calibration contrast changes task return across policies but did not say what the contrast changes. E25 walks between the two rules at a fixed 256-row budget, drawing λ of the rows from the early window and the rest from the middle window, on four E23 policies: the two with the largest E23 contrast (9512, 9516) and the two with the smallest (9514, 9515). Four policies cannot support pooled significance—an exact two-sided sign-flip test over four units bottoms out at $p = 0.125$—so this is a mechanism probe with reset seeds as the unit and the two groups compared descriptively.

Every choice was fixed in two stages. Stage A built the dose panels, compiled 40 artifacts, and recorded outcome-blind diagnostics; stage B authorized the task outcomes only after those diagnostics were committed. The grid extends the locked E23 family rather than replacing it: $\lambda = 1$ and $\lambda = 0$ reproduce the E23 phase-early and phase-middle selections, and their action dumps match the stored E23 dumps exactly in all 16 checked cells, so the endpoints are the controllers E23 measured.

Table 20: Hailo mean return by calibration mixture, 256 fresh paired seeds per cell. The step column is the mean over $\lambda \geq 0.25$ minus the $\lambda = 0$ value. Groups are labelled by their prior early–middle contrast, not by these outcomes.

Policy	Group	λ=0	0.25	0.5	0.75	1	Step
9512	large contrast	−17.022	−0.762	−0.725	−0.695	−0.729	+16.29
9516	large contrast	−16.987	−0.861	−0.984	−0.983	−0.975	+16.04
9514	small contrast	−0.892	−1.133	−0.972	−0.913	−0.939	−0.10
9515	small contrast	−0.954	−0.869	−0.803	−0.780	−0.795	+0.14

Table 20 gives the grid. The registered rule required a Spearman correlation of at least 0.9 between λ and mean return in both large-contrast policies, and a range within the 0.5 practical tolerance in both small-contrast policies. The second half passes, at 0.241 and 0.174. The first misses, at 0.700 and 0.400, so the family misses.

It misses for a reason recorded in the stage B lock before any outcome was read. The stage A error field is already a step: Hailo post-clip error p99 falls from 2.264 and 1.868 at $\lambda = 0$ to 0.22–0.35 at every larger mixture, and the early-state tail fraction falls from 0.66–0.69 to exactly zero at $\lambda = 0.25$ and stays there. The lock predicted that return would follow the same shape and that a monotonicity rule would then fail even for a real and large effect, because the ordering of the four upper mixtures is set by noise. That is what happened. The rule was not modified.

Three results survive the miss. The step separates the groups by more than two orders of magnitude on a seed block disjoint from E23. The small-contrast policies are flat at every mixture although their $\lambda = 0$ error tail is substantial—p99 0.776 and 0.860, early-state tail fraction 0.484 and 0.328—so tail presence does not explain collapse while tail magnitude tracks it, which is the gap E23 left open. And the same calibration defect costs 16.3 return on Hailo but 0.30 on EdgeTPU for policy 9512, so the consequence is a property of the artifact rather than of the calibration rule alone.

Held-out replication. The four policies above were the two largest and two smallest E23 contrasts, so their behaviour was partly anticipated. A second family, locked separately because E25 forbids adding policies after its outcomes were read, runs the same grid on the four E23 policies E25 did not use. All four exceed the practical threshold in E23, so this family has no flat arm and tests correspondence rather than presence.

Its endpoint is the step size, not a monotone trend: E25 had already shown that a monotonicity rule tests the shape assumption rather than the mechanism. The rule required every step to clear the 0.5 threshold and the step to rank with the E23 contrast at Spearman 0.8 or better.

Table 21: Held-out replication on the four policies not used above. Steps are the mean over $\lambda \geq 0.25$ minus the $\lambda = 0$ value on Hailo. The prediction column is the prior early–middle contrast, registered before the outcomes.

Policy	Predicted	Observed step	Ratio	EdgeTPU step
9510	6.23	6.71	1.08	2.55
9513	7.36	8.24	1.12	0.27
9517	11.00	12.96	1.18	0.35
9511	13.02	15.05	1.16	0.60

Both conditions pass (Table 21). Every step clears the threshold, and the step ranks with the E23 contrast at Spearman 1.000. The locked rule is that threshold, not a p-value; for reference, one of the 24 orderings of four policies is perfect, so the one-sided exact permutation probability of a perfect *positive* ordering is $1/24 = 0.042$, and the direction was fixed in the lock. A two-sided reading would admit the perfect negative ordering as well and give $2/24 = 0.083$. The registered point prediction also holds: observed steps are 1.08 to 1.18 times the predicted values, consistently slightly larger and none off by more than a fifth.

One further reading was registered before those outcomes. The $\lambda = 0$ error p99 ranks against the E23 contrast at Spearman 0.8, with 9513 and 9517 transposed, and the lock stated that a return step

ranking better than the tail would show the tail is not the whole mechanism. The step ranks at 1.000 and does not fail where the tail fails, so it is not.

Together the two families cover all eight policies on one task, network class, and training recipe. They do not estimate prevalence, isolate calibration support as the sole cause, or establish a calibration ranking that transfers.

C.3 Fresh full-episode FetchPush task outcome

We separately locked a full-episode total-artifact audit for the fixed FetchPush policy. A TQC+HER teacher (Kuznetsov et al., 2020; Andrychowicz et al., 2017) was distilled to a 256–256 ReLU behavior-cloning student. Before backend outcomes, a disjoint 128-seed capability block gave FP32 terminal/ever success 0.703/0.797, versus 0.094/0.094 under zero action. We then ran FP32 and both online artifacts for 50 steps on 512 fresh paired reset seeds. Success means achieved-goal distance at most 0.05 m. Dense return, the sole primary endpoint, equals the negative sum of post-step goal distances to numerical precision. Table 22 reports the complete paired family.

Table 22: Fresh full-episode FetchPush total-artifact audit. Return degradation is FP32 minus backend, so positive values are adverse; the two rows form one Holm family. Success entries are FP32/backend. Distance is backend minus FP32. Secondary task endpoints do not rescue the return family.

Backend	FP32/backend return	Return degradation [95% CI]	Holm p	Terminal success	Ever success	Final-distance Δ, m [95% CI]
EdgeTPU	$-4.333/-4.579$	0.246 [-0.109, 0.605]	0.186	0.732/0.676	0.809/0.771	0.0044 [-0.0075, 0.0162]
Hailo	$-4.333/-5.211$	0.878 [0.538, 1.232]	< 0.001	0.732/0.590	0.809/0.666	0.0158 [0.0047, 0.0271]

Hailo's locked marginal interval exceeds the 0.5-return practical threshold, but the conservative post-outcome simultaneous band makes its main-text audit status indeterminate (Figure 2). Terminal and ever success losses are both 0.143, with 95% intervals [0.092, 0.193] and [0.096, 0.189]. EdgeTPU return is inconclusive and its final-distance interval crosses zero, although its secondary terminal-success loss is 0.057 [0.006, 0.107]. The result adds actual task-performance evidence for one contact-manipulation policy and shows backend dependence. It neither isolates error direction nor validates CLAE as a full-episode selector.

Calibration-content breadth. A separately locked family held the FetchPush behavior-cloning policy, backend path, artifact count, and 512 new paired starts fixed while changing only among phase-early, phase-middle, and fixed-random calibration rules. FP32 mean return was -4.504 and terminal success was 0.672. Table 23 reports the two pre-specified early-minus-middle primary rows; fixed-random is shown to expose the complete three-rule family.

Table 23: FetchPush calibration-content audit. Return contrast is phase-early minus phase-middle; positive favors early. The two backend rows form one Holm family. Hailo uses the CPU path's automatic level-0 fallback.

Backend	Early return	Middle return	Fixed-random return	Early $-$ middle [95% CI]	Holm p	Early/middle success
EdgeTPU	-4.373	-4.431	-4.622	0.058 [-0.295, 0.406]	0.746	0.711/0.711
Hailo	-4.979	-6.687	-5.424	1.708 [1.259, 2.151]	< 0.001	0.670/0.504

Only Hailo clears both inference and the fixed 0.5-return practical threshold. This is second-task and second-architecture breadth for one policy on one tested path, not a universal calibration ranking or causal vendor comparison.

C.4 Fresh full-episode FetchReach task replication

We prospectively repeated the same total-artifact protocol on the fixed FetchReach SAC policy. A disjoint 128-seed capability block gave FP32 terminal/ever success 1.000/1.000, versus 0.031/0.031 under zero action. The outcome block used 512 new reset seeds, both online artifacts, a 50-step cap, and dense return as the sole primary endpoint. E17 and E18 were not pooled after observing their results. Table 24 reports the complete paired family.

Table 24: Fresh full-episode FetchReach total-artifact replication. Conventions and multiplicity match Table 22. EdgeTPU is detected but not certified practically large because its interval crosses the locked 0.5-return threshold; Hailo's full interval exceeds it.

Backend	FP32/backend return	Return degradation [95% CI]	Holm p	Terminal success	Ever success	Final-distance Δ, m [95% CI]
EdgeTPU	$-0.701/-1.136$	0.435 [0.350, 0.528]	< 0.001	1.000/0.885	1.000/1.000	0.0152 [0.0121, 0.0186]
Hailo	$-0.701/-2.254$	1.553 [1.395, 1.718]	< 0.001	1.000/0.654	1.000/0.816	0.0363 [0.0322, 0.0406]

Terminal-success losses are 0.115 [0.088, 0.143] and 0.346 [0.307, 0.387]; Hailo ever-success loss is 0.184 [0.150, 0.217]. Both artifacts therefore reduce actual reaching-task performance for this one capability-qualified policy, with a larger Hailo effect. This is total-artifact evidence: it does not identify error direction as the cause or validate CLAE as a full-episode selector.

C.5 Unforced Ant Task Evidence

Unforced online task-return audit. E14 used the same fixed policy and artifact hashes but a disjoint block of 320 fresh reset seeds (81101–81420), no external force, and a common 1,000-step cap. The sole primary endpoint was paired undiscounted task return, $R_{\mathrm{FP32}} - R_{\mathrm{backend}}$, with a two-backend Holm family locked before the outcomes were opened. Both rows pass the positive-interval criterion and the additional requirement that the complete interval exceed 250 return units (Table 25).

Table 25: Ant task-return audit without constructed perturbations. Positive return and progress deltas mean lower backend performance. Return is the sole primary endpoint; progress and termination decompose the same effect. This single-policy contrast does not isolate CLAE or error direction.

Backend	Return FP32/back.	Return Δ [95% CI]	Δ/FP32	Holm p	Progress Δ m [95% CI]	Term. FP32/back.
EdgeTPU	5539/4872	667 [526, 802]	12.0%	< 0.001	32.6 [25.8, 39.6]	0.059/0.097
Hailo	5539/4908	631 [514, 749]	11.4%	< 0.001	31.4 [25.5, 37.4]	0.059/0.053

The forward-reward components account for most of the return differences: EdgeTPU and Hailo lose 653 [514, 793] and 627 [512, 747] units, respectively. Episode-length intervals include zero for both artifacts, Hailo terminates slightly less often than FP32, and the return difference persists among pairs where both arms reach the common cap. Thus, different episode lengths do not explain the return difference. Thus, these two fixed artifacts change task performance for this selected policy, while the links from CLAE or error direction alone to task outcome remain unresolved.

C.6 Family-adjusted task-audit bounds

The procedure and its post-outcome role are specified in Appendix A.3.

Table 26: Marginal and family-adjusted intervals from the same replicates. Widening the band for the two-backend family changes one status of six.

Audit	Backend	δ_T	Marginal	Status	Family-adjusted	Status
Ant	EdgeTPU	250	[531, 804]	materially degraded	[512, 823]	materially degraded
Ant	Hailo	250	[513, 748]	materially degraded	[497, 764]	materially degraded
FetchPush	EdgeTPU	0.5	$[-0.11, 0.61]$	indeterminate	$[-0.17, 0.66]$	indeterminate
FetchPush	Hailo	0.5	[0.53, 1.23]	materially degraded	[0.48, 1.28]	**indeterminate**
FetchReach	EdgeTPU	0.5	[0.35, 0.52]	indeterminate	[0.34, 0.53]	indeterminate
FetchReach	Hailo	0.5	[1.39, 1.72]	materially degraded	[1.37, 1.74]	materially degraded

Table 26 gives both columns, and FetchPush Hailo is the one cell that moves. Its marginal lower bound clears the 0.5 tolerance and its family-adjusted lower bound does not, so under the stricter reading the audit is indeterminate rather than materially degraded. The other five statuses are unchanged, and no cell moves in the direction of a stronger claim.

The locked 250-return Ant contract remains the reported decision threshold; a post-outcome sensitivity audit only asks whether that choice drives the label. At thresholds 150, 200, 250, 300, and

350, both backend rows remain materially degraded because their family-adjusted lower bounds are 512 and 497 return units. Thus the observed Ant label is insensitive over this range, without turning any member of the grid into a universal Ant threshold.

The two calibration-family ranges are a second family the rule reads together, so they get their own band from the same construction. Reset seeds are shared across all sixteen artifacts, so one resample drives both backends, and the estimand is each backend's best-to-worst mean-return range, which is selection-aware by construction. The EdgeTPU range is 0.489, marginal [0.459, 0.519] and joint [0.455, 0.523]; the Hailo range is 14.926, marginal [14.300, 15.553] and joint [14.209, 15.644]. Neither status moves: the EdgeTPU range remains indeterminate heterogeneity and the Hailo range materially heterogeneous.

Two caveats. These intervals are independent recomputations rather than copies of the locked intervals. Relative to the main tables, marginal endpoints move by at most 5 return units for Ant and 0.008 for the Fetch audits; the calibration-range endpoints move by at most 0.018. The comparison between marginal and family-adjusted columns here uses identical replicates and is the point of the table. And a family here is the two backends of one audit, or the two calibration ranges, not the three audits together, because the audits use different tasks, tolerances, and seed blocks and were locked separately.

C.7 Cross-task Ant extension: four locked attempts

E23, E25, and E26 all use FetchReach SAC policies from one recipe, so the calibration evidence rests on one task family. E27 was locked to ask whether the pure-middle calibration step appears in a locomotion task with a different action dimension, observation dimension, and episode length. It never reached that question.

The lock fixed eight policy seeds, an FP32 capability screen of 128 held-out seeds from a block disjoint from every calibration and outcome block, a 1,000-step episode cap, the five gates of Table 27, and a requirement that at least six policies qualify, with failing policies reported rather than replaced. It also fixed the order: capability first, then panel and artifact hashes, then a separate outcome lock.

The six-policy minimum matches the E23 family this extension mirrors, whose locked primary inference is an exact two-sided sign-flip test over policies. Six is the smallest cohort that can reach $p < 0.05$ there: five units bottom out at 0.0625 and six at 0.03125. The E27 lock fixed the number before training and never changed it, though it records the number without this derivation.

The work divides into two phases, and the boundary matters. In an outcome-blind development phase we found and corrected a recipe defect and evaluated the training budget, using FP32-only pilot blocks disjoint from the capability block. In the confirmatory phase the recipe, the eight policy seeds, and the capability block were run against gates and a cohort minimum that had been fixed before any policy was trained and were never revised. No backend artifact existed at any point in either phase, so "prospectively locked" is exact with respect to artifact outcomes while the capability recipe itself went through development.

Both development rounds were disclosed in the lock before the screen ran. The first omitted `buffer_size` from the recipe, so training used the script default of 10^5 rather than the 10^6 the reference policy Ant-73102 used at the same budget and architecture; a budget pilot on a disjoint seed block showed native termination of 0.281–0.672 against a 0.20 gate for all seven policies then trained. The second matched the reference recipe and reached 0.062–0.078 on four policies, at parity with the reference policy's 0.059, but four stalled. Raising the budget from 10^6 to 3×10^6 steps then raised return without improving the binding gate: five of eight policies terminated *more* often at the larger budget, so the return and termination gates trade off and a further increase was not indicated. No gate was relaxed and no seed was added.

The screen qualified four policies against a required six (Table 27). Native termination rate was binding throughout: the three near misses clear every other gate and fail only there, while 9616 fails broadly. The zero-action arm returns 994.5. A 64-episode pilot on a different seed block had predicted the same four qualifiers, and 9611 landed on 0.203 in both. A descriptive boundary sensitivity does not change the stop: counting 9611's 0.203 as qualified would give five of eight, still below the locked minimum of six.

Table 27: Initial Ant capability screen, all five gates. All gates, thresholds, policy seeds, and the six-of-eight cohort rule were fixed before the 128-seed confirmatory capability block and were never revised. Thresholds: mean return $\geq$ 3000, advantage over a zero-action arm $\geq$ 2000, median episode length $\geq$ 900, native termination rate $\leq$ 0.20, non-finite episodes $= 0$. Bold marks a failed gate.

Policy	Mean return	Zero-action adv.	Median length	Native term.	Non-finite	Status
9610	5745	4751	1,000	0.125	0	qualified
9611	4865	3870	1,000	**0.203**	0	stopped
9612	5711	4716	1,000	0.172	0	qualified
9613	4059	3065	1,000	**0.461**	0	stopped
9614	4278	3284	1,000	**0.297**	0	stopped
9615	5792	4797	1,000	0.141	0	qualified
9616	**2114**	**1119**	**522**	**0.664**	0	stopped
9617	5739	4745	1,000	0.133	0	qualified

The cause was not the budget. E27 trained on the default Ant observation, which includes contact forces, while every other Ant experiment in this paper fixes the 27-dimensional variant that excludes them. We report the screen as it ran rather than replacing it, because the corrections are separate locked experiments.

Three further attempts. A second family compiled artifacts from E27's four survivors and read outcomes. Every artifact collapsed at every calibration dose—FP32 returns 5726–5900 against INT8 arms at 17–3015—so its locked null measured a floor rather than the absence of the mechanism. The diagnosis is the same mis-specified observation: 54 of its 105 coordinates are identically zero and a few reach ± 16.8, so per-tensor integer input quantization takes its scale from the outliers and leaves a typical informative coordinate about 16 of 255 levels. We added an outcome-blind gate on exactly that quantity, requiring at least 32 levels, which the paper's working Ant artifact clears at 82.4.

A third family fixed the environment and qualified six of eight at native termination 0.008–0.148 with the reference recipe unchanged at 10^6 steps, so the budget never was the constraint. It then stopped at the new gate: five of eighteen cells fell below 32. A fourth added a conditioning pre-screen to policy selection, computed on the FP32 calibration panel alone and conservative against the compiled gate on eighteen cells, trained sixteen policies, and took the first six by seed among those clearing both screens. Its artifacts pass at 35.1–57.5 levels and all thirty-six arms repeat exactly.

Outcome of the fourth family. It misses. The locked rule required an exact two-sided sign-flip $p \leq 0.05$ over the six per-policy Hailo steps and a median step above the 250-return tolerance; the observed p is 0.1875 and the median step is -443. Four of six steps are negative, so pure middle-phase calibration outperformed the mixtures, opposite to FetchReach. The pattern beneath is heterogeneous rather than a reversed calibration effect: the half-and-half mixture is the worst arm for two policies whose pure doses are similar, two fall monotonically with the early-phase fraction, one is flat, and one policy's artifacts all return negative reward at zero termination and full-length episodes, which is a different failure from falling.

What the family does establish is a prospective failure of a locally measured predictor. Before any outcome we recorded that the Hailo post-clip error tail separated the calibration bands by $+0.206$ on average, roughly a third of the FetchReach separation, and predicted a small positive step below tolerance. The steps are large and mostly negative, so the prediction was wrong in sign and magnitude. The same tail tracked the FetchReach step closely, which is what makes the failure informative rather than merely a null.

These four families do not establish whether the calibration cliff transfers to locomotion, and the last one describes only Ant policies whose observation geometry clears the pre-screen. Two families stopped prospectively, at the capability and effective-level gates. A third reached outcomes but the INT8 floor made its intended contrast uninterpretable. The fourth reached outcomes and missed its rule, with a prewritten predictor wrong in sign and magnitude.

C.8 CALIBRATION GEOMETRY IN ANT POLICIES

Four locked Ant families left the cross-task scope question unanswered, and two of them died on quantization headroom rather than on the calibration question. CAL13 therefore holds three things fixed instead of two. Its cells are E30's six outcome policies, three anchors each, every anchor inside a narrow span-coverage band and clearing the locked effective-levels floor of 32, so coverage and headroom are both controlled while the subset centre moves. The build specification is byte-identical to CAL11's and CAL12's, and the rollout is the committed E14 Ant runner on E30's 256 reset seeds with a 1,000-step cap. No FP32 arm was run: the Ant policy checkpoints are not recoverable, and a within-policy range over INT8 cells does not need one. This does not mean every source representation is missing: the six source ONNX actors, eighteen calibration arrays and executed HEFs, and actor/runner I/O contracts remain available with digest checks. Analysis replay is verified; task replay and recompilation have retained inputs but are not re-executed in this provenance audit. Original-checkpoint training or FP32-framework replay is unavailable, and byte-identical recompilation is not promised (source-availability record, Appendix I.4).

Stage B required an outcome-blind tail-separation check. Earlier FetchReach audits associated task degradation with post-clipping error tails, motivating a check that the matched construction retained sufficient tail separation. The median within-policy p99 tail separation is 0.734, against 0.330 and 0.441 for the same statistic on CAL12's bands and 0.206 for the dose-axis family that missed, so the gate passed and Stage B was executed.

Table 28: Ant calibration-geometry results by policy. The statistic is the observed within-policy best-to-worst mean-return range minus the median range under a permutation null that shuffles episode-level cell labels. The last column expresses the reused 250-return yardstick as a percentage of that policy's own FP32 mean.

Policy	Observed range	Null median	Statistic	Range / FP32	250 / FP32
9642	622.8	38.2	584.6	12.36%	4.96%
9644	497.5	29.7	467.8	10.43%	5.24%
9648	235.6	15.3	220.2	4.49%	4.76%
9641	232.9	35.4	197.5	4.68%	5.02%
9645	155.6	22.4	133.2	4.39%	7.05%
9643	114.7	21.8	92.9	2.29%	5.00%

Table 28 gives the six policies. **The pre-registered rule misses.** It required an exhaustive two-sided sign-flip at or below 0.05 over the six per-policy statistics and a median observed range above the yardstick. The sign-flip gives 0.03125, its floor at six units, but the median observed range is 234.2 against 250, so the family misses on materiality.

The miss has the opposite character to the dose-axis family's on the same six policies. There, four of six steps were negative, the sign-flip gave 0.1875, and a prewritten error-tail prediction was wrong in sign and magnitude. Here all six statistics share a sign, every observed range exceeds its permutation-null median of 15.3 to 38.2, and the observed median misses by 15.8 return units. Positive excess range does not identify an early–middle calibration preference. The gate miss is not a population upper bound or equivalence result.

The yardstick is the weak link, and the lock said so before the outcome. The 250 value was derived for Ant-73102 from a two-row Holm design MDE of 227–231 and 4.5% of that policy's FP32 mean; it is reused here, not re-derived. Because the six policies have different FP32 means it is not a constant relative standard: 4.76% of seed-9648's mean but 7.05% of seed-9645's. We do not re-derive it after the fact, since the lock demoted the value without pre-specifying a replacement and choosing one now would select a threshold from the result. The FP32-relative column is reported instead; the task-audit decision rule in Section 3 likewise distinguishes absolute tolerances from relative losses.

CAL13 uses three cells per policy and deliberate constructions rather than natural sampling rules. Its observed median fails the registered gate on these cells; it does not bound the population effect or show that calibration choice is immaterial for Ant.

Null-assumption audit (post-outcome sensitivity). The registered two-sided sign-flip result is retained, but subtracting a null median does not make a range statistic symmetric about zero. Exhaustive sign enumeration alone does not establish null invariance (Hemerik & Goeman, 2018). A separate two-sided sign-only test yields 0.03125 conditional on six independent, nonzero fair signs; those assumptions are not consequences of median centering. We additionally construct a cohort-level null directly: independently permute the three cell labels within each paired episode and policy, then average policy ranges minus the fixed registered null medians. This assumes within-seed cell-label exchangeability, stronger than equal means. With 10,000 permutations and a plus-one correction, the two-sided equal-tail value is $2/10{,}001 = 0.000200$ for each cohort separately; neither has an upper-tail exceedance. This Monte Carlo resolution is not an exact exhaustive probability. The sensitivity neither validates the original symmetry assumption nor changes either cohort's materiality miss, and does not pool the cohorts. The separate assumptions, seeds, source hashes and results are in the Ant null-transformation record (Appendix I.4).

C.9 Held-out Ant calibration-geometry replication

CAL14 repeats the CAL13 construction on the six E30-qualified Ant policies whose calibration-geometry task outcomes had never been read. Each policy again has three cells matched on span coverage and effective integer levels, evaluated on 256 reset seeds with the same 1,000-step runner: 18 artifacts and 4,608 episodes. Before compilation, the lock derives a 219.97-return materiality yardstick for this cohort by applying the paper's existing $250/5539$ relation to each policy's recorded FP32 mean and taking the median. The threshold applies only to CAL14; it is not retrofitted to CAL13.

Table 29: Held-out Ant calibration-geometry results. The statistic is the observed within-policy range minus the median label-permutation range. The primary rule requires both an exhaustive two-sided sign-flip $p \leq 0.05$ and a median observed range above the pre-locked 219.97 yardstick.

Policy	Observed range	Null median	Statistic
9649	180.6	21.2	159.3
9650	738.5	40.1	698.4
9651	60.2	10.3	49.8
9652	480.9	27.9	453.0
9654	111.3	22.5	88.8
9655	232.3	16.4	216.0

All six statistics in Table 29 are positive, so the sign-flip test reaches its six-unit floor, $p = 0.03125$. The primary rule still misses: the median observed range is 206.44, below 219.97. Thus CAL14 gives a held-out positive excess-range replication but not a materiality pass or a replicated calibration preference. CAL13 and CAL14 have nearly identical range-to-threshold ratios, 0.937 and 0.938, despite using a borrowed threshold in the first cohort and a prospectively derived threshold in the second. Descriptively the disjoint cohorts are each 6/6 positive in excess range; they were not designed to be pooled, and we perform no post-hoc pooled test.

Policy influence (post-outcome sensitivity). The magnitude of the mean-range sensitivity is heterogeneous. Omitting 9650 changes the mean observed range from 300.630 to 213.050, and the mean excess from 277.555 to 193.386. Nevertheless, every leave-one-policy-out permutation calculation retains zero upper-tail exceedances and the same 0.000200 Monte Carlo resolution (Table 30). The small permutation value is therefore not dependent on retaining the largest-range policy in these calculations. All remaining signs stay positive; the two-sided sign-only value becomes 0.0625 for each five-policy subset, conditional on independent fair signs. These are distinct tests of distinct null structures, not interchangeable significance claims. The registered 219.97 yardstick applies to the full-cohort median, not these subset means, and its original miss is unchanged.

The construction-side Stage-A mechanism statistic does not explain that agreement. It is 0.7336 for CAL13 and 0.1930 for CAL14, almost fourfold apart, while their task range-to-threshold ratios differ by only 0.001. This is another failure of a pre-compile statistic to predict task consequence, not evidence that the statistic is useless for every construction.

Table 30: Post-outcome policy influence in the held-out Ant cohort. Each subset reuses the existing within-episode cell-label permutation stream and fixed policy-specific null centers, renormalized over retained policies. No policy is excluded from the registered result. Equal-tail permutation values include the plus-one correction; sign-only values additionally assume independent fair signs.

Omitted policy	Mean range	Mean excess	Permutation p	Sign-only p
None	300.630	277.555	0.000200	0.03125
9649	324.645	301.201	0.000200	0.06250
9650	213.050	193.386	0.000200	0.06250
9651	348.723	323.097	0.000200	0.06250
9652	264.572	242.464	0.000200	0.06250
9654	338.498	315.308	0.000200	0.06250
9655	314.294	289.874	0.000200	0.06250

Two analyzer fields missing from the Stage-B configuration were added after the episode file existed. They supply only descriptive FP32 means and the recorded CAL13 comparison string; neither enters the sign-flip, median range, threshold, or primary verdict. Re-running the integrity step under the amended lock reused every shard and reproduced the episode table byte-for-byte. As in CAL13, the three cells are deliberate constructions on a conditioned policy cohort, no FP32 arm is rerun, and the miss does not establish a general Ant or locomotion law.

D Compiler Paths and Calibration Construction

D.1 Compiler robustness summary

The Hailo spread might be specific to the CPU-only level-0 build path. Outcome-informed complete-family CPU reruns retain large ranges at explicitly requested levels 1–3: 14.567, 14.782, and 14.561, versus 14.926 at level 0 (Appendix D.3.1). Using all 256 rows for quantizer statistics gives a range of 14.969 at level 0; using all 256 for both statistics and AdaRound gives 14.420 at level 3 (Appendix D.4). A separately locked GPU-visible task bridge tests the same phase-early/phase-middle contrast in two separately compiled production-default level-3 pairs on fresh starts. Their return contrasts are 14.670 and 14.675, and both CI lower bounds (13.808 and 13.811) exceed the locked 0.5 threshold. The GPU task bridge is restricted to selected pairs, not the complete family. These checks are not independent policy replications.

D.2 Effective calibration-subset audit

We next hold the calibration states fixed and vary their order. On the tested CPU Hailo DFC 3.33.1 level-0 path, builds supplied with all 256 rows match their same-order explicit first-64 controls in measured actions and task outcomes, for both original and reversed order. In this selected audit, reversal raises terminal success from 69.9% to 97.1% and improves return by 1.232 [1.075, 1.392]. We infer an *effective first-64 subset* from these behavioral identities, not from compiler internals (Appendix D.2).

The selected order effect persists when level 1 is explicitly requested in a locked audit. A separate prospective audit tests three policy–multiset combinations; only one meets the registered task-effect criterion, so the family fails its two-of-three rule. Yet full-256 and same-order first-64 builds match in paired return and terminal success in all six controls. Across the level-0/1 audits, all ten full-versus-first-64 comparisons match on their tested endpoints. This supports repeatable subset dependence, while the task effect varies with the multiset and policy (Figure 5).

On the tested Hailo DFC path, statistics and learned stages use different row budgets. At levels 2–3, learned stages use all supplied rows; full-256 and first-64 builds differ on the fixed SDK-emulator panel across tested GPU-visible/hidden builds. The quantizer-statistics budget `calibset_size` defaults to 64. At level 0, Hailo's eight-rule mean returns span 14.926 for the historical stride-4 64-row subset, 6.732 for the default head 64, and 14.969 for all 256 rows. The full-set range is 0.3% above the historical range; count and subset change together. The two 64-row subsets yield ranges differing by more than twofold. A separate three-rule 64-row EdgeTPU control has range 0.475. Row count alone does not explain the observed task spread (Appendices D.3.2 and D.4).

Span coverage averages fractions of reference-panel coordinate ranges covered by calibration states. For one FetchReach policy on this CPU level-0 path, a prospectively locked eight-set ladder correlates with mean return ($\rho = 0.905$, exact $p = 0.0046$; range 6.5225), but also changes the state-selection anchor. A locked two-band follow-up varies subset centres. Low-band (width 0.0184) mean returns span 9.547 [9.245, 9.844]. Band means differ by 7.201 [7.024, 7.379], distinct from within-band ranges. Radius and density also vary, so the comparison does not isolate a geometric cause. This score alone cannot justify ranking or approval (Appendix D.5).

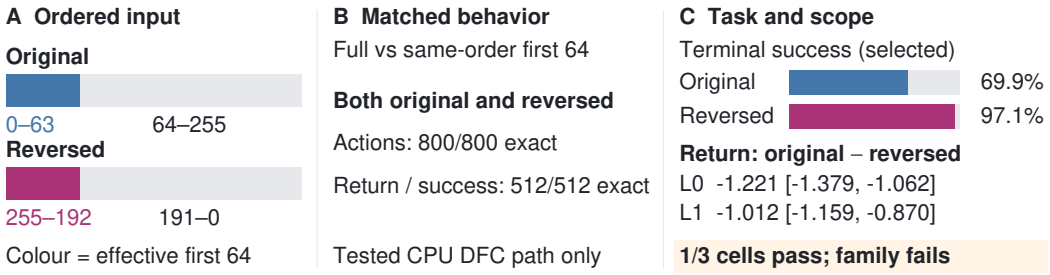

Figure 5: Compiler-path audit. **A:** Same states, reversed order. **B:** Full versus same-order first-64: exact actions on 800 states and return/success on 512 paired episodes. **C:** Success: selected 512-start order-reversal block. Return (original minus reverse; 95% CIs): disjoint 512-start level-0/1 block. Ten identity comparisons: two fixed-panel pairs, two selected-block task pairs, six breadth controls. The prospective breadth family fails (Appendix D.2).

An outcome-blind compiler-path probe asked which rows effectively determine an artifact when the tested CPU Hailo DFC 3.33.1 path receives 256 ordered FetchReach states. All optimizer runs reported 64 entries. On an 800-state hardware panel, the full-array executable exactly matched the executable compiled from rows 0–63; shuffling rows 64–255 changed nothing. Reversing the full array instead exactly matched the executable compiled from original rows 255–192. These checks fixed the hypothesis before task outcomes were authorized.

Calibration-count control. A separate fresh three-rule audit matched EdgeTPU to the 64 entries reported by the tested Hailo path while retaining the original 256-entry EdgeTPU family. The locked primary estimand—the 64-minus-256 change in the EdgeTPU early-minus-middle contrast—is -0.0368 [-0.0660, -0.0114], opposite the registered positive direction and far below 0.5. The supporting fixed-family ranges in Table 31 show that count matching does not explain the selected path-specific spread; they do not show that count never matters or isolate a vendor effect.

Table 31: Fresh three-rule calibration-count control on a separate 512-seed paired block (not the 1,024-seed eight-artifact family). Ranges are selection-aware best-minus-worst return within each fixed three-rule family.

Path/family	Entries	Return range [95% CI]	Role
EdgeTPU original	256	0.4886 [0.4408, 0.5378]	reference count
EdgeTPU matched	64	0.4753 [0.4261, 0.5238]	count control
Hailo tested path	64	14.7235 [13.8332, 15.6109]	supporting path row

The subsequent lock used one 256-state multiset, its exact reversal, both explicit first-64 subsets, and FP32 on fresh reset seeds 134101–134612. The single primary estimand was full-original minus full-reverse paired return; the two-sided 95% bootstrap interval had to exclude zero and the absolute estimate had to exceed 0.5. Neither the conditions nor the seed block could be extended.

Table 32: Locked same-multiset DFC audit on 512 paired FetchReach starts. The full-array rows exactly reproduce their corresponding explicit first-64 rows for every episode.

Condition	Mean return	Terminal success
FP32	-0.711	100.0%
Full original 256	-2.067	69.9%
Explicit original first 64	-2.067	69.9%
Full reversed 256	-0.835	97.1%
Explicit reversed first 64	-0.835	97.1%

Table 32 shows that original-minus-reverse return is -1.232 [-1.392, -1.075], so the locked criterion passes; terminal-success difference is -0.272 [-0.315, -0.231]. Both full-versus-explicit

controls have zero nonzero return pairs and zero discordant success pairs. This associates the task effect with the path's measured full-array/first-64 equivalence. It does not show that order matters when every supplied row affects the artifact or estimate prevalence over calibration multisets, policies, tasks, compilers, or vendors. The release bundle pins the arrays, executables, complete episode table, decision, and historical lock by SHA-256 and provides an offline statistical replay.

Optimization-level factorial control. The Hailo artifacts above were built on a CPU-only DFC 3.33.1 host whose automatic fallback selected level 0; this was not a user-selected setting. A later outcome-blind audit forced level 1 on that host. At level 1, full-256 and explicit-first-64 artifacts remain action-identical on all 800 fixed states in both orders, while original and reverse still differ on every state. Their action-difference RMSE falls from 0.10370 at level 0 to 0.09430 at level 1.

We then locked the construction-matched four-cell family in Table 33 before opening 512 fresh paired starts (136101–136612), disjoint from the earlier same-multiset audit. All four Hailo Executable Format (HEF) files were built in one process from the same source Hailo Archive (HAR); within each order, both levels received the same reshaped calibration array.

Table 33: CPU-forced optimization-level by calibration-order task control. The level-0 estimates use a new seed block and therefore need not equal the earlier same-multiset estimate.

DFC level	Calibration order	Mean return	Terminal success
0	Original	-2.0676	71.29%
0	Reverse	-0.8465	95.31%
1	Original	-1.9157	70.31%
1	Reverse	-0.9038	94.34%

Original-minus-reverse return is -1.2210 [-1.3789, -1.0623] at level 0 and -1.0118 [-1.1594, -0.8696] at level 1. The registered signed interaction is $+0.2092$ [0.1249, 0.2934]: statistically nonzero but below the prospectively reused 0.5-return materiality yardstick. Both raw success gaps are exactly $-123/512$ (-24.02 percentage points), with interaction 0.00 points [-4.10, 4.30]. Level 0 is therefore a modest moderator, not a complete explanation of this selected order effect. The control covers one policy, multiset, CPU DFC path, and forced level-1 build; the broader robustness audit below covers the headline family on the same CPU path; a GPU-visible host is audited separately in Appendix D.3.2.

Headline-family optimization-level robustness. A later informed audit rebuilds the complete seed-9500 eight-rule family and the eight policies' early/middle cells at forced levels 0 and 1 on the same CPU DFC 3.33.1 path. It covers 48 artifact conditions and 24,576 episodes. Before the level-1 outcomes were interpreted, the matched level-0 cells exactly reproduced all 12,288 historical outcome rows, while every level-1 artifact changes its 800-state action field. At level 1, the selected best–worst eight-rule return range is 14.5666 [13.9194, 15.2111] and its Holm-adjusted family test has $p < 0.001$; terminal success spans 96.2% to 23.2%. A post-outcome paired bootstrap that reselects both extremes estimates the range change as -0.3596 [-0.5348, -0.2052]. It is not a locked equivalence test, and its interval extends just beyond -0.5. The level-1 point range retains 97.6% of the level-0 range. All eight early-minus-middle policy contrasts are positive, with mean 8.4692 [4.2834, 12.3827] and median 8.6046. Their raw exact two-sided sign-flip value is $p = 0.0078125$, the attainable minimum for eight policies. It remains 0.0078125 after Holm adjustment across the two CAL8 primary tests because the companion Friedman p-value is smaller. Seed 9514 alone crosses zero (-0.0376 to $+0.1288$); the level-1-minus-level-0 change in the policy mean is -0.1261 [-0.4214, 0.1066] ($p = 0.5234$). The lower sign-flip value is therefore the arithmetic consequence of one near-zero sign change, not an independent strengthening of the evidence. Because the level-1 cells were chosen after the level-0 findings, this is a robustness factorial rather than an independent replication; it neither isolates Bias Correction from every level-dependent transform nor replaces the separate GPU-path audit of Appendix D.3.2. Tables 34 and 35 expose every level-0/1 family row and policy contrast used in this interpretation.

CAL3 fixed breadth before opening outcomes: two disjoint random 256-state multisets from the seed-9500 panel and one random multiset from independently trained seed 9510, each in original and reversed order. Every cell also included the two explicit first-64 controls. The family required

Table 34: Complete matched-level seed-9500 calibration family. Returns and terminal-success rates use the same 1,024 paired starts at each level. The level-0 rows reproduce the historical episode table exactly; level 1 changes every fixed-panel action field. The selected extremes change from even episodes/phase middle at level 0 to phase early/phase late at level 1; retaining 97.6% of the spread is therefore not tied to one artifact pair.

Calibration rule	Level-0 return	Level-1 return	Level-0 success	Level-1 success
Episode head	−0.967	−1.183	92.68%	87.60%
Episode tail	−0.840	−0.834	96.39%	96.09%
Even episodes	−0.820	−0.841	96.58%	96.19%
Odd episodes	−0.901	−0.898	94.63%	94.14%
Phase early	−0.837	−0.821	95.70%	95.80%
Phase middle	−15.747	−15.383	21.39%	23.34%
Phase late	−15.628	−15.387	21.97%	23.24%
Fixed random	−1.106	−0.845	92.38%	95.41%
Best–worst return range	14.926	14.567	–	–

Table 35: Phase-early-minus-phase-middle return contrasts for all eight policies. Levels 0–1 are the matched rebuilds; levels 2–3 are the learned-stage extension on the same 256-start blocks.

Policy seed	Level 0	Level 1	Level 2	Level 3
9510	6.229	6.555	6.534	6.454
9511	13.024	13.105	13.042	13.012
9512	16.608	16.562	16.389	16.439
9513	7.358	6.319	6.590	6.510
9514	−0.038	+0.129	+0.083	+0.134
9515	0.146	0.105	0.069	0.105
9516	14.438	14.324	14.218	14.235
9517	10.996	10.654	10.127	10.976
Policy mean	8.595	8.469	8.382	8.483
Policy median	9.177	8.605	8.359	8.743

at least two of three Bonferroni-adjusted intervals to exclude zero and exceed 0.5 in absolute return, including one seed-9500 cell and seed 9510; no extension was allowed.

Table 36: Prospective row-order breadth audit on 256 fresh paired starts per cell. Intervals are Bonferroni-adjusted over the three cells; return differences are full original minus full reverse.

Policy/multiset	Return Δ	Adjusted interval	Success Δ	Cell
9500 / random A	−0.099	[−0.135, −0.065]	−0.012	Miss
9500 / random B	+0.115	[+0.074, +0.162]	+0.012	Miss
9510 / random A	+7.465	[+5.636, +9.375]	+0.320	Pass

Table 36 has only one passing cell, so the locked two-of-three family misses. In all six full-versus-explicit comparisons, paired return and terminal success match exactly. The tested path's first-64 behavior is therefore reproducible, but CAL2's selected-case magnitude is not: the two random seed-9500 effects are small and opposite-signed, whereas seed 9510 has a large effect. The supported conclusion is conditional task sensitivity to an unreported effective subset, not a general benefit or harm from reversing calibration rows. The CAL3 release bundle tracks all arrays, 12 executables, 3,840 episode rows, and a clean-clone replay.

D.3 Compiler-level and calibration-budget robustness audits

D.3.1 Forced optimization levels through the learned stages

CAL8 stopped at the compiler-effort levels. CAL9 extends the same 24 base cells, the eight seed-9500 rules plus phase-early and phase-middle for seeds 9510–9517, to the levels at which DFC

activates learned optimization. Reading the `get_flavor_config` optimization-level dispatch in the installed DFC 3.33.1 package shows that an explicitly requested optimization level bypasses the vendor default heuristic that had produced our level-0 fallback, so no GPU host is required; CAL9 therefore runs on the same CPU host as every earlier family, and level effects are not confounded with a host change. A GPU-visible host is audited separately in Appendix D.3.2. The ladder is not cumulative: level 1 runs Bias Correction, level 2 replaces it with quantization-aware fine-tuning, and levels 3–4 run AdaRound instead.

Stage A compiled 120 artifacts, five levels per cell, and gated every one on the consumed-entry count, the absence of an automatic level reduction, and the completion line for that level's learned stage. Levels 1–3 changed the emulated 800-state action field in all 24 cells, and level 4 was bit-identical to level 3 in all 24 because both AdaRound defaults clamp to the family's 64 available rows; level 4 is therefore a construction control, not a task condition. On hardware, 96 executables repeated bit-exactly, the 24 matched level-0 fields reproduced the historical level-0 dumps exactly, and levels 1–3 each changed the field in all 24 cells. Table 37 records the default, supplied, and effective row budgets for each learned stage.

Table 37: DFC learned-stage row budgets in the optimization-level and calibration-budget audits. The default column is the learned stage's `dataset_size`; Bias Correction has no separately parameterized training size. Level 4 is a construction control, and the final row is the unclamped production-default AdaRound build. All optimization-level audit rows receive the same stride-4 64-row calibration input for the statistics pass and, where applicable, the learned stage.

Level / stage	Default rows	Rows supplied	Rows actually used
0 / none	–	64	–
1 / Bias Correction	–	64	64 calibration rows
2 / fine-tuning	1,024	64	64 (clamped)
3 / AdaRound (CAL9)	256	64	64 (clamped)
4 / AdaRound control	1,024	64	64 (clamped)
3 / AdaRound (CAL10 full)	256	256	256

The task family retained the original 1,024-seed and 256-seed blocks. Because level is a factor inside the two CAL8 headline tests rather than a new family, each test was required to hold at both learned levels 2 and 3 and combined as an intersection-union, which controls size at α without further adjustment and leaves Holm running over two tests. Both headlines hold: the eight-rule family test has an adjusted $p < 10^{-300}$ with a least-favourable range interval lower bound of 13.968, and the eight-policy contrast has an adjusted $p = 0.0078$ with a least-favourable median contrast of 8.359. A pre-specified per-level split family under four-test Holm also passes all four tests, so the conclusion does not depend on the multiplicity construction. Table 35 gives all 32 per-policy contrasts. All 24,576 historical rows reproduced exactly, although the rebuilt executables were byte-different from their predecessors in all 48 shared cells: DFC output is not byte-reproducible across build sessions even when behavior is, so artifact hash equality is not a valid reproduction criterion for this compiler and measured action-field equality is.

Table 38: Forced optimization levels on the complete eight-rule seed-9500 family. Ranges are selection-aware best-to-worst mean-return bootstrap intervals over the same 1,024 paired starts; success reports the maximum–minimum over all eight rules. Level 4 is omitted because it is construction-identical to level 3 at this budget. Every row uses the historical stride-4 64-row statistics set; the head-64 arm in Table 42 is a different 64-row subset. Level-0/1 intervals are independently recomputed for this audit rather than copies of the earlier tables; endpoints differ by at most 0.010 and no status changes. Appendix C.6 documents a separate recomputation whose range endpoints differ by at most 0.018.

Level	Learned stage	Best–worst range [95% CI]	Retained	Success max–min
0	none	14.926 [14.308, 15.546]	–	96.6–21.4%
1	Bias Correction	14.567 [13.926, 15.213]	97.6%	96.2–23.2%
2	Fine-tuning	14.782 [14.149, 15.430]	99.0%	95.8–22.5%
3	AdaRound	14.561 [13.968, 15.186]	97.6%	96.4–23.0%

Table 38 shows no monotone erosion: level 2 retains more of the range than level 1 does. Every level changes per-rule returns decisively while none moves the cross-policy contrast detectably, all three interaction intervals covering zero. CAL9 remains an outcome-informed audit of known level-0 results on one CPU path, and its learned stages ran clamped from their default training sizes to the 64 rows this family supplies, which CAL10 addresses.

D.3.2 GPU-VISIBLE DFC HOST: LEVEL-DEPENDENT SUBSET AND BUILD-PATH EFFECTS

The preceding CAL9 and row-order Hailo families were built on a CPU host, because an explicitly requested optimization level bypasses the vendor's GPU-and-dataset heuristic and no GPU is therefore required. That leaves one confound the earlier audits cannot remove: the 64-row cap and the first-64 equivalence might be artifacts of a build host on which the compiler never saw a GPU. This audit removes it.

The factorial was compiled on an NVIDIA RTX A5000 (compute capability 8.6, driver 580.173.02) with DFC 3.33.1 and TensorFlow 2.18 under CUDA 12.5.1. It crosses six visibility-and-level modes—automatic and forced levels 0 and 1, with the GPU visible and hidden—against four calibration conditions: 24 primary executables plus six additional full-set, original-order rebuilds, one per mode, giving 30 executables in total. Forcing the level under both visibility settings separates GPU visibility from the requested level, so the two cannot be confounded. Newer Blackwell parts are excluded on both sides of the boundary: an RTX 5070 Ti at capability 12.0 and a GB10 at 12.1 both fall outside the sm_60–sm_89 range this TensorFlow build supports, while the A5000 at 8.6 is inside it.

The GPU path is genuinely engaged. Under GPU visibility the automatic optimization level rises from 0 to 1, no CPU fallback is reported, and the compiler selects GPU 0. A separately locked Nsight Systems trace of one GPU-visible automatic-level optimization call attributes 14,921 CUDA kernel events to the DFC Python process, including named TensorFlow GPU operations such as `AddV2_GPU_DT_FLOAT_DT_FLOAT_kernel` (1,463 launches). An earlier one-second sampling trace of the same cell had observed 286 MiB of framebuffer memory but no positive SM sample, and its pre-registered verdict was indeterminate rather than negative, since one-second averages can miss short kernels. The optional TensorFlow `Created device` log line is absent in this build; that is a logging-level artifact and is not evidence of CPU execution.

The construction result is unchanged by any of this. All 30 builds report 64 consumed entries, including the 18 that supply 256 rows and every GPU mode. Each executable was dumped twice on the locked 800-state panel and repeated within tolerance. Full-256 and explicit-first-64 action fields are exactly identical—maximum absolute difference 0.0—in all six modes and in both supply orders, 12 comparisons of 12. The order effect also survives the host change at its CPU magnitude: original against reverse differs on 800 of 800 states in every mode, with maxima of 1.110 at level 0 and 1.063 at level 1.

One endpoint needs its magnitude rather than its verdict. Comparing GPU-visible against GPU-hidden builds at a matched forced level, 6 of 8 pairs fall outside the audit's 10^{-9} tolerance. The separations are 1.19×10^{-7} to 2.38×10^{-7}, the latter being the float32 spacing at unit magnitude, so the tolerance is stricter than float32 round-trip noise; the order effect measured on the same panel is 1.110, more than six orders of magnitude larger. We report this as a numerical magnitude and draw no behavioural conclusion from it.

For levels 0–1, the audit covers one model, one accelerator target, and one compiler version, and takes no Hailo-8 action beyond the fixed-state dumps. It establishes that their effective first-64 behaviour is not caused by hiding the GPU from the compiler.

Learned-level extension. A separately locked 20-artifact factorial crosses forced levels 2 and 3 with GPU visibility and the same four calibration conditions; each mode also rebuilds full-original once. Every build still reports 64 rows for quantizer statistics, but fine-tuning uses 256 rows for a full condition and 64 for an explicit-prefix condition, while AdaRound uses 256 and 64 respectively. All artifacts are then evaluated through one common SDK-quantized emulator on the locked 800-state panel, so the comparison changes construction, not the inference path.

The level-0/1 identity does not extend to these learned stages. Full-256 and explicit-first-64 fields differ on all 800 states in all four modes, with maximum action differences 0.0866–0.1495. Full-

original and full-reverse also differ on all states, with maxima 0.9999–1.0392. Table 39 reports the initial two-build comparison that motivated the repeated-build audit below.

Table 39: Initial two-build learned-level construction audit on one common 800-state SDK-emulator panel. "Rebuild pair" compares two full-original builds; the other columns report maximum absolute action differences. The material reference fixed before compilation is 1.0628; values above 10^{-6} but below it are intermediate.

Level	Build visibility	Rebuild pair	Full vs first 64	Original vs reverse
2	GPU visible	0.0000	0.0866	1.0235
2	GPU hidden	0.0000	0.1102	1.0392
3	GPU visible	0.01574	0.1416	0.9999
3	GPU hidden	0.0000	0.1495	0.9999

Repeated-build distribution. The initial rebuild maximum is an order statistic, not a "floor." A later locked audit rebuilds 20 cells ten times each in one runner session: the two bridge phase conditions at level 3 and all four row-order conditions at levels 2 and 3, each crossed with GPU visibility. Every cell produces ten distinct optimized-HAR hashes and ten distinct HEF hashes. Nevertheless, 16 cells produce one SDK-emulator action field across all ten builds; every GPU-hidden cell and every level-2 cell is action-deterministic on this panel. The four exceptions are all GPU-visible level-3 cells. They produce two or three action fields, with exact-pair rates 0.400–0.800 and maximum pairwise differences 0.00336–0.01575.

The build is the statistical unit; the 45 within-cell pairs are dependent descriptions. For each of ten matched condition–level comparisons, the locked statistic subtracts mean within-visibility rebuild RMSE from mean between-visibility paired-build RMSE. The equivalence margin is 0.0019685. Table 40 summarizes the family-adjusted percentile intervals; the complete per-comparison records ship with the field bundle.

Table 40: Ten-build GPU-visible/hidden build-path audit on the common SDK-emulator field. Δ is between-visibility minus within-visibility mean RMSE. "Above/equiv." counts intervals wholly above the locked 0.0019685 margin or wholly below it. Intervals use ten paired blocks and are approximate; six are zero-width because both sampled arms are deterministic.

Family	Level	Comparisons	Δ range	Above/equiv.
Bridge phase	3	2	0.00354–0.00557	2 / 0
Row order	2	4	2.4×10^{-8}–0.01236	2 / 2
Row order	3	4	0.00268–0.00576	4 / 0

All ten exact block-permutation tests have raw $p = 0.001953$ and Holm-adjusted $p = 0.01953$, but the two equivalence verdicts follow from their adjusted intervals relative to the margin, not from interpreting a nonzero permutation statistic as absence. A separate portable data-path replay reproduces every one of the 20 cell summaries and ten comparisons without a mismatch; it reuses the locked analyzer functions and therefore checks inputs and wiring rather than providing an independent implementation of the formulas.

Fixed-state hardware bridge. Eight production-default level-3 FetchReach HEFs were each run twice on the same locked 800-state panel; all repeats are bit-exact. Four of five emulator and hardware maximum comparisons agree to the reported digit. In particular, the GPU-visible phase-early rebuild differs by 0.015746 on both paths, showing that this rebuild difference reaches the accelerator rather than stopping at an intermediate representation. The exception is the GPU-visible/hidden phase-middle pair: its maximum is 0.010085 in the emulator and 0.850470 on hardware, an 84.33-fold underestimate. The hardware RMSE is 0.015545, versus 0.171893 for the early–middle reference contrast, so the large maximum is concentrated rather than panel-wide.

The 200-build audit therefore characterizes a GPU-visible/hidden build-path effect on the optimized-HAR SDK-emulator action field, not a hardware device effect. The fixed-state gate validates only eight artifacts and by itself opens no task outcome.

Locked task bridge. Under a separate precommitted outcome lock, the same eight HEFs and an FP32 arm were evaluated on 512 fresh paired FetchReach starts. The primary estimand is phase-early minus phase-middle return, computed separately for the two separately compiled GPU-visible pairs; the intersection-union rule requires both 95% lower bounds to exceed 0.5. Table 41 reports every arm.

Table 41: Production-default level-3 task bridge. Return is phase-early minus phase-middle paired return; success gives the two condition rates. GPU-visible rows form the locked primary intersection-union, while GPU-hidden rows are secondary.

Build path	Repeat	Return [95% CI]	Success, early/middle
GPU visible	0	14.670 [13.808, 15.551]	95.7% / 22.5%
GPU visible	1	14.675 [13.811, 15.555]	96.0% / 22.5%
GPU hidden	0	14.689 [13.827, 15.566]	95.7% / 22.3%
GPU hidden	1	14.689 [13.827, 15.566]	95.7% / 22.3%

Both primary repeats pass, with 500/512 and 499/512 positive pairs. FP32 mean return is -0.670 with 100% success; it is contextual rather than part of the primary. The visibility interactions are -0.0186 [-0.0691, 0.0157] and -0.0140 [-0.0647, 0.0206], which do not establish path equivalence. The GPU-visible rebuild interaction is 0.00461 [0.00251, 0.00685], about 0.03% of the 14.67 contrast, while the GPU-hidden interaction is exactly zero on all 512 pairs. Thus build nondeterminism reaches task outcome in one arm but is negligible relative to the selected contrast in this audit.

The task result closes the automatic-level-0 and CPU-only-host explanations for this one contrast: both builds use 256 statistics rows, 256 AdaRound rows, and the production-default level-3 GPU-visible path. It is outcome-informed robustness evidence, not an independent discovery or a rerun of the complete eight-rule family. One post-outcome amendment supplies a missing descriptive scope field that no calculation reads; all nine outcome shards were reused and the episode table reproduced byte-for-byte. Generalization across policies, tasks, GPUs, Hailo targets, DFC versions, or physical robots remains open.

D.4 Quantizer statistics budget versus optimization level

Every earlier Hailo family supplied 64 calibration rows, a stride-4 uniform subsample of a 256-row representative set. CAL10 supplies the full 256 rows for all eight seed-9500 rules and crosses optimization level $\{0, 3\}$ with the quantizer statistics count $\{64, 256\}$, 32 executables. Two budgets are distinguished here that earlier families held together: the statistics pass, whose vendor default consumes 64 rows and which produces the reported effective first-64 behavior, and the learned stage's training set, which draws on all supplied rows. With 256 rows supplied, AdaRound trains on 256 in both statistics arms, so no CAL10 build is clamped and the level-3 arms are production-default builds. At a statistics count of 64 the compiler consumes the contiguous head block of the 256-row set, a different multiset from the stride-4 subsample the earlier families used; the budget contrast therefore changes count and consumed subset together and isolates no pure count effect.

Table 42: Calibration-budget by optimization-level build family. Ranges are selection-aware best-to-worst mean-return intervals on the original 1,024-seed block. The benign-six column restricts the family to the rules other than phase-middle and phase-late, a restriction declared before outcome access.

Level	Statistics rows	Best–worst range [95% CI]	Benign-six range	Benign-six status
0	64 (head block)	6.732 [6.549, 6.906]	1.826 [1.692, 1.957]	materially heterogeneous
0	256 (all)	14.969 [14.350, 15.582]	0.215 [0.174, 0.260]	compatible with limited heterogeneity
3	64 (head block)	6.634 [6.444, 6.827]	1.752 [1.627, 1.877]	materially heterogeneous
3	256 (all)	14.420 [13.825, 15.045]	0.146 [0.100, 0.196]	compatible with limited heterogeneity

Table 42 reports the four cells. Both full-budget headline tests pass under two-test Holm, with adjusted $p < 10^{-300}$ and interval lower bounds of 14.350 and 13.825 against the 0.5 threshold. Terminal success is 97.2% versus 21.4% at level 0 and 96.2% versus 22.9% at level 3, about 28–30 cm of mean post-step goal-distance separation. A production-default build therefore reproduces

the family spread, which closes the reading that the effect is an artifact of the automatic level-0 fallback on this path.

Comparing the three sampling choices at fixed policy, seeds, panel, and device, the headline range is 14.926 for a 64-row stride-4 subsample, 6.732 for a contiguous 64-row head block, and 14.969 for all 256 rows. The two 64-row constructions yield substantially different spreads, while the stride-4 subset and full set yield similar spreads. Supplied row count alone is therefore insufficient to characterize the observed outcome. The default-head value remains 13.5 times the 0.5 practical threshold. The head block compresses returns toward a middle range: the two collapsed rules improve from -15.75 and -15.63 to -7.65 and -7.63, while every benign rule degrades, for example even-episodes from -0.78 to -2.74.

The pre-registered bimodality restriction is regime-dependent. At the full budget the six benign rules span 0.215 and 0.146, well inside the 0.5 practical threshold, so the family is six practically equivalent rules plus two catastrophic ones. In the starved head-block arms the same six span 1.826 and 1.752 and calibration choice is material across the whole family. Bimodality is a property of the well-fed regime, not of the family.

The repeated 21.39% phase-middle success rate at level 0 is not copied from CAL8. CAL10's separate 1,024-row episode table also contains 219 successes, but its HEF and fixed-panel action hashes differ from CAL8's; the corresponding mean returns are -15.746492 and -15.746516, respectively. The full-budget rule-wise fidelity–return results and uncertainty audit are reported once in Appendix B.2, Tables 9 and 10.

D.5 Prospectively locked calibration-coverage audits

The historical FetchReach family suggested a cheap input-side diagnostic. For a 64-row calibration set S, let ℓ_j, u_j be coordinate extrema on a fixed 800-state representative panel and define span coverage as

$$c(S) = \frac{1}{d}\sum_{j=1}^{d} \mathrm{clip}\left(\frac{\max_{s\in S} s_j - \min_{s\in S} s_j}{u_j - \ell_j}, 0, 1\right).$$

This statistic was defined after the historical outcomes were known and separated their two collapsed rules from the benign six. The audit below is its first prospectively locked test, not an independent discovery or a validated qualification criterion.

The outcome-blind construction forms a 798-state pool from the eight historical representative sets and fixes eight 64-row cells with target coverage $\{0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.75, 0.90\}$. Candidate sets combine rows near an anchor with a fixed farthest-point ordering; the closest candidate to each target is selected before compilation. Thus pool, row count, policy, compiler path, build specification, device, and evaluation seeds are fixed, but anchor and distribution geometry vary with the selected subset. The design tests the deterministic coverage-targeted construction and removes phase identity; it does not isolate scalar coverage while holding every other distributional property fixed.

All eight CPU DFC 3.33.1 level-0 builds consume exactly the supplied 64 rows, emit no clamp or level-reduction warning, and produce distinct 800-state action fields. Both hardware executions of each field are bit-exact. Stage 1 then uses the same 1,024 reset seeds for every cell; the maximum initial-state mismatch is zero.

Table 43 gives the eight locked cells and the two historical comparators. The primary test correlates target coverage with the eight cell means. It passes all three gates: Spearman $\rho = 0.9048$ exceeds 0.7, the two-sided exact permutation value over all $8! = 40{,}320$ orderings is 0.00456, and the best-to-worst range is 6.5225 against 0.5. The four targets at or above 0.50 have an observed range of only 0.066, consistent with a plateau rather than a material gain from additional coverage in this construction.

Both recorded directional predictions also hit. Neither of the two prospective cells at or above the historical 0.638 floor collapses under the fixed -5-return rule. Three of the four cells below 0.47 do not collapse, so low coverage is not sufficient for failure. The first result is only a two-cell, one-policy candidate screen, not a sensitivity guarantee.

Table 43: Locked span-coverage ladder and historical comparators. The bottom two rows were known before the ladder and are descriptive, not members of its eight-cell primary test. Return is the 50-step dense FetchReach return; higher is better.

Cell	Coverage	Mean return	Success	Role
Target 0.10	0.0996	−7.470	2.8%	Locked ladder
Target 0.20	0.2010	−4.355	40.6%	Locked ladder
Target 0.30	0.3081	−2.247	66.3%	Locked ladder
Target 0.40	0.4011	−1.657	72.3%	Locked ladder
Target 0.50	0.4987	−0.982	93.4%	Locked ladder
Target 0.60	0.6016	−1.013	92.8%	Locked ladder
Target 0.75	0.7491	−1.014	89.6%	Locked ladder
Target 0.90	0.8973	−0.947	93.6%	Locked ladder
Phase middle	0.4640	−15.747	21.4%	Historical
Phase late	0.4617	−15.628	22.0%	Historical

The historical comparison supplies the stronger boundary. At nearby coverage, phase-middle/late are about 14.7 return units worse than the constructed 0.499 cell, approximately 29.5 cm in mean per-step goal distance. Span coverage can therefore flag a calibration set for audit before compilation, once a representative panel is available, but cannot approve an executable. Exact paired task outcomes remain the qualification evidence.

D.5.1 Separating calibration geometry from scalar coverage

CAL11 varied anchor and coverage together, so its dose response could not be attributed to the scalar. CAL12 separates them with two coverage bands of six anchors each, every cell being that anchor's 64 nearest pool rows from the same 798-state pool, at the same build specification, device, and 1,024 starts. Within a band coverage is nearly constant while the subset centre moves; accepted anchors must not duplicate an existing row set and must lie at least 0.30 apart in normalised coordinates. Two constructions that instead hold the anchor fixed are degenerate in this pool: four farthest-point rows lift coverage from 0.49 to 0.88, and thinning the anchor's distance ordering gives a narrow, non-monotone 0.60–0.71. Both decisions are interval-based, so each uses a Bonferroni-adjusted 97.5% bootstrap interval.

Table 44 gives the twelve cells. Both within-band ranges are material: 9.547 [9.245, 9.844] in the low band and 1.779 [1.630, 1.933] in the mid band. The between-band contrast, a roughly fourfold coverage change, is 7.201 [7.024, 7.379]. This between-band mean contrast and the within-band range are different estimands, and neither isolates a pure coverage or geometry effect. Inside the low band terminal success runs from 0.8% to 62.1% with no monotone relation to the statistic.

CAL11's proposed screen at coverage ≥ 0.638 remains untested: this band is empty here. Its insufficiency prediction is reinforced by low_435's 62.1% success at coverage 0.1304. Coverage cannot completely order or approve these cells: mid_288 returns -2.746 at 60.5% success and coverage 0.4642, inside the band whose CAL11 cell was benign.

The endpoints low_720 and low_435 (coverage 0.1236/0.1304) differ by 9.547 return and 61.3 percentage points of success. Pool, count, policy, build specification, device, and starts are fixed; coverage and other subset properties are not. These deliberate constructions on one policy and one CPU level-0 path demonstrate the limits of the tested scalar summary, not a universally causal geometry variable.

Table 44: Coverage-matched calibration bands. Coverage spans 0.0184 in the low band and 0.0798 in the mid band; subset centre, radius, and density also vary.

Band	Cell	Coverage	Mean return	Success
Low	low_760	0.1155	−8.397	1.6%
Low	low_506	0.1200	−6.352	48.1%
Low	low_720	0.1236	−12.619	0.8%
Low	low_435	0.1304	−3.072	62.1%
Low	low_190	0.1309	−8.219	2.1%
Low	low_709	0.1339	−12.567	0.8%
Mid	mid_659	0.4401	−1.189	95.4%
Mid	mid_288	0.4642	−2.746	60.5%
Mid	mid_450	0.4840	−0.977	93.4%
Mid	mid_448	0.4879	−0.967	93.4%
Mid	mid_498	0.4987	−1.041	91.9%
Mid	mid_456	0.5199	−1.100	95.4%

E Qualification Decisions under a Fixed Task Budget

E.1 Completed prospective task-decision benchmark

The prospective task-decision cohort is separate from fixed-panel screening. It comprises 8 fresh SAC policies in FetchReachDense-v4 and retains all after FP32-only capability screening. Each policy has 16 artifacts (8 calibration rules, compiler levels 0 and 3). Final checkpoints, disjoint calibration/fidelity/audit/holdout seed blocks, tolerance $\delta = 0.5$, runner, analyzer and comparator were locally sealed before the relevant outcomes. Staged permissions preserve the initial planning record; this is not a claim of public preregistration. Every task action is executed on Hailo-8 in a simulated FetchReach loop, not a physical robot.

The audit uses 32 paired reset seeds per policy, with all decisions sealed before the separately invoked 512-seed holdout. Both blocks use shared episode-bootstrap indices across the within-policy artifact family, 20,000 replicates and two-sided $\alpha = 0.05/16$. Loss is FP32 minus artifact return. Approve if the upper bound is below tolerance, reject if the lower bound exceeds it, otherwise defer. Holdout uses the same interval rule on its disjoint block and labels outcomes compatible/degraded/unresolved, not ground truth. Full-horizon traces, reward identities, paired initial states, RPC transcripts and pre/post fingerprints all pass re-verification.

Table 45: Prospective audit decisions. 32 shared reset seeds per policy require 544 task episodes across FP32 and 16 artifact arms; all methods reuse these episodes. $A/R/D$ are approval/rejection/deferral counts; A_D/A and R_C/R are conditional conflicts with degraded/compatible holdout intervals. U_A/U_R are unresolved approvals/rejections, not errors or confirmed correctness.

Method	A	R	D	A_D/A	R_C/R	U_A/U_R
Family-adjusted paired CI	66	32	30	0/66	0/32	0/0
Mean cutoff	81	47	0	0/81	0/47	3/2
Margin $\tau = 0$	81	47	0	0/81	0/47	3/2
Margin $\tau = 0.1$	78	45	5	0/78	0/45	0/2
Margin $\tau = 0.2$	76	45	7	0/76	0/45	0/2
Margin $\tau = 0.3$	65	43	20	0/65	0/43	0/1
Margin $\tau = 0.5$	2	41	85	0/2	0/41	0/1
Margin $\tau = 1$	0	33	95	undefined	0/33	0/0
Matched-count diagnostic	64	34	30	0/64	0/34	0/0

Holdout classifies 78 artifacts as compatible, 45 as degraded and 5 as unresolved. All tested methods have zero observed approval/rejection conflicts (Table 45); mean cutoff nevertheless approves 3 unresolved artifacts. Unresolved records stay in denominators, not in confirmed-correct counts. A zero denominator is undefined, not zero risk. The primary margin $\tau = 0.3$ was fixed before outcomes; the full grid is reported without selecting a winner. Matched-count allocation borrows CI deferral counts and is only a diagnostic.

Table 46: Registered CI-minus-fixed-margin contrasts, percentage points. Intervals resample policies, not artifacts or episodes.

Estimand	Difference	Interval
Unconditional approval conflict	0.0000	[0.0000, 0.0000]
Unconditional rejection conflict	0.0000	[0.0000, 0.0000]
Deferral	7.8125	[-3.9062, 25.0000]

The 8 policies are the top-level units. The 20,000-replicate policy bootstrap allocates $\alpha = 0.05/3$ across the registered contrasts (Table 46). Both conflict intervals degenerate because every observed policy contribution is zero; this establishes neither zero population risk, equivalence nor false-approval control. CI defers more in aggregate, but its difference interval spans zero. This completed prospective comparison does not establish CI superiority. All intervals are approximate.

The secondary fidelity analysis of these same policies and holdout episodes is in Appendix B.3; it is not an independent cohort. The timing intervention reuses the policies with fresh resets (Appendix F.1).

Audit collection uses 4,352 episodes (256 shared FP32, 4,096 artifact); holdout uses 69,632 (4,096 FP32, 65,536 artifact). All methods reuse these episodes. The machine-readable record separates FP32/artifact worker times and collection wall time, including setup and pre/post probes; these are not pure inference latencies or total training/compilation costs. It retains every policy-specific full decision–holdout table, denominator, unresolved rate and secondary interval. All registered arms completed without a task-collection failure.

The complete machine-readable fresh-policy comparison record is retained in the reproduction bundle. This benchmark is described in both manuscripts; it is shared evidence, not independent replication, and its evidence is retained locally without a dependency on another submission.

E.2 Retrospective replay and matched-deferral sensitivity

At matched deferral, retrospective replay shows no uniform CI advantage.

This reanalysis uses the already observed complete eight-artifact family per path. It evaluates decisions, not best-artifact selection, and does not revise the original audit. A fixed random seed generates 200 overlapping resplits of the 1,024 episode seeds. Each split reserves 512 seeds for holdout evaluation and uses nested budgets $n \in \{8, 32, 128, 256\}$ from the other half. Every method receives the same paired losses for eight artifacts and FP32, costing $9n$ task episodes per path; holdout evaluation is additional. Artifacts are unchanged across the split, so this does not test unseen policies or artifacts.

Four rules are compared. *Mean cutoff* applies the 0.5-return tolerance to the audit mean alone. *Fitted MSE* chooses the lower-MSE-is-compatible threshold with the fewest disagreements with those audit labels, breaking ties toward fewer approvals; it is an optimistic task-supervised comparator, not an offline-only method. *Paired CI* uses a marginal 95% interval, and *Family CI* applies Bonferroni over the eight artifacts. Both use 2,000 percentile bootstrap replicates of complete paired episodes and defer when the interval meets or spans the tolerance. The MSE rule additionally uses the existing 800-state action panel. No holdout outcomes fit a rule.

Holdout comparison uses the same tolerance with eight-artifact Bonferroni bootstrap intervals. An *approval conflict* is an audit approval whose holdout lower bound exceeds 0.5. An approval with an unresolved holdout interval is reported separately, rather than counted as a confirmed correct decision. Table 47 gives counts and deferral rates.

Table 47: Retrospective decision replay, 200 resplits and eight decisions per path per split. Conflict entries are conflicting approvals/all approvals; defer is a percentage of all 1,600 decisions. Hailo unresolved is the percentage of approvals with unresolved holdout status. Resplits overlap; these counts are not independent validation trials. $9n$ excludes holdout episodes and the fixed-state MSE panel.

			EdgeTPU		Hailo		
n	Episodes	Rule	Conflict	Defer (%)	Conflict	Unresolved (%)	Defer (%)
8	72	Mean cutoff	98/1106	0.0	0/967	11.8	0.0
8	72	Fitted MSE	96/1188	0.0	0/956	11.9	0.0
8	72	Paired CI	8/804	48.0	0/713	12.3	30.4
8	72	Family CI	4/740	53.4	0/665	13.1	33.6
32	288	Mean cutoff	32/1162	0.0	0/1076	10.6	0.0
32	288	Fitted MSE	32/1226	0.0	0/1059	10.8	0.0
32	288	Paired CI	2/702	48.0	0/610	4.9	36.9
32	288	Family CI	0/529	64.2	0/459	5.0	46.3
128	1152	Mean cutoff	0/1188	0.0	0/1141	12.1	0.0
128	1152	Fitted MSE	0/1200	0.0	0/1142	12.1	0.0
128	1152	Paired CI	0/1019	13.8	0/967	2.3	14.5
128	1152	Family CI	0/929	26.6	0/856	0.9	21.5
256	2304	Mean cutoff	0/1200	0.0	0/1165	13.9	0.0
256	2304	Fitted MSE	0/1200	0.0	0/1165	13.9	0.0
256	2304	Paired CI	0/1148	3.6	0/1017	2.4	11.4
256	2304	Family CI	0/1046	10.8	0/989	0.7	13.2

At eight seeds, EdgeTPU mean-cutoff and fitted-MSE approvals conflict with holdout degradation in 8.9% and 8.1% of approvals, respectively. Marginal and family CIs reduce this to 1.0% and 0.5%, with 48.0% and 53.4% deferral. At 128 seeds, all four methods have zero observed conflicts on this path. No method approves a holdout-degraded Hailo artifact at any tested budget; its large

failures are easy to separate. However, some Hailo approvals have unresolved holdout status, as the table reports. EdgeTPU has six unresolved holdout decisions across all resplits; the complete CSV records their approval counts for every method and budget.

The value demonstrated is an empirical tradeoff between early approval and deferral in this finite family. Neither the overlapping resplits nor a zero-conflict cell establishes nominal error control, future reliability, or safety. In particular, small-sample percentile intervals still make conflicting approvals. The design was chosen after the original outcomes were known and has no prospective status. The separately locked fresh-policy benchmark is now complete (Appendix E.1); it does not retroactively change this replay's status. Initial planning records and subsequent execution permissions remain distinct.

Supporting matched-deferral sensitivity. An additional, explicitly post-outcome analysis preserves the original resplits, bootstrap stream, holdout intervals, and all four original methods' counts. A mean-margin rule approves when $\widehat{D} < 0.5 - \tau$, rejects when $\widehat{D} > 0.5 + \tau$, and otherwise defers. The fixed grid is $\tau \in \{0, .01, .02, .03, .04, .05, .075, .1, .125, .15, .2, .25, .3, .4, .5, .75, 1, 2, 4, 8, 16, 32\}$. For each path and budget, we report the full grid and select the nearest aggregate deferral rate to paired CI, breaking ties toward smaller τ; no holdout outcome enters this choice. A second comparator exactly matches each resplit's CI deferral count by deferring that many artifacts nearest the tolerance in absolute audit-mean distance, with ties resolved in fixed artifact order. This comparator borrows the CI count and is an allocation diagnostic, not a standalone deployable method.

For example, the nearest grid point at eight EdgeTPU seeds is $\tau = 0.3$: it defers 42.6% versus CI's 48.0%, with $0/801$ approval conflicts but $46/118$ rejection conflicts. This grid comparison is only approximately matched in aggregate; the exact-count analysis below answers the allocation question at identical per-resplit deferral (Table 48).

Table 48: Exact matched-deferral sensitivity. Paired CI/mean-margin counts are shown in that order, not as fractions. A is approvals, A_D approvals with holdout degradation, and R_C rejections with holdout compatibility. U counts all unresolved holdout decisions. Each row has 1,600 decisions; rejection count is $1{,}600 - A - \text{deferred}$. The complete record separately reports approval/rejection/deferral rates, both conflict denominators, all nine decision–holdout cells, and unresolved approvals/rejections.

Path	n	Defer (%)	A: CI/mean	A_D: CI/mean	R_C: CI/mean	U
EdgeTPU	8	48.0	804/790	8/8	0/15	6
EdgeTPU	32	48.0	702/729	2/0	0/3	6
EdgeTPU	128	13.8	1019/1061	0/0	0/0	6
EdgeTPU	256	3.6	1148/1150	0/0	0/0	6
Hailo	8	30.4	713/698	0/0	0/9	194
Hailo	32	36.9	610/605	0/0	0/0	194
Hailo	128	14.5	967/965	0/0	0/0	194
Hailo	256	11.4	1017/1017	0/0	0/0	194

At eight EdgeTPU seeds and 48.0% deferral, approval-conflict rates are $8/804 = 1.0\%$ (CI) and $8/790 = 1.0\%$ (matched mean); rejection-conflict rates are $0/28$ and $15/42$. At 32 seeds, the mean comparator has zero approval conflicts versus $2/702$ for CI, but $3/103$ rejection conflicts versus $0/130$. Neither allocation dominates both errors at every budget. At eight Hailo seeds both have zero approval conflicts, but their unresolved approvals remain in the record; zero conflict is not proof of correctness. Each decision batch still costs $9n$ task episodes plus 9×512 holdout episodes; repeated resplits reuse those stored episodes. This is a *retrospective approval–deferral trade-off under a fixed task-audit budget*, not verified false-approval control, nominal coverage, new-policy reliability, or established cost superiority. The analysis and full grid are retained in the matched-deferral record (Appendix I.4).

E.3 Post-outcome audit-prefix sensitivity

This is a post-outcome sensitivity of the same prospective cohort, not a new cohort. The registered benchmark uses 32 audit seeds per policy. It does not prospectively test the smaller 8-seed budget at which the earlier, different-path retrospective replay showed more cutoff conflicts. Those corpora

and execution paths are not interchangeable. This supporting analysis selects the original first 8 audit seeds per policy in their fixed seed order, without reshuffling or outcome-based selection. The decision rules, margin grid, tolerance, family adjustment, bootstrap settings, and separate 512-seed holdout remain unchanged. The subset is outcome-independent, but choosing to analyze it is post-outcome; it does not replace the registered result or constitute another prospective cohort.

Table 49: Audit-prefix sensitivity on 8 policies and 128 artifacts, using 8 paired audit episodes per arm. A/R/D are approval/rejection/deferral counts. AC/A and RC/R show conflicts with resolved holdout status divided by the corresponding decisions, not by all artifacts. UA counts approvals with unresolved holdout status; 5 holdout statuses are unresolved overall. A dash denotes no decisions of that type.

Method	A	R	D	AC/A	RC/R	UA
Family-adjusted paired CI	61	24	43	1/61	0/24	1
Mean cutoff	79	49	0	3/79	3/49	1
Mean margin $\tau = 0$	79	49	0	3/79	3/49	1
Mean margin $\tau = 0.1$	76	45	7	2/76	2/45	1
Mean margin $\tau = 0.2$	68	40	20	2/68	0/40	1
Mean margin $\tau = 0.3$	60	40	28	1/60	0/40	0
Mean margin $\tau = 0.5$	10	39	79	0/10	0/39	0
Mean margin $\tau = 1$	0	30	98	—	0/30	0
Matched-count diagnostic	57	28	43	1/57	0/28	0

Table 49 reports the complete comparator grid. Both the fixed margin and the matched-count diagnostic reproduce CI's 1 approval and 0 rejection conflicts. The former defers 28 rather than 43 cases, while the latter borrows CI's deferral count and is not a standalone rule. Thus the observed advantage over non-abstaining cutoff is not exclusive to CI. This comparison does not identify deferral quantity as its causal mechanism, establish population superiority, or validate false-approval control. Holdout intervals are not ground truth; unresolved approvals remain unresolved. Policies, not artifacts or overlapping budget prefixes, are the top-level replication units.

E.4 Fidelity-based selection is not task qualification

The retrospective selection audit reuses this same cohort, not an independent validation. Minimum RMSE uses only the calibration-unused target-runtime panel; minimum audit loss uses the audit block, never holdout. Each selects an artifact rather than approving or deferring it. Primary comparisons hold compiler level fixed; pooling levels is secondary. Exact ties use registered cell order. The holdout oracle is hindsight-only and is not an implementable comparator. The degraded minimum-RMSE choices occur in the same policy across scopes, without a holdout-compatible alternative in the candidate set. This does not establish passing over a confirmed-compatible artifact. Table 50 retains every policy-cohort contrast, including those favorable to RMSE. Within-level

Table 50: Retrospective selection contrasts. Positive loss means worse return for minimum RMSE. Policy-bootstrap intervals adjust over this contrast family and are conditional on the observed panel and audit selections.

Scope	Comparator	Mean loss difference	Policy-family interval
Level 0	Audit minimum	0.0511	[-0.0001, 0.1077]
Level 0	Fixed random	-0.0793	[-0.2035, -0.0110]
Level 3	Audit minimum	0.0379	[-0.0050, 0.1052]
Level 3	Fixed random	-0.0038	[-0.0131, 0.0009]
Joint levels (secondary)	Audit minimum	0.0633	[0.0051, 0.1298]

RMSE-minus-audit intervals include zero; this is not equivalence. The joint-level secondary difference excludes zero but remains within practical tolerance. RMSE selection also outperforms the designated fixed-random comparator at the lower level. These favorable results for fidelity are retained: non-monotonicity does not imply poor minimum-score selection. The methods consume different information and budgets, so no equal-cost superiority follows. Holdout status is uncertain

evidence, not ground truth; approximate conditional intervals provide no population safety guarantee.

F Action-Source Timing and Matched-Magnitude Controls

Action replacement schedules affect return. A separately locked intervention on 8 previously evaluated policies replaces 10 of 50 actions with FP32 outputs computed at the states reached during the intervention. Replacing the first 10 rather than the middle 10 actions yields mean return gains of 4.074 [2.056, 6.156] and 4.194 [2.101, 6.238] at compiler levels 0 and 3, respectively. Intervals use family-adjusted policy bootstrap. Replacement count is matched; removed error magnitude is not (Appendix F.1). Matched-norm, equal-count injections at post-clipping action-error norms 0.05, 0.10, 0.20 yield 6 positive middle-minus-early return intervals below the 0.5-return tolerance, missing the registered materiality criterion (Appendix F.2).

F.1 Action-source timing intervention

This is a development-informed intervention on the completed task-decision benchmark's known policy cohort, not replication on task-unseen policies. The design, schedules, seed blocks, analyzer and success criteria were locally sealed before intervention outcomes, with explicit execution permissions. Local sealing is not public preregistration.

All 8 policies and compiler levels 0 and 3 are retained. Each arm uses 128 fresh paired reset seeds and 50 steps. The base artifact uses middle-phase calibration; early-phase artifacts are retained as unmodified references. FP32 is shared within each policy. No new training or compilation is performed.

Early correction replaces steps 0–9; middle correction replaces steps 20–29. The fixed dispersed schedule is $\{4, 5, 9, 13, 15, 24, 29, 38, 39, 48\}$. Indices are zero-based. Correction applies FP32 actions in the selected steps and artifact actions elsewhere; reciprocal injection instead applies artifact actions early and FP32 elsewhere. Both predictors run at every current hybrid observation, including the unchosen action. Actions use the carried clipping contract. There is no replay of actions from a different trajectory and no inference-cost or deployable fallback-policy claim.

Registered inference. The primary contrast is early-minus-middle correction. Early-minus-dispersed correction and FP32-minus-early-injection are registered secondary contrasts, not replacements for a missed primary. Joint episode resampling is paired within policy; policy means are the cohort-level units. The approximate bootstrap allocates $\alpha = 0.05/6$ across the contrast-by-level family. A prediction holds only if its cohort lower bound exceeds $\delta = 0.5$ in both levels. All three hold; this is not a finite-sample coverage theorem or population-prevalence estimate (Table 51).

Table 51: Registered action-timing contrasts. Entries are cohort mean return differences and approximate Bonferroni-adjusted policy-bootstrap intervals. Positive values have the direction stated in the first column.

Contrast	Level 0	Level 3
Early minus middle correction	4.074 [2.056, 6.156]	4.194 [2.101, 6.238]
Early minus dispersed correction	3.187 [1.020, 6.050]	3.356 [1.109, 6.144]
Loss from early artifact injection	1.611 [0.927, 2.288]	1.624 [0.938, 2.287]

Heterogeneity and scope. Positive cohort effects are not a claim that every policy benefits materially. Figure 6 retains every policy mean alongside the cohort intervals. Equal replacement counts do not equate the removed action-error norms, state exposure or feasible post-clipping perturbations. The result supports a schedule-dependent intervention effect; it does not identify a unique mechanism, establish scalar-fidelity selection failure, isolate AdaRound causality, or certify physical-robot safety.

Collection provenance. Collection was interrupted and resumed under the original execution contract. Completed rows were verified and reused without treating them as additional observations; the partial trace was excluded. Full recovery provenance is retained in Appendix I.3.

Retrospective magnitude audit. All 104 retained jobs (13,312 episodes) were re-read without new rollouts. We clip both current-state action outputs under the carried contract before computing their Euclidean difference. Table 52 reports per-episode sums, averaged within policy and then equally across policies. Early correction has smaller shadow-error norm and squared-norm sums on its corrected steps than middle correction, despite its larger return benefit. On the unmodified artifact trajectory, the early-window mean norm sum is instead larger. These are different state exposures, not causal adjustment: an intervention changes subsequent states and errors. The audit therefore does not isolate timing from error magnitude or direction, and squared action norm is not physical energy.

Table 52: Descriptive post-clipping error sums. Hybrid rows compare each correction arm on its own correction window; reference rows compare windows on the same unmodified artifact trajectory. Entries are not paired task effects or confidence bounds.

		Norm sum		Squared-norm sum	
Trajectory	Level	Early	Middle	Early	Middle
Hybrid correction	0	5.1796	8.6541	5.8836	13.5083
Artifact reference	0	14.3829	13.0829	24.1373	23.2698
Hybrid correction	3	4.7924	8.4891	5.7628	13.6475
Artifact reference	3	14.3932	13.1596	24.2367	23.6429

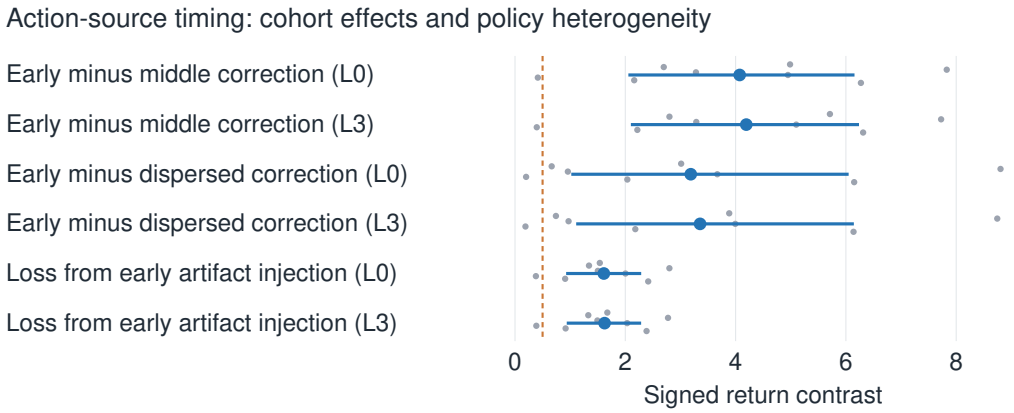

Figure 6: Action-source timing intervention. Blue dots and bars show cohort means and approximate family-adjusted policy-bootstrap intervals. Grey dots retain every policy mean, with vertical offsets for visibility, not uncertainty. The dashed line is the carried practical threshold. Positive cohort effects do not imply a material effect in every policy; correction counts, not removed error norms, are matched. Numerical results are in Table 51.

F.2 Matched-magnitude timing control and supporting interpretation

The subsequent matched-magnitude experiment retains the known policy cohort, both compiler levels and all fixed doses. It is a locally sealed, development-informed synthetic intervention with fresh paired reset seeds, not an independent new-policy replication or a naturally compiled controller. Only the injection window uses FP32 plus the normalized current-state device-minus-FP32 direction; other steps use FP32. Float32 casting and final clipping precede the norm-feasibility check. Direction and visited states are not held fixed across windows.

All 104 jobs / 13,312 episodes completed and passed the original raw simulator, RPC and action-contract verification. This supporting audit recomputes raw rewards, applied norms and the un-

changed registered bootstrap. Absolute loss (Table 54) and magnitude context are explicitly post-outcome; neither replaces the registered materiality test (Table 53).

Registered result. The signed contrast is middle-injection minus early-injection return; positive means early injection is more harmful. The locally sealed criterion required every dose-by-level lower bound to exceed $\delta = 0.5$. It fails: all 6 adjusted intervals are positive, but even their largest upper bound is 0.0371. The tested Euclidean norms are 0.05, 0.1, 0.2 action units. The original dose grid and failed P1 are unchanged; this is not a revised success criterion.

Table 53: Registered matched-norm timing contrasts: middle-minus-early return, with approximate family-adjusted policy-bootstrap intervals. The full dose-by-level family is retained.

Applied norm	Level 0	Level 3
0.05	0.0068 [0.0059, 0.0078]	0.0075 [0.0054, 0.0091]
0.1	0.0137 [0.0118, 0.0158]	0.0148 [0.0105, 0.0180]
0.2	0.0266 [0.0232, 0.0306]	0.0294 [0.0195, 0.0371]

Post-outcome absolute-loss audit. We compare each injection arm with its shared paired FP32 reference, rather than infer absolute harm from the difference between two interventions. All doses, levels and both windows form a new 12-contrast family. The paired episode and equal-weight policy bootstrap, streams and replicate counts are carried forward, with family-adjusted quantiles recomputed for this family. Intervals remain approximate; this is a post-outcome sensitivity, not a prospective qualification result. Cohort mean losses span 0.0010–0.0488; all upper bounds are at most 0.0619, below the 0.5-return tolerance. Thus the small timing contrast is not accompanied by large cohort-average losses in both arms at these doses. This does not qualify every policy or establish population equivalence.

Table 54: Post-outcome absolute task loss, FP32 minus injection return. Approximate intervals share one dose-by-level-by-window family. Negative losses mean improved return relative to FP32.

Level	Norm	Early injection	Middle injection
0	0.05	0.0096 [0.0050, 0.0129]	0.0028 [-0.0014, 0.0064]
0	0.1	0.0211 [0.0112, 0.0276]	0.0074 [-0.0018, 0.0146]
0	0.2	0.0488 [0.0281, 0.0619]	0.0222 [0.0028, 0.0361]
3	0.05	0.0085 [0.0068, 0.0101]	0.0010 [-0.0007, 0.0031]
3	0.1	0.0189 [0.0149, 0.0226]	0.0041 [0.0008, 0.0085]
3	0.2	0.0451 [0.0344, 0.0530]	0.0157 [0.0101, 0.0243]

Post-outcome magnitude context. Norms use the complete clipped action vector, not physical actuator energy. Table 55 separates new-hybrid shadow differences from historical unmodified-artifact and early-injection trajectories. Quantiles pool equal window-step counts per policy: finite traces, not independent samples or confidence bounds. Old and new trajectories remain separate. Historical early-injection and unmodified-artifact early prefixes coincide by construction and are not independent observations. The maximum tested norm, 0.2, is below historical unmodified-artifact early-window medians 1.5333–1.5415. This documents a scale limitation, not proof that magnitude alone caused the earlier large task differences: seeds, feedback trajectories and intervention backgrounds differ. No further dose or cohort was selected from these outcomes.

Table 55: Post-clipping device-minus-FP32 norm distributions, kept separate by execution context. For matched runs these are unscaled shadow norms at the current hybrid state, not the applied fixed norm. E/M denote early/middle windows.

Context	Level	Dose/window	Mean	Q05	Median	Q95
Matched hybrid	0	0.05/E	0.5323	0.0883	0.2520	1.7882
Matched hybrid	0	0.05/M	0.1930	0.0785	0.1793	0.3492
Matched hybrid	0	0.1/E	0.5496	0.0931	0.2624	1.7953

Table 55 (continued)

Context	Level	Dose/window	Mean	Q05	Median	Q95
Matched hybrid	0	0.1/M	0.2029	0.0795	0.1868	0.3619
Matched hybrid	0	0.2/E	0.5850	0.1047	0.2884	1.8079
Matched hybrid	0	0.2/M	0.2185	0.0840	0.2054	0.3894
Matched hybrid	3	0.05/E	0.4940	0.0595	0.1691	1.7846
Matched hybrid	3	0.05/M	0.1353	0.0560	0.1262	0.2464
Matched hybrid	3	0.1/E	0.5104	0.0618	0.1808	1.7929
Matched hybrid	3	0.1/M	0.1410	0.0499	0.1330	0.2591
Matched hybrid	3	0.2/E	0.5477	0.0721	0.2121	1.8045
Matched hybrid	3	0.2/M	0.1549	0.0583	0.1427	0.3055
Old artifact	0	–/E	1.4383	0.2170	1.5415	2.2399
Old artifact	0	–/M	1.3083	0.1123	1.5140	2.3901
Old injection	0	–/E	1.4383	0.2170	1.5415	2.2399
Old artifact	3	–/E	1.4393	0.1749	1.5333	2.2490
Old artifact	3	–/M	1.3160	0.0872	1.5195	2.4167
Old injection	3	–/E	1.4393	0.1749	1.5333	2.2490

G Falsification Control: Dynamics-Aware Proxy and Selector Tests

This appendix documents the falsification control of Section 3.3: whether a deliberately strong local proxy can replace paired task evidence. The main text uses it to define the proxy, to summarize its local directional evidence in Table 1, and for one negative result, that the proxy's prospective uses miss their registered criteria (Appendix G.4). No qualification status and no fidelity–return result depends on it, so a reader interested only in those can skip it.

G.1 Definitions, endpoints, and mathematical scope

G.1.1 Endpoint Levels and Dynamics-Aware Proxy Contract

Related work on local diagnostics. Finite-horizon Gramians describe forced directional propagation (Zhou et al., 1996; Brockett, 2015), and finite-time Lyapunov exponents describe state-dependent growth (Krishna et al., 2023; Aurell et al., 1997). Our auxiliary proxy injects measured errors at every step, allowing cross-time reinforcement or cancellation. Adversarial perturbations likewise show controlled-norm direction effects (Moosavi-Dezfooli et al., 2018; Huang et al., 2017; Pattanaik et al., 2018), and smooth-policy regularization changes the policy to reduce sensitivity (Mysore et al., 2021); our error is measured rather than optimized and the policy is not retrained. Reachability tools analyze a neural controller with its plant (Ivanov et al., 2019; Xiang & Johnson, 2018; Tran et al., 2020; Wang et al., 2024); runtime methods monitor or predict violations (Xiang, 2022; Yu et al., 2025). Unlike these certificate-oriented methods, our local score only estimates amplification; paired task tests measure consequence. Adaptive conformal prediction has also been combined with incremental learning for OOD safety monitoring (Lin et al., 2025); calibrating our alerts would still calibrate only the surrogate.

Table 56: Strong-proxy stress-test ledger. Retrospective local information does not transfer to prospective decisions; rows are separate families.

Audit	Status	Main result	What it establishes
Norm-matched backend ID	Retrospective positive	CLAE − trace: 0.257 [0.163, 0.349]	Local directional information
Fresh alert prioritization	Prospective miss	CLAE − trace: 0.061 [−0.043, 0.156]	Gain not established
Calibration selection	Prospective miss	No task-loss ordering selected	No artifact decision
Unseen-policy trust gate	Prospective miss	Target met in 1/4 cells	No portable transfer
Ant direction control	Locked reversal	Control − measured: −327/ − 931 (EdgeTPU/Hailo)	Lower score, worse return

Figure 7 separates the three quantities used in the local audit. CLAE is a linearized score, the alert is a nonlinear short-horizon trajectory-separation proxy, and the adverse task outcome is a fall, termination, recovery failure, or return loss. Only paired task outcomes support qualification.

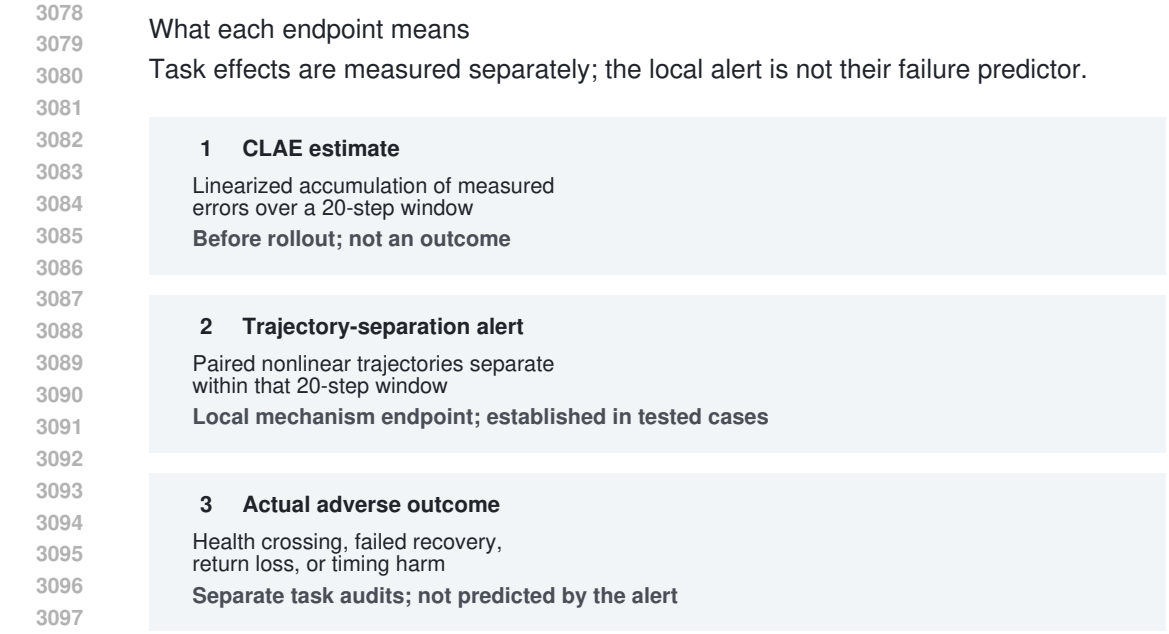

Figure 7: Three evidence levels. The 20-step alert is neither a fall nor a demonstrated predictor of task loss.

Statistical unit. The local effect is the within-pair alert-rate difference between measured and alignment-rotated fields, conditional on a fixed policy, nominal episode, and initial state. Windows are repeated workload units and the nominal episode is the inferential unit, so intervals resample complete episodes. Policy seeds are fixed cases, not a population sample.

Measured error and amplification score. For FP32 policy π_{fp}, compiled path π_b, and nominal state s_t, define raw and post-clip effective errors as

$$e_{b,t} = \pi_b(s_t) - \pi_{\mathrm{fp}}(s_t), \qquad e^{\mathrm{eff}}_{b,t} = \mathrm{clip}(\pi_{\mathrm{fp}}(s_t) + e_{b,t}, \mathcal{A}) - \mathrm{clip}(\pi_{\mathrm{fp}}(s_t), \mathcal{A}).$$

Open-loop baselines use $e_{b,t}$ alone. The closed-loop analysis uses $e^{\mathrm{eff}}_{b,t}$ and nominal finite-difference Jacobians A_t and B_t:

$$\delta s_{t+1} \approx A_t \delta s_t + B_t e^{\mathrm{eff}}_{b,t}, \qquad \Phi(j,i) = A_{j-1} \cdots A_i.$$

Writing the predicted terminal deviation as $\delta s_H = \sum_{k=0}^{H-1} \Phi(t_0 + H, t_0 + k + 1) B_{t_0+k} e^{\mathrm{eff}}_{b,t_0+k}$, coherent closed-loop amplification energy is

$$\mathrm{CLAE}^{\mathrm{coh}}_{b,t_0}(H; W) = \delta s_H^\top W \, \delta s_H, \qquad W \succeq 0.$$

Contributions are summed before squaring, so they may reinforce or cancel. Every result states its metric; the default $W = I$ mixes coordinates of different units, and changing it can reorder selected windows (Appendix G.5.7). The direction-blind trace comparator, $\|e^{\mathrm{eff}}_t\|_2^2 \, \mathrm{tr}(G_t^{(H)})/d_a$, keeps error size and average local gain but discards direction and cross-time interaction. Neither score is a failure probability.

CLAE is exact for a linear time-varying system and first-order local for a smooth nonlinear one only while the perturbed trajectory remains in a small nominal tube. It is not an $H = 20$ safety certificate: uniform derivative bounds are unestimated and the worst-case constant can be vacuous (Appendices G.1.3 and G.5.6).

Short-horizon alert. With d_{health} the signed margin to the Gymnasium health boundary (Farama Foundation, n.d.), an alert is raised at the first $k \le H$ where the nominal-to-injected state distance exceeds $d_{\mathrm{health}}(s^0_{t+k})$, or where the injected rollout crosses the boundary outright. The first

branch mixes position and velocity units and is a heuristic rather than a distance to falling; the second is an actual crossing. At natural measured-error scale ($\eta = 1$), every primary alert uses the heuristic branch and actual crossings are zero. The locked horizon is $H = 20$. Coordinate-normalized and Mahalanobis replays preserve the sign but change magnitude and path order, while strict height/angle-only alerts are nearly absent (Appendix G.5.5). The alert comes from nonlinear MuJoCo states, but score and intervention share a Gramian and target the same local amplification; this is a nonlinear mechanism stress test, not independent validation of the score.

Numerics, eligibility, and scope. Finite-difference Jacobians include policy, wrapper, action clip, and simulator. The horizon, margin bins, finite-difference steps, and cross-path thresholds were fixed before injection outcomes. The cross-path precheck (Gate 0 in the release files) asks only whether error directions are distinguishable; failure implies neither safety nor absence of a within-path effect. "Prospectively locked" means committed before the outcome run, not registered with an external service. Stable-Baselines3 PPO actors (Raffin et al., 2021; Schulman et al., 2017) were compiled for EdgeTPU and Hailo-8 (Google Coral, 2020a;b; Hailo, n.d.). The local set uses Walker2d seeds 5100, 5400, 5700, and 5800; Hopper-6100 is saturation stress. These are specified cases rather than a policy sample. Gate 0's low-cosine fraction correlates with effect size in a six-seed check (Pearson 0.670; Spearman 0.657), so headline magnitudes are selection-enriched. Appendix A.2 gives provenance; Appendices G.2.1 and G.2.1 summarize selection and its sensitivity, with full inventories in the separate technical record. Unless stated otherwise, local intervals use 5,000 episode-cluster bootstrap replicates.

G.1.2 Alignment-Rotation Construction Details

The local mechanism audit uses a deterministic coefficient-preserving eigenbasis permutation as the alignment-rotated intervention. For each nominal index t, the construction uses the effective error vector e_t^{eff} and the nominal single-pulse amplification Gramian $G_t^{(H)} = {M_t^{(H)}}^\top M_t^{(H)}$:

1. Eigendecompose the symmetrized Gramian: $G_t^{(H)} = V \Lambda V^\top$, with eigenvalues sorted from high to low. Eigenvector signs are canonicalized by making the first entry with absolute value above 10^{-12} positive.
2. Project the error into the nominal Gramian eigenbasis: $c = V^\top e_t^{\text{eff}}$.
3. Sort coefficient indices by decreasing $|c_i|$ and assign those coefficients to eigenvectors in increasing-gain order. Equivalently, form a permutation matrix P_t that maps the largest-magnitude coefficient onto the lowest-eigenvalue direction, the second largest onto the second lowest, and so on.
4. Return $e_t^{\text{rot}} = V P_t c$.

This is an anti-alignment control rather than a Haar-random rotation. Because the eigenvalues are sorted in decreasing order, assigning the largest squared coefficients to the smallest eigenvalues minimizes $\sum_i \lambda_i |c_{\sigma(i)}|^2$ over coefficient-preserving permutations, by the rearrangement inequality. Thus the control intentionally depletes alignment with the nominal high-gain eigendirections while preserving the construction error norm, coefficient signs, and coefficient multiset. This norm equality precedes feasible action re-clipping: the nonlinear runner applies $u_t^c = \text{clip}(\pi_{\text{fp32}}(s_t^c) + e_t^c, \mathcal{A})$. The applied perturbation relative to the clipped FP32 action at that same state is $d_t^c = u_t^c - \text{clip}(\pi_{\text{fp32}}(s_t^c), \mathcal{A})$; $\|e_t^{\text{rot}}\|_2 = \|e_t^{\text{eff}}\|_2$ does not imply equal $\|d_t^c\|_2$ across conditions, whose states can also diverge. We do not claim that it is the global minimum over all norm-preserving rotations; it is the minimum-energy member of this deterministic coefficient-permutation family under the single-pulse Gramian. Exact eigenvalue or coefficient-magnitude ties are resolved by the deterministic order returned by the eigensolver and sort routine. Eigenvalue ties do not change the Gramian energy within the tied eigenspace, and equal coefficient magnitudes give the same squared-energy assignment.

Jacobian, margin, and horizon locks. MuJoCo Jacobians use raw Gymnasium observation coordinates. Walker2d and Hopper omit global root-x; reconstruction sets that translationally invariant coordinate to zero and logs nominal-next-state error. The closed-loop observation Jacobian A_t uses central finite differences with step 10^{-4} through the frozen observation wrapper, policy, action clip,

and simulator step. The wrapper is the identity for the main tasks. The action-input Jacobian B_t uses step 10^{-4} in post-clip action coordinates, with an inward one-sided difference at action bounds. On the initial Walker2d panel, the A_t sweep at 10^{-5}–10^{-3} kept local-ρ median/p95 within 1.0239–1.0253/1.2354–1.2357. The matched A_t/B_t score audit preserves global ranking but misses its strict tail-stability rule (Table 75). Analytic LQR, pendulum, and cart-pole gates reproduce energy rankings with Spearman 1.0.

For $\rho_t = \max_i |\lambda_i(A_t)|$, locked bins are contractive ($\rho_t < 0.80$), intermediate ($0.80 \le \rho_t < 0.95$), near-marginal ($0.95 \le \rho_t \le 1.05$), mildly unstable ($1.05 < \rho_t \le 1.20$), and extreme unstable ($\rho_t > 1.20$). The primary horizon was selected before intervention outcomes from $H \in \{5, 8, 10, 12, 15, 20, 25, 30, 40, 50, 80\}$: the locked rule chose the smallest horizon with at least 1,000 reliable windows, median amplification in $[1.5, 2.0]$, and p95 in $[3, 8]$. $H = 20$ was first eligible (1693 windows; median 1.605; p95 5.431); $H = 25$ and 30 were also eligible. Sensitivity interventions use $H = 10$ and 40 without changing the primary horizon.

G.1.3 CLAE Exactness and Local Remainder

The following statements delimit the score's mathematical interpretation; they do not establish task prediction or qualification. The separate technical record preserves the original proofs, linear construct check, and illustrative worst-case constant. The linear check is positive evidence where the model is exact, not evidence that the nonlinear task lies in that regime.

Proposition 1 (Linear exactness)**.** *For the corresponding linear time-varying system, coherent CLAE equals the exact squared terminal deviation.*

Proposition 2 (Local nonlinear remainder)**.** *Let $g_k(s, e)$ be the post-clip closed-loop map. On tubes $\|s - s_k^0\|_2 \le r_s$, $\|e\|_2 \le r_e$, assume $\|Dg_k\|_{\mathrm{op}} \le \Gamma$ and that Dg_k is L-Lipschitz in the product norm. Define $q_0 = 0$ and $q_{k+1} = \Gamma(q_k + 1)$. If $\max_k \|e_k^{\mathrm{eff}}\|_2 \le \epsilon$, $\epsilon \le r_e$, and $q_k \epsilon \le r_s$ for every $k \le H$, so the nominal-to-perturbed segments remain in these tubes, then the terminal nonlinear-to-linear error is at most $C_H(L, \Gamma)\epsilon^2$.*

Proof outline. For the linear system, unrolling the perturbation recursion gives the sum defining δs_H in Appendix G.1.1; its quadratic energy is coherent CLAE. For the nonlinear system, use the product norm $\|(z, e)\| = \|z\|_2 + \|e\|_2$, where $z_k = s_k^e - s_k^0$ and $z_0 = 0$. Induction and the derivative bound keep $\|z_k\|_2 \le q_k \epsilon$ inside the assumed tube. Taylor expansion gives $z_{k+1} = A_k z_k + B_k e_k + r_k$, with $\|r_k\|_2 \le (L/2)(q_k + 1)^2 \epsilon^2$. Unrolling the remainder recursion with $\|A_k\|_2 \le \Gamma$ yields

$$\|z_H - z_H^{\mathrm{lin}}\|_2 \le C_H(L, \Gamma)\epsilon^2, \quad C_H(L, \Gamma) = \frac{L}{2} \sum_{k=0}^{H-1} \Gamma^{H-1-k}(q_k + 1)^2.$$

Limits of the bound. The smooth local model requires the stated tubes and a fixed post-clip active regime; a clipping-boundary change can violate the assumptions. Construction norm matching does not ensure equal applied norms after re-clipping (Appendix G.1.2). We estimate neither a uniform operator bound Γ nor the derivative Lipschitz constant L on these tubes. The local spectral radius is not such an operator bound and cannot be substituted into the proposition. Even the original illustrative substitution gives a large worst-case constant, not a useful numerical certificate. Appendix G.5.6 reports empirical fidelity and its failures; neither the proposition nor agreement on selected local windows licenses a transferable task-qualification rule.

G.2 Eligibility and Policy-Selection Limits

G.2.1 Selection and Claim Scope

Gate 0 tests whether backend error directions are distinguishable enough for a comparative geometry claim. It is neither a safety screen nor the qualification protocol's provenance Stage 0. It does not test whether an individual backend can amplify errors.

Under the recorded repeatability/quantization-floor filtering, an excluded fraction above 0.5 makes the comparison insufficient. Otherwise a median cosine below 0.4 or an IQR containing zero gives a strong pass; a median below 0.7 or a cosine-below-0.3 fraction at least 0.25 gives a pass. The remaining cases are no-go for this comparison.

Table 57: Original selection denominator, recomputed from filtered pair metrics. These are direction-diversity decisions, not task qualifications.

Policy	Gate 0	Median cosine	Fraction cosine < 0.3
Walker2d-5100	Strong pass	-0.000	0.668
Walker2d-5200	No-go	0.832	0.205
Walker2d-5300	No-go	0.852	0.167
Walker2d-5400	Weak pass	0.727	0.259
Walker2d-5500	No-go	0.912	0.122
Walker2d-5600	No-go	0.908	0.217
Hopper-6100	Strong pass	0.029	0.614
Hopper-6200	No-go	0.910	0.173
Hopper-6300	No-go	0.982	0.114

The original 9-policy denominator contains 2 strong passes, 1 weak pass, and 6 no-go cases (Table 57). The bounded Ant sweep, fresh Ant holdout, Humanoid inventory, and later targeted strong-pass Walker2d replications do not enlarge that denominator.

Sensitivity and excluded settings. The local threshold sweep leaves the strong and no-go classifications unchanged at the registered low-cosine cutoff, while the weak pass is threshold-adjacent. Relaxing that cutoff can promote some original no-go settings; that stress analysis does not redefine eligibility. Capped no-go diagnostics nevertheless show positive within-backend measured-minus-rotated effects. Their smaller panels are not eligible cross-backend replications. Gate 0 is not a monotone effect-size predictor; selecting on directional separation enriches the reported cases rather than defining a representative policy sample.

Capability and replication limits. Humanoid provides a boundary case, not a clean mechanism replication: directional diversity can coexist with action saturation and an empty near-margin pilot population. A separate numerically accepted checkpoint failed visual promotion. Neither is counted as a clean replication. The later strong-pass Walker2d replications and the weak-pass case remain in the technical record's quantitative inventory; their estimands are paired within-panel contrasts, not comparisons of raw prevalence across policies. Hopper is retained as a clipping/saturation stress case. These local-policy limitations do not change the separately capability-screened task cohorts.

G.2.2 Ant fresh-holdout audit

The prospective Ant-8800 audit used 40 fresh reset seeds for the mechanism panel and a disjoint 100-episode block for the FP32 lifetime audit. The earlier Ant pilot was excluded from every estimate. The FP32 termination rate was 0.03, mean episode length was 972.24, and mean return was only 774.59. This checkpoint demonstrates long-horizon survival, not locomotion competence. The fresh panel contained 3744 states from 39 qualifying episodes and 2964 valid $H = 20$ windows. The registered near-margin population retained 831 windows from 27 episodes; the mildly-unstable secondary retained 73 windows from 28 episodes. Table 58 reports the primary and secondary endpoint rows separately.

Table 58: Prospective Ant-8800 fresh-holdout outcomes. Intervals use 5,000 nominal-episode cluster-bootstrap replicates. Alert delta is real minus rotated. Restricted-time delta is rotated minus real, so positive values mean earlier events under the real measured field. Health-margin delta is rotated minus real minimum signed margin, so positive values mean greater margin consumption under the real field. The mildly-unstable rows are descriptive and do not replace the registered near-margin primary.

Backend	States/episodes	Real/rotated alert	Alert delta [95% CI]
Near-margin primary			
EdgeTPU	831/27	0.0120/0.0096	0.0024 [0.0000, 0.0080]
Hailo	831/27	0.0084/0.0072	0.0012 [0.0000, 0.0038]
Mildly unstable, descriptive			
EdgeTPU	73/28	0.8493/0.7123	0.1370 [0.0263, 0.2903]
Hailo	73/28	0.7945/0.7123	0.0822 [0.0135, 0.1754]

Backend	Time delta [95% CI]	Margin delta [95% CI]
Near-margin primary		
EdgeTPU	0.042 [−0.008, 0.121]	−0.000004 [−0.000376, 0.000396]
Hailo	0.032 [0.004, 0.077]	−0.000186 [−0.000550, 0.000007]
Mildly unstable, descriptive		
EdgeTPU	3.123 [1.354, 5.436]	−0.00222 [−0.01180, 0.00844]
Hailo	2.507 [1.235, 3.978]	0.00716 [−0.00926, 0.02367]

All pre-outcome gates passed. Gate 0 had median EdgeTPU–Hailo cosine -0.1447, IQR $[-0.3981, 0.2306]$, and a cosine-below-0.3 fraction of 0.8186. The applied rotated/real norm ratios were 1.000004 for EdgeTPU and 0.999997 for Hailo. Nevertheless, the near-margin alert endpoint was sparse: EdgeTPU had 10/831 real and 8/831 rotated events, while Hailo had 7/831 and 6/831. Both primary alert intervals include zero, and neither continuous health-margin row establishes the hypothesized effect. The mildly-unstable alert and timing contrasts are positive, but their rotated event prevalence is already 0.7123; they are regime-limited secondary evidence rather than a clean-task replication. The locked no-extension rule was followed.

Canonical values use the corrected nominal-episode bootstrap and injected health-margin replay. The original state-bootstrap and nominal-margin outputs are retained in the release bundle; neither correction changed the binary primary endpoint or its verdict.

The capability-boundary figure, full Gate-0 and eligible-replication matrices, Humanoid inventory, and threshold/no-go diagnostics remain in the separate auxiliary technical record. Their original populations, positive and negative results, and registered statuses are unchanged.

G.2.3 FP32 Policy Lifetime and Reporting Limits

In the earlier full-episode backend-runtime audit, every observed FP32 episode reached the health boundary within the registered 1,000-step cap (Table 59). The original registry excludes a policy from pooled deployment claims when its FP32 termination rate exceeds 0.10. All 5 audited policies exceed that threshold; no Walker2d seed remains eligible for the registered pooling. Per-row contrasts and arithmetic pooled summaries remain descriptive. Both FP32 and backend arms have saturated binary incidence: a zero paired incidence difference is not evidence of safety or equivalence.

Table 59: FP32 lifetime evidence retained for the reporting downgrade. Counts use unique paired reset seeds, not duplicated baseline rows across backends. Detailed return and lifetime distributions remain in the technical record.

Policy	Health-boundary events / episodes	Median time (steps)
Walker2d-5100	100/100	205.0
Walker2d-5400	100/100	190.0
Walker2d-5700	100/100	206.5
Walker2d-5800	100/100	224.0
Hopper-6100	100/100	297.5

The later Humanoid capability/visual gate is not retroactively imposed on these earlier Walker2d/Hopper panels. That later search asked for a clean high-dimensional mechanism case; a healthier policy could still provide no near-margin pilot population (Appendix G.2.1). This lifetime audit leaves the paired local mechanism estimand unchanged and does not reclassify the separately capability-screened task cohorts.

What the remaining diagnostics mean. Persistent full-episode backend exposure and the nominal-boundary-censored post-window recoverability audit are different experiments. Binary incidence can be saturated in the former while the latter remains non-identifying. Restricted mean survival time (RMST) and time-to-boundary are secondary diagnostics relative to an already fragile FP32 baseline, not safety certificates.

Return and survival time are coupled in these tasks. The technical record preserves the complete paired timing-implied return, residual, and per-step decomposition; it does not treat them as independent confirmation. The RMST-implied component uses each paired FP32 return/time rate, and per-step differences are computed within episodes rather than by dividing mean residuals by mean duration. Observed timing shifts have both signs; longer survival under an artifact is not declared a safety improvement or assigned a post hoc cause.

The descriptive mechanism-to-deployment rank links also remain in the technical record. Their permutation analysis preserves both backend rows within each policy block rather than treating those rows as independent. The positive alert-prevalence association remains a regime diagnostic, not a mechanism proof. All original tables, positive and negative rows, and historical interpretation limits are preserved there.

G.3 Local directional and retrospective ranking evidence

G.3.1 Norm-Matched Cross-Backend Discrimination

The paired AUC-delta analysis asks whether coherent CLAE is a better within-backend screen for trajectory-separation alerts. The cross-backend discrimination audit asks a different stored-data question: for the same nominal window, can a score identify which backend produced the event when EdgeTPU and Hailo have similar post-clip effective error norms? We use the registered primary norm-ratio band $\|e_{\mathrm{Edge}}\|/\|e_{\mathrm{Hailo}}\| \in [0.8, 1.25]$ and report the stricter $[0.9, 1.1]$ band as sensitivity. Accuracy is computed only on discordant windows where exactly one backend has an alert; ties count as 0.5. Intervals use 5,000 episode-cluster bootstrap replicates. Table 60 gives the per-seed coverage, discordant counts, and pooled contrast used for this claim.

This audit satisfies the registered E3b success rule: at least two adequately powered Walker2d rows have CLAE accuracy intervals above 0.5, and the pooled paired CLAE-minus-trace accuracy interval is above zero. It remains a stored-alert audit rather than an actual-crossing result, so it supports the screening-score mechanism claim without changing the alert-endpoint limitation. Here "adequately powered" denotes the registered state/episode coverage gate, not a formal power analysis. The Walker2d-5100 row has a degenerate 1.000 [1.000, 1.000] CLAE interval because the discordant subset is small and all 11 discordant cases are won by CLAE. The pooled delta, 0.257 [0.163, 0.349], is less sensitive to that small discordant subset. The pooled primary-band row includes the underpowered Walker2d-5700 discordant cases in the pooled bootstrap; the registered powered-row count still excludes that row. The 0.8–1.25 band also leaves some residual norm information, as reflected by trace accuracy above chance in Walker2d-5100. The exact-tie statement is an idealization

Table 60: Stored-data cross-backend discrimination under near-matched backend error norms. Trace-isotropic scores can only use residual norm differences at a fixed state; under exact norm matching they tie by construction. CLAE-terminal ρ is the Spearman correlation, over all norm-band states, between the EdgeTPU-minus-Hailo coherent CLAE score difference and the EdgeTPU-minus-Hailo terminal observation-deviation difference. The registered powered label requires at least 150 norm-band states and eight episodes; it is a coverage rule, not a formal power calculation or a minimum discordant-count requirement.

A: Coverage and score–outcome association

Panel	States / episodes	Discordant	CLAE-terminal ρ	Powered
Walker2d-5100	545 / 20	11	0.625	yes
Walker2d-5400	799 / 22	55	0.472	yes
Walker2d-5700	72 / 11	24	0.209	no
Walker2d-5800	178 / 9	46	0.385	yes
Pooled primary band	–	136	–	3/3 seed rows
Pooled $[0.9, 1.1]$ sensitivity	–	78	–	2/2 seed rows

B: Accuracy and paired advantage

Panel	CLAE accuracy	Trace accuracy	CLAE − trace
Walker2d-5100	1.000 [1.000, 1.000]	0.727 [0.444, 0.938]	0.273 [0.062, 0.556]
Walker2d-5400	0.691 [0.551, 0.807]	0.527 [0.448, 0.612]	0.164 [0.000, 0.294]
Walker2d-5700	0.875 [0.793, 1.000]	0.500 [0.294, 0.650]	0.375 [0.200, 0.688]
Walker2d-5800	0.761 [0.632, 0.864]	0.457 [0.341, 0.548]	0.304 [0.143, 0.467]
Pooled primary band	–	–	0.257 [0.163, 0.349]
Pooled $[0.9, 1.1]$ sensitivity	–	–	0.269 [0.147, 0.392]

for exact norm matching; the pooled advantage remains positive in the stricter 0.9–1.1 sensitivity band.

The norm band matches stored post-clip effective error norms at a fixed state. Appendix G.3.3 summarizes why matching stored norms does not match executed trajectories. Its separate-harness controls change direction and scale together, so they limit the interpretation of matching without isolating direction or re-adjudicating this stored-backend comparison.

G.3.2 FETCH GOALENV GOAL-DEVIATION AUDIT

We ran a bounded GoalEnv audit only as a protocol-portability check, using finite-horizon achieved-goal deviation and terminal goal-distance degradation rather than health-boundary or fall-risk labels. FetchPush uses the rebuilt standardized-input backend contract after the earlier folded-standardization dumps failed runtime fidelity. Table 61 keeps the result at summary level: FetchPush is positive on achieved-goal deviation for both backends, while terminal goal-distance degradation is clear for Hailo and weak/null for EdgeTPU; these rows are not used as main task-level replication evidence.

FetchPush is the only contact-manipulation row with a well-matched applied-norm audit, but its endpoint evidence is mixed across goal endpoints. This appendix audit is intentionally separate from the main Walker2d/Hopper health-divergence evidence. FetchSlide is omitted from this compact table because its alignment-rotated/real norm ratio was only 0.85, outside the intended norm-matching regime; the archived exploratory outputs remain in the release bundle and are not used as evidence.

G.3.3 WALKER2D CONTROL EVIDENCE AND SCOPE

The larger-panel replication and the full control ladder are preserved in the separate auxiliary technical record, with their original tables and figures. The replication revisits existing eligible policies, not a new policy cohort. Here we retain the comparisons that qualify the interpretation of the local proxy: deliberate low-gain redirection is not a neutral random baseline, and matching an error norm does not match a closed-loop trajectory.

Table 61: Compact Fetch GoalEnv protocol-portability audit. Values are paired real-minus-alignment-rotated deltas. Reach uses the earlier $0.01\,\mathrm{m}$ threshold summary; Push uses the standardized-input backend contract at $\tau = 0.02\,\mathrm{m}$. This is a goal-deviation audit, not a health-boundary divergence table.

A: Protocol and applied error norm

Task	Backend	H/n	Event τ	Rot./real norm
FetchReach	EdgeTPU	10/196	0.01 m	0.9996
FetchReach	Hailo	10/196	0.01 m	0.9943
FetchPush	EdgeTPU	10/519	0.02 m	0.9798
FetchPush	Hailo	10/519	0.02 m	0.9683

B: Paired outcome differences

Task / backend	Achieved-goal event Δ	Goal-distance event Δ	Max achieved-goal Δ
FetchReach / EdgeTPU	0.000	0.000	0.0018 [0.0015, 0.0021]
FetchReach / Hailo	0.362	0.296	0.0092 [0.0051, 0.0135]
FetchPush / EdgeTPU	0.046 [0.004, 0.088]	0.004 [-0.029, 0.034]	0.0035 [0.0016, 0.0051]
FetchPush / Hailo	0.069 [0.024, 0.114]	0.050 [0.023, 0.075]	0.0071 [0.0024, 0.0124]

Interpretation: FetchReach/EdgeTPU, weak continuous sanity; FetchReach/Hailo, positive no-contact sanity; FetchPush/EdgeTPU, achieved-goal positive, terminal weak/null; FetchPush/Hailo, positive on both endpoints.

Fixed Construction and Multi-Construction Sensitivity The alert comparisons in this subsection use the same Walker2d panel and short-horizon trajectory-separation alert, not a task-failure endpoint. Table 62 retains the basis of the main-text percentages. For the fixed seeded eigenbasis permutation, we divide the random-minus-anti-aligned alert-rate difference by the measured-minus-anti-aligned difference. This is a descriptive ratio of contrasts, not a proportion of task loss causally explained by randomness.

Table 62: Fixed-construction basis of the descriptive contrast shares. The ratio uses unrounded paired alert rates; it is not the multi-construction estimate in Table 63.

Backend	Measured−anti-aligned	Random−anti-aligned	Ratio
EdgeTPU	0.263	0.224	85.1%
Hailo	0.080	0.062	77.2%

Table 63 answers a separate question. It repeats the random construction at 5 seeds and integrates construction and nominal-episode uncertainty. The panel contains 1429 paired windows from 10 episodes, with horizon $H = 20$. These post-outcome measured-minus-random contrasts must not be substituted into the fixed-construction ratio above. The raw-alert difference is small; the Hailo normalized and Mahalanobis intervals favor measured over random errors, while the corresponding EdgeTPU intervals include zero. Health-only contrasts vanish on this panel. The comparison therefore does not establish a general adverse ordering of measured errors over random controls.

Table 63: Five-construction measured-minus-random permutation alert contrasts on the Walker2d-5100 panel. Intervals use 5,000 hierarchical construction-draw and nominal-episode bootstrap replicates. These are descriptive endpoint sensitivities, not a prospectively locked family.

Endpoint	EdgeTPU contrast [95% CI]	Hailo contrast [95% CI]
Raw full L2	0.037 [0.018, 0.057]	0.019 [0.013, 0.026]
Coordinate-normalized L2	0.004 [−0.010, 0.019]	−0.099 [−0.132, −0.069]
Mahalanobis	0.005 [−0.013, 0.025]	−0.100 [−0.143, −0.061]
Health only	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]

Input Matching Is Not Execution Matching A separate-harness closed-loop check replaced each recorded error vector by a random direction proportional to that step's own norm. Without rescaling, these rollouts had a larger action gap from unperturbed FP32 than the backend rollouts. Matching the realized gap required reducing the global scale. The intervention thus changes direction and scale together and does not isolate a causal effect of direction. Its Ant and out-of-family HalfCheetah results are limitations of the matching criterion, not replications within the registered Walker2d families or revisions of their verdicts.

The full execution table and calibrated isotropic-noise comparison are retained in the technical record, including the negative device-specificity result. They do not establish a chip-specific signature. Construction norms, post-clipping applied norms, and trajectory-level gaps are distinct quantities; the input-level control comparisons above should not be read as effect-matched rollout experiments. The source experiments, selection rules, and registered statuses are unchanged by this editorial compression.

G.3.4 Screening Ranking Utility

Retrospective ranking evidence. On stored real-condition alert labels, paired CLAE-minus-trace AUC intervals are wholly positive in 3/4 primary task/backend contrasts; the remaining primary interval includes zero. The supporting and descriptive rows also include a negative point estimate and intervals spanning zero, so the increment over trace is not uniform. The paired intervals resample nominal episodes rather than treating their windows as independent. The complete family, including Holm-adjusted tests and the supporting/descriptive roles, remains in the separate technical record.

In the earlier, separate ranking panels, CLAE exceeds the better of action MSE and maximum action error in PR-AUC in 14/14 panels (margins 0.039–0.333). This compares with open-loop errors, not with trace-isotropic ranking. These panels and the paired AUC audit do not have identical input populations and are not pooled. PR-AUC depends on event prevalence; auxiliary top-budget recall and NDCG can saturate or tie in high-prevalence panels. Neither positive ranking performance nor ties establish calibrated probabilities or task qualification.

The technical record also retains the complete capped runtime comparison and its descriptive CLAE-decile diagnostic, not a general monotonicity claim. The fresh fixed-budget audit in Appendix G.4.1 tests whether the local ranking advantage persists under outcome-blind selection; it remains distinct from task-level qualification.

G.3.5 Exploratory calibration-axis evidence

We first varied only the PTQ calibration set for a fixed Walker2d-5100 policy, backend, precision, graph, and evaluation population. The reference, held-out episode, near-margin, and broad-random variants were fixed before new outcome labels were inspected, but the injection follow-up remained exploratory. EdgeTPU's held-out variant reproduces the reference TFLite export and measured output dumps, not its compiled file bytes. This is a negative control within the tested artifact–panel cells, not a general invariance claim.

On the fixed-state error panel, Hailo's broad-random variant has lower mean FP32 error norm than the reference (0.0625 versus 0.0807). On the eligible injection subset, it has higher median CLAE, measured alert fraction (0.568 versus 0.407), and real-minus-rotated alert difference (0.330 versus 0.080). These are local trajectory-separation alerts, not task-return losses. The within-artifact real-minus-rotated contrasts have positive episode-cluster intervals, but no between-variant inferential family was locked; the ordering across artifacts is descriptive.

Across these Hailo variants, marginal median CLAE and measured alert fraction share the ordering held-out $>$ broad-random $>$ near-margin $>$ reference. Yet the near-margin variant's median paired CLAE ratio is below the reference value even though its marginal median is higher. The median of statewise ratios and the ratio of marginal medians answer different distributional questions.

The complete geometry, ranking, and exploratory injection tables remain in the separate technical record. These are naturally compiled calibration variants, not synthetic rotations, but calibration changes direction together with other quantizer details. The result shows that scalar error norm alone does not determine the measured field or local alert rate; it neither isolates a causal mechanism nor

establishes a general calibration-selection rule. The prospective selection test and later fixed-contrast replication in Appendix G.4.2 address different, explicitly separated questions.

G.4 Prospective selector families

Three prospective CLAE families—alert prioritization, calibration selection, unseen-policy trust gating—miss registered criteria (Appendix G.4). Alerts are heuristic trajectory separations, not falls. Random-direction comparisons are descriptive local controls, not explanations of task loss (Table 62). Ant signed-permutation controls have lower CLAE but worse return than measured fields: constructed interventions, not naturally compiled artifacts (Appendix G.6).

Table 56 summarizes the local evidence and the three prospective selector families reported here.

G.4.1 Fresh fixed-budget hold-out audit

We directly tested the proposed local-audit use on a fresh Walker2d-5400 panel. Before generating injected outcomes, we fixed 4,096 nominal states, the reliable near-marginal population, both frozen hardware artifacts, and a budget of 256 windows per method and backend. The outcome-blind lock retained 2,257 common candidates over 22 episodes, then committed the complete CLAE, trace-isotropic, action-MSE, and SHA-ranked random selections. No panel or budget extension was permitted after outcomes. Table 64 gives the absolute 256-window yields, and Table 65 gives the locked paired comparisons.

Table 64: Raw full-state alert yield among the 256 freshly selected windows per method and backend. Random selection is descriptive. Every selected rollout is valid, and no cell contains an actual health-boundary crossing.

Backend	CLAE	Trace-isotropic	Action MSE	Random	Crossings
EdgeTPU	1.000	0.930	0.727	0.328	0
Hailo	1.000	0.949	0.797	0.316	0

Table 65: Equal-backend pooled CLAE yield gains on the fresh fixed-budget hold-out. Intervals resample the 22 nominal episodes jointly across methods and backends. Holm p-values apply only to the two locked raw-endpoint comparisons.

Endpoint	CLAE − action MSE [95% CI]	CLAE − trace [95% CI]	Locked result
Raw full L2	0.238 [0.158, 0.320], $p_{\mathrm{Holm}} < 0.001$	0.061 [−0.043, 0.156], $p_{\mathrm{Holm}} = 0.264$	family miss
Coordinate-normalized L2	0.301 [0.225, 0.373]	0.006 [−0.033, 0.045]	sensitivity
Mahalanobis	0.371 [0.289, 0.455]	0.098 [0.037, 0.158]	sensitivity
Health only	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	no events

CLAE clearly improves yield over action MSE, and the sign is positive on both backends. It does not establish an increment over trace-isotropic ranking: the raw interval crosses zero, while CLAE and trace both approach ceiling. The normalized CLAE-minus-trace point estimates also differ in sign across backends (−0.035/0.047), although their equal-backend mean is slightly positive. In contrast, the post-outcome Mahalanobis sensitivity is positive, 0.098 [0.037, 0.158]. This predictor comparison is distinct from the measured-versus-random intervention contrast: a dimension-aware endpoint can retain direction-aware ranking while reversing the Hailo intervention ordering. Because the prospective rule required both primary contrasts, this audit misses and is not promoted as validated prioritization utility. It tests local trajectory-separation alerts, not deployment decisions or adverse outcomes.

G.4.2 PROSPECTIVE FIXED-FAMILY SELECTION AND INDEPENDENT REPLICATION

We then locked 12 Hailo artifacts and 20 fresh nominal episodes. Episodes 0–9 supplied screening scores; episodes 10–19 supplied validation outcomes. The split, artifact family, score, selection rule, outcome, two-test Holm family, and prevalence guard were fixed before validation rollout. Calib-C had the lowest screening median CLAE and was committed as the selected switch before any validation outcome was generated. Table 66 reports the complete fixed family.

Table 66: Prospective Hailo calibration-family follow-up on Walker2d-5100. Screening medians use 1373 windows from 10 episodes; every validation rate uses the same 1396 starts from 10 held-out episodes. Calib-C was selected before validation outcomes.

Artifact	Screening median CLAE	Validation alerts / 1396
Calib-A reference	0.003328	558 (0.400)
Calib-B held-out	0.001777	517 (0.370)
Calib-C near-margin (selected)	0.001178	512 (0.367)
Calib-D broad random	0.001959	559 (0.400)
V1 episode block 1	0.003369	558 (0.400)
V2 episode block 2	0.002029	544 (0.390)
V3 episode block 3	0.002847	571 (0.409)
V4 episode block 4	0.003748	599 (0.429)
V5 bootstrap 1	0.002207	551 (0.395)
V6 bootstrap 2	0.003660	550 (0.394)
V7 lowest health margin	0.001736	553 (0.396)
V8 broad random P-4096	0.002714	596 (0.427)

Across the fixed family, screening median CLAE and validation alert fraction have Spearman $\rho = 0.564$ (permutation $p = 0.060$, Holm $p = 0.120$). The selected-minus-reference alert-rate difference is -0.033, with episode-cluster CI $[-0.070, -0.004]$, but its exact ten-episode sign-flip test gives $p = 0.070$ and Holm $p = 0.120$. The interval weights windows within resampled episodes, whereas the exact test uses paired episode signs. Reference prevalence is 0.400 and all 1396 starts are common-valid. Neither registered test passes, so E8 does not confirm screening-guided artifact selection.

As a secondary mechanism check, real-minus-rotated event deltas are 0.284 for Calib-A, 0.277 for selected Calib-C, and 0.241 for maximum-screening-CLAE V4. These positive contrasts replicate the within-artifact direction effect on the fresh panel but do not alter the E8 selection decision.

After inspecting that miss, we prospectively locked a separate independent-episode replication of only the selected Calib-C versus Calib-A contrast. The artifacts and selection were unchanged; 30 SHA-derived reset seeds were disjoint from E8 and earlier panels, and no extension or replacement was allowed. Before rollout, all 30 episodes supplied 3971 eligible windows. The new-panel CLAE medians again favored Calib-C (0.001101 versus 0.003348) although its mean effective error norm was larger (0.0342 versus 0.0297). Table 67 keeps the original family decision separate from this independent replication.

Table 67: Held-out and independent-episode results for the fixed Calib-C versus Calib-A contrast. The original selection-test interval is 95% and belongs to its two-test Holm family. The later replication uses its prospectively locked 97.5% interval and $\alpha = 0.025$; it does not change the original selection-family decision.

Panel	Episodes	Calib-A	Calib-C	C−A	CI; exact p
E8 held-out	10	0.400	0.367	-0.033	[-0.070, -0.004]; 0.070
Independent replication	30	0.406	0.381	-0.026	[-0.048, -0.004]; 0.015

The replication passes its support and prevalence guards, CI rule, and exact episode sign-flip threshold; 20 episode contrasts are negative, six positive, and four zero. It supports repeatability of the already selected contrast on new episodes. It neither confirms the original artifact-family rank endpoint nor establishes a general calibration-selection rule.

G.4.3 PROSPECTIVE UNSEEN-POLICY TRUST GATE

The metric construction is defined in Appendix G.5.7.

Prospective trust-gate test. Conditioning the fidelity audit of Appendix G.5.6 on no-alert rows requires the outcome. We therefore precommitted an outcome-blind trust gate and tested it on two policy seeds that were never inspected while the specification was chosen. Walker2d-5100 and Walker2d-5400 were the development panels; Walker2d-5700 and Walker2d-5800 carried the inference. The gate retains a window when a linear excursion score — the largest normalized predicted deviation reached inside the window — stays below an absolute threshold frozen at the pooled development median over 11,074 windows. The primary endpoint was the median relative error between $\mathrm{CLAE}^{\mathrm{coh}}(H;W)$ and the observed nonlinear terminal deviation measured in the same W, with the health-coordinate metric primary. The locked rule required a median relative error at most 0.10 in all four confirmatory cells and an improvement over the ungated value in all four. The configuration, thresholds, endpoint, and success rule were committed before any confirmatory outcome existed.

Table 68: Prospective trust-gate outcome on never-inspected policy seeds, primary health-coordinate metric. Entries are the median relative error between predicted and observed terminal deviation, with retained coverage. The no-alert column is outcome-dependent and descriptive.

Panel	Backend	Ungated	Trust-gated (coverage)	No-alert (coverage)
Walker2d-5700	EdgeTPU	2146.1	0.429 (0.130)	0.464 (0.145)
Walker2d-5700	Hailo	241.5	0.235 (0.217)	0.336 (0.248)
Walker2d-5800	EdgeTPU	1.19	0.168 (0.177)	0.159 (0.318)
Walker2d-5800	Hailo	13.87	0.096 (0.148)	0.104 (0.354)

The family misses (Table 68). Gating improves fidelity in all four cells, but only one cell reaches the precommitted 0.10 target. The frozen threshold also retains just 13–22% of windows instead of the roughly 50% seen in development, because the confirmatory panels are far more nonlinear: their measured-field alert fractions are 0.752–0.855 and 0.646–0.682, versus 0.443–0.464 and 0.293–0.312 in development.[2] A secondary Jacobian-drift gate does not transfer at all, retaining no Walker2d-5700 window at its tighter threshold. No threshold, gate, metric, panel, or endpoint was substituted after the outcome was read.

Two things follow. Ungated coherent CLAE in a task-relevant metric is not usable on high-prevalence panels, where its median relative error reaches three orders of magnitude. And an outcome-blind operating region does capture most of what the outcome-dependent no-alert conditioning captures, but not at a transferable accuracy level or coverage. Metric-aware, trust-gated CLAE is therefore not a validated local predictor with a portable operating point. This is the third precommitted local-score test in this paper to miss, alongside the fixed-budget audit of Appendix G.4.1 and the calibration-selection family of Appendix G.4.2.

The Auxiliary Technical Record, §9 (Numerical Robustness and Fidelity), preserves the complete original numerical audit, including all construction-seed, linearization, finite-difference, metric-estimation and trust-gate results. The registered statuses are unchanged.

G.5 RECOVERY, CLIPPING, NOISE, METRIC, AND NUMERICAL LIMITATIONS

The local-alert, recoverability, and sustained task-execution audits answer different questions. Table 69 keeps their distinct statuses; the complete original tables, health-normal diagnostics, and figure are preserved in the Auxiliary Technical Record, §7 (Outcome and Recoverability Results). The unforced Ant task-return audit (Appendix C.5) and constructed direction control (Appendix G.6) retain their original designs and results.

[2] The E24 fractions use every valid real-condition window with finite predicted and observed terminal vectors. The archived linearization table instead restricts to the near-marginal population with complete condition and paired censoring support.

G.5.1 Recoverability and Boundary Audits

The initial audit continued under FP32 for 180 steps after an $H = 20$ measured-error injection. Asymmetric nominal-boundary censoring and saturation made its binary alert-to-termination analysis uninterpretable, not evidence of equivalence. The separately locked follow-up used a common censor and a continuous signed health-margin endpoint. Its full-population Walker2d criterion misses; the lower-margin sensitivity did not replace that primary. Row signs vary, and the full-population decile audit does not establish the proposed predictive link.

The separately locked near-boundary audit selected low positive-margin, horizon-valid starts. Its binary primary fails the prevalence gate, while its continuous secondary resolves in the reverse direction. The post-outcome health-normal diagnostic also finds opposite signs for nominal first-order prediction and nonlinear signed margin. These are different limitations, not a pooled null result or a validated alert-to-fall bridge.

Table 69: Recovery-audit status summary. Continuous margins are rotated minus real; positive is the hypothesized direction. Intervals retain their original episode-cluster analyses; these rows are not one inferential family.

Audit	Evidence	Status
Initial continuation	Asymmetric censoring and saturated termination	Binary link uninterpretable
Common-censor follow-up	$\Delta = -0.000886$ $[-0.002308, 0.000578]$; 3/8 positive row estimates	Registered miss
Near-boundary primary	Actual rates real/rotated: 0.0033/0.0055; margin $\Delta = -0.00066$ $[-0.00087, -0.00040]$	Binary non-identifying; reverse continuous secondary

G.5.2 Ant Prechecks and Stress-Conditioned Evidence

The Ant policy passed its FP32 capability gate. Its health-normal-preserving control construction then failed the locked post-clipping and energy-separation gates and stopped before held-out outcomes. That construction stop is not an adverse-outcome null; the subsequent total-artifact audits use different estimands. Capability, unforced-return, and direction-control blocks use disjoint episodes, so their FP32 baseline means need not coincide.

Stress-conditioned total-artifact audit. The online stress audit restarted each arm from the same post-pulse state. Both backend intervals satisfy the registered $(-5, +5)$-step RMST absence rule (Table 70). This is a narrow bound for the selected policy and force, not backend equivalence or an unforced prevalence estimate.

Table 70: Ant stress-conditioned online RMST, FP32 minus backend. Each original paired interval lies within the registered absence band; this is not an unforced or direction-only contrast.

Backend	Paired RMST difference	95% interval
EdgeTPU	2.20	$[-0.41, 4.79]$
Hailo	-2.03	$[-4.98, 0.90]$

The prospectively selected score comparison remains a separate secondary: EdgeTPU is inconclusive, while Hailo has a positive backend-specific interval. Its small method-specific support and lack of a pre-specified adjusted two-backend secondary family limit that result; it does not validate a deployment selector. The full selection table and stopped prechecks, including the separately inventoried pre-outcome stop, remain in the technical record.

These checks limit the interpretation of the local proxy. The registered noise gate misses, clipping prevents a clean Hopper replication, and the main-panel alerts do not represent actual health-boundary crossings. Changing the alert definition also changes effect sizes and backend ordering. The separate auxiliary technical record preserves the full noise and stress sweeps, margin-regime figures, and original explanations; the experiments and their registered statuses are unchanged.

G.5.3 NOISE ROBUSTNESS

The noise audit adds common-random-number isotropic actuation noise to both paired rollout conditions. Its prospectively locked success rule required a positive real-minus-alignment-rotated alert contrast on both backends at $\kappa = 1$, measured in each backend's own error scale. Table 71 retains this primary comparison: the Hailo interval covers zero, so the registered criterion is not met. The absolute noise scales differ between backends; these are not equal-noise comparisons. Reporting that difference does not change the miss. The locked subsample contains 512 windows from 10 episodes, with 5 noise seeds per window; it is not the original main panel.

Table 71: Registered primary actuation-noise comparison, in each backend's own error units. Contrasts and intervals use the episode-weighted analysis, not a subtraction of pooled alert rates. The complete sweep is archived.

Backend	Absolute noise σ	Real−rotated	Paired interval
EdgeTPU	0.01309	0.0810	[0.0630, 0.1021]
Hailo	0.03367	0.0029	[-0.0058, 0.0125]

G.5.4 HOPPER CLIPPING-MITIGATION AUDIT

Hopper is reported as a saturation-stress case because clipping couples the alignment-rotated condition with applied-delta survival. We tested whether clipping-mitigation variants restore both applied-norm matching and a backend-consistent event contrast. Scaling the injected error down to $\eta = 0.25$ or $\eta = 0.5$ preserves positive real-minus-rotated event deltas on the original Hopper primary stratum, but the alignment-rotated/real actual-applied norm ratio remains low. Restricting to a near-marginal interior-action subset makes actual-applied norms match by construction, but the event contrast becomes backend-dependent rather than uniformly positive. Neither variant satisfies both conditions, so Hopper remains a high-saturation stress panel rather than a second norm-matched Walker2d replication.

G.5.5 ALERT COMPOSITION AND ENDPOINT SENSITIVITY

Table 72 retains the main-panel alert audit. An actual crossing requires $\min_k d_{\text{health}}(s^{\delta}_{t+k}) \leq 0$ within the evaluation horizon. Every alert in these panels is instead heuristic-only: raw full-observation deviation exceeds the nominal health-coordinate margin without an actual crossing. These observations support short-horizon trajectory separation, not observed falls or task failure.

Table 72: Composition of trajectory-separation alerts in the main panels. Actual crossing uses the task health-coordinate margin; heuristic-only alerts are triggered only by the raw full-observation deviation exceeding the nominal health margin. All rows use P-orig.

Setting	Backend	Condition	n	Alerts	Actual crossing	Heuristic only
Walker2d-5100	EdgeTPU	real	1429	556	0	556
Walker2d-5100	EdgeTPU	align-rot.	1429	180	0	180
Walker2d-5100	Hailo	real	1429	582	0	582
Walker2d-5100	Hailo	align-rot.	1429	468	0	468
Walker2d-5400	EdgeTPU	real	1115	289	0	289
Walker2d-5400	EdgeTPU	align-rot.	1115	208	0	208
Walker2d-5400	Hailo	real	1115	324	0	324
Walker2d-5400	Hailo	align-rot.	1115	238	0	238
Hopper-6100	EdgeTPU	real	363	283	0	283
Hopper-6100	EdgeTPU	align-rot.	363	191	0	191
Hopper-6100	Hailo	real	363	279	0	279
Hopper-6100	Hailo	align-rot.	363	154	0	154

Changing the distance metric. The post-outcome replay in Table 73 preserves positive measured-minus-anti-aligned contrasts under the normalized and Mahalanobis variants in the tested cells.

Magnitudes change, however, and both variants reverse the raw backend ordering in the first Walker2d panel. Health-coordinate-only alerts are nearly absent. Positive signs under these variants therefore do not validate a backend risk ranking or turn the original mixed-coordinate alert into evidence of health-boundary failure.

Table 73: Post-outcome alert-definition sensitivity on the original Walker2d-5100/5400 panels. Every entry is the measured-minus-anti-aligned alert-rate point estimate. Coordinate-normalized L2 uses empirical per-coordinate scales; Mahalanobis uses a regularized calibration covariance and distance to the nearest health-boundary hyperplane.

Setting	Backend	Raw full L2	Normalized L2	Mahalanobis	Health only
Walker2d-5100	EdgeTPU	0.263	0.021	0.027	0.000
Walker2d-5100	Hailo	0.080	0.040	0.069	0.000
Walker2d-5400	EdgeTPU	0.073	0.035	0.042	0.001
Walker2d-5400	Hailo	0.077	0.041	0.032	0.001

A continuous health-coordinate check. On the larger Walker2d replay panels, the endpoint $\max_k \max_i |\Delta h_{i,k}|/m_i(s_k)$ measures height/angle deviation relative to the remaining coordinate-specific health margin. Table 74 retains its paired summaries: real-error ratios usually exceed alignment-rotated ratios, but their absolute magnitudes are small and actual crossings remain absent. This is construct-validity support for a local deviation measure, not a fall result. The archived stress-dose sweep is kept separate from these natural-scale observations and from the original main panels.

Table 74: Walker2d health-coordinate margin-ratio endpoint. The ratio is the maximum over the horizon of height/angle deviation divided by the corresponding remaining health-boundary margin. Actual crossings are reported as real/rotated counts.

Setting	Backend	Pairs	Real p50	Rotated p50	Paired delta p50	Frac. real $>$ rotated	Crossings
Walker2d-5100	EdgeTPU	2772	0.0022	0.0004	0.0014	0.960	0/0
Walker2d-5100	Hailo	2772	0.0031	0.0006	0.0014	0.917	0/0
Walker2d-5400	EdgeTPU	2239	0.0004	0.0001	0.0002	0.922	0/0
Walker2d-5400	Hailo	2239	0.0004	0.0002	0.0003	0.916	0/0

G.5.6 Construction variability and local fidelity

The primary alignment-rotated construction is deterministic conditional on the nominal Gramian and measured error. A separate 5-seed audit varies the randomized secondary controls and recomputes predictor-level geometry, without rerunning nonlinear outcomes. Constructed effective vectors remain norm matched to numerical precision; this is not a claim about perturbations actually applied along diverging trajectories. Full construction ranges remain in the Auxiliary Technical Record, §9 (Numerical Robustness and Fidelity).

Linearization is local, and no-alert conditioning uses outcomes. The coherent formula is exact for the stated linear time-varying model, not for nonlinear rollouts. The stored-rollout audit compares coherent energy with squared terminal observation deviation, using log-scale fit and rank summaries. Full-panel fidelity varies across the tested policies; restricting to no-alert windows improves several diagnostics, but selects on observed outcomes. It cannot supply an outcome-blind operating region or a task qualification rule. The complete paired full-panel/no-alert table and figure remain in the technical record. The prospective trust-gate test in Appendix G.4.3 addresses that distinction.

Finite-difference sensitivity. The joint state/action-Jacobian audit is post-outcome and uses the same fixed canonical population across steps. All tested rank correlations meet 0.95, but 1 AUC shift exceeds 0.02 and 3 top-decile overlaps fall below 0.70; the strict audit misses (Table 75). The locked implementation is retained: stable bulk ranking does not imply invariant tail selection.

Table 75: Post-outcome joint finite-difference sensitivity on the fixed Walker2d-5100 canonical near-margin population (1429 windows/10 episodes). Rows vary both A_t and B_t relative to the locked 10^{-4} score. The strict numerical audit misses because one AUC shift exceeds 0.02 and three top-decile overlaps are below 0.70.

Backend	Step	Rank ρ	Δ AUC	Top-10% Jaccard	Rot./real p50
EdgeTPU	10^{-5}	0.954	-0.023	0.673	0.174
EdgeTPU	10^{-3}	0.954	0.001	0.702	0.180
Hailo	10^{-5}	0.969	-0.011	0.580	0.218
Hailo	10^{-3}	0.969	0.001	0.653	0.216

G.5.7 Metric choice and a prospective trust-gate miss

The local score measures predicted terminal deviation as $\mathrm{CLAE}^{\mathrm{coh}}(H; W) = \delta s_H^\top W \delta s_H$. The original Euclidean score uses $W = I$, which mixes observation coordinates with different units. The inverse-variance and Mahalanobis alternatives are dimensionless; the health-coordinate metric retains torso height and angle, weighted by their inverse FP32 panel variances. Metrics use only nominal FP32 observations, not injected outcomes.

Table 76: Effect of the state metric on the coherent score, Walker2d-5100 and Walker2d-5400 real condition. Spearman and top-decile Jaccard are computed against the published $W = I$ ranking. Inverse-variance and Mahalanobis are dimensionless; health-coordinate keeps only the torso height and angle.

		Inv. variance		Mahalanobis		Health coord.	
Panel	Backend	ρ	Jaccard	ρ	Jaccard	ρ	Jaccard
Walker2d-5100	EdgeTPU	0.909	0.500	0.774	0.432	0.423	0.231
Walker2d-5100	Hailo	0.942	0.378	0.892	0.364	0.748	0.279
Walker2d-5400	EdgeTPU	0.870	0.879	0.865	0.886	0.745	0.667
Walker2d-5400	Hailo	0.772	0.879	0.786	0.866	0.725	0.661

Changing the metric changes top-decile membership in a panel-dependent way (Table 76); the score's numerical definition matters for the fixed-budget selector. Cross-panel and split-episode metric-estimation sensitivities found less variation than changing the metric in those tested settings; the full results remain in the technical record, not as new task evidence. Neither a favorable local fit nor a stable bulk ranking establishes a portable selection rule.

The prospective trust-gate result is reported in Appendix G.4.3.

G.6 Full-episode constructed counterexample

Direction-controlled unforced task return. E16 used the same fixed policy and artifacts on 512 new reset seeds (83101–83612). The measured arm applied the current backend error field. The control arm applied an episode-keyed signed permutation, fixed for all 1,000 steps and shared across backends for a given seed. At a common state this orthogonal transform preserves the raw error norm and every cross-time inner product. Before outcomes, both backends passed locked post-clip scale and nominal-CLAE gates: median control/measured effective norm was 1.000, while median control/measured CLAE was 0.659 and 0.617 for EdgeTPU and Hailo.

The sole primary contrast was control-minus-measured return, so positive meant that the measured direction was more harmful. Table 77 shows the opposite on both backends. Both two-sided Holm-adjusted tests are significant, but both intervals are entirely negative; the registered positive family therefore misses rather than supporting the intended bridge.

Table 77: Ant return under constructed error-direction controls. Return Δ is signed-permutation control minus measured field; positive was the locked hypothesis. The common-cap column is descriptive and restricts to pairs in which both arms reach 1,000 steps.

Backend	Return FP32/meas./ctrl.	Return Δ [95% CI]	Holm p	Term. meas./ctrl.	Common-cap Δ (n)
EdgeTPU	5470/4827/4500	-327 $[-476, -178]$	< 0.001	0.104/0.176	-112 (381)
Hailo	5470/4897/3966	-931 $[-1082, -781]$	< 0.001	0.066/0.277	-364 (350)

The controls also lose 12.7 and 40.2 m of forward progress and terminate 38.0 and 124.6 steps earlier on average. The negative common-cap contrasts show that termination alone does not create the ordering. The exact transform invariant is below 3.1×10^{-16} at a common state; after the arms visit different states, their mean effective-error norms are 0.391/0.380 (measured/control) on EdgeTPU and 0.478/0.492 on Hailo. This is evidence that a synthetic same-size direction intervention can change actual task performance, not that measured artifacts are unusually harmful. It is also inconsistent with the proposed transfer from lower nominal CLAE to safer full-episode behavior in this policy. Signed permutation can reassign actuator components and is not a naturally occurring compiler variant, so the result does not estimate natural artifact or policy-population prevalence.

State screening does not establish episode prediction. In the earlier lifetime panels, binary termination is degenerate because every FP32 and backend episode terminates within the cap. First-window score versus termination timing has mixed signs across panels. Neither those diagnostics nor the outcome-conditioned no-alert fidelity subset establishes a transferable episode-level predictor. This limitation does not negate the separate paired task-return results in Appendix C.5. The original experiments, selection rules, and registered verdicts are unchanged.

H Auxiliary Fixed-Panel and Toolchain Checks

H.1 Output Identity, Panel Conformance, and Task Qualification

Output equality, panel conformance, and task qualification answer different questions (Table 78). A fingerprint checks output identity only on the queried inputs. The auxiliary classifier screen instead counts disagreements in serve/defer decisions on a fixed panel and compares that count with a declared tolerance. It does not measure class accuracy or closed-loop task loss. Accepting a panel pair is therefore not deployment acceptance or task qualification.

Table 78: Three separate contracts. Agreement or conformance in one row does not certify the next.

Evaluation	Question answered	Not established
Output fingerprint	Output identity on the queried inputs	Application importance of differences
Fixed-panel screen	Serve/defer disagreement within the panel tolerance	Accuracy, return, or success preservation
Paired task audit	Task loss relative to a specified tolerance	Other tasks, distributions, or devices

For a fixed artifact pair with fixed inputwise disagreement indicators, the screen's error bound is over uniform random input order without replacement. The allocated tail bounds cover the entire stopping event, not only each individual look. This is a pairwise finite-panel guarantee, neither family-wise control nor a guarantee about new deployment inputs. Emulator and target-runtime disagreement fields are separate measurement contracts.

The historical emulator corpora contain no material pair, so their absence of wrong decisions does not establish material-pair sensitivity. Later device and material-corpus follow-ups are separate evidence, not a relabelling of those corpora. Query reductions in stored-output replay do not recover already-paid inference cost; cache-aware model evaluations and end-to-end device time require their own accounting. Savings depend on the tested disagreement distribution: near the tolerance, the fixed rule can use almost the full panel, and sequential execution can be slower than exhaustive evaluation.

The separate auxiliary technical record preserves the complete stopping boundaries, operating-characteristic calculation and figure, replay results, device-cost follow-ups, and their limitations, with a local evidence index. The underlying records also remain indexed in Appendix I.4. They are the same measurements reported in the MLSys provenance study, not independent replications. None substitutes for the paired closed-loop task audit in Section 3.

H.2 Construction-only vendor probes

Is the order effect vendor-specific? Outcome-blind converter probes say the row-consumption rule explains it. They reuse one Fashion-MNIST classifier and one 256-row calibration pool and read neither labels nor task outcomes. Feeding the pool to the Edge TPU path's TFLite converter (TF 2.18, fixed weights, no device), the quantization parameters are identical under reversal and a random permutation in this tested conversion, which consumes every representative row rather than a fixed-size prefix. Consuming all rows does not by itself imply order invariance; the observation is parameter identity for these tested TFLite reorderings. The TFLite input scale is 0.005888 for all 256 rows, 0.005792 for a uniform stride-4 subsample of 64, and 0.005346 for the contiguous first 64, so a uniform subsample lands within 1.6% of the full-set scale where a head block is 9.2% away, echoing Appendix D.4.

A separately locked OpenVINO POT 2022.3.1 probe records every index requested from its data loader. All eight consumption predictions match: uncapped runs access every supplied row, while setting `stat_subset_size=64` accesses the first 64 positions in supplied-array order. The locked estimand preserves all 207 per-parameter $(\text{scale}, \text{zero point})$ pairs, of which 202 are per-channel entries that a per-node reading would collapse to five. Exact equality matches only two of six predictions: the two predicted-different comparisons pass, while all four identity predictions fail bit-exact equality in the same one-to-two-ULP regime as the identical rerun. A disclosed post-outcome sensitivity first matches the invariant zero-point multiset $\{0:5,\ 127:202\}$ and then aligns

scales within each group. It matches six of six: reversal and both capped-prefix controls are one ULP apart against the two-ULP repeat floor, while the two different-set comparisons are 254,590 and 433,494 grouped-scale ULPs. This sensitivity does not replace the locked exact result and establishes no task consequence. The tested MYRIAD plugin refused all eight quantized graphs at a `FakeQuantize` layer although each graph compiled on CPU, so no accelerator-output arm exists.

The provenance requirement is therefore toolchain-shaped. In these probes, TFLite needs the supplied-set identity, OpenVINO needs that identity plus the statistics cap and supplied order when capped, and the tested DFC path needs the identity of the rows it actually consumed. TFLite comparisons use name-free parameters because serialized conversions are not deterministic; the OpenVINO repeat-floor analysis is explicitly post-outcome because even those parameters move slightly. TensorRT is not comparable, as CAL4 builds explicit Q/DQ engines whose scales come from the graph rather than a calibration pass.

I Reproducibility, Provenance, and Evidence Index

I.1 Environment, execution, and source availability

Humanoid S1 policy. The Humanoid-7100 S1 smoke policy is used only for appendix inventory and pre-rollout interpretability checks. It uses the same 256–256 ReLU PPO actor class, eight vectorized environments, external VecNormalize, rollout length 2,048, minibatch size 1,024, five PPO epochs, learning rate 10^{-4}, clip range 0.15, target KL 0.05, and $\gamma = 0.99$. The policy bundle records a 3M-step stable-KL checkpoint resumed for 2M additional steps, with a VecNormalize count of 5,015,248. The 100-episode FP32 evaluation recorded in the bundle has mean return 5914.8 and mean length 1,000.

Compiler and runtime versions. Table 79 summarizes the compiler/runtime provenance used by the paper-facing measurements. These versions are stored in the Gate-0 reports and backend-output manifests that accompany the measured action dumps. Because we treat hardware as part of artifact identity, we also fix the execution sites. Compilation runs on a UDOO Bolt host; the two backends execute on separate Raspberry Pi 5 Model B hosts, Edge TPU artifacts on a Coral USB accelerator and Hailo artifacts on a Hailo-8 over PCIe. Compiled files move between machines by hash, and the recorded digests are re-verified on the runtime host before any dump is accepted.

Table 79: Compiler and runtime provenance recorded for paper-facing backend measurements. Runtime labels are normalized from the run manifests; exact host strings are retained in the released artifacts.

Backend	Compiler	Recorded runtime labels
EdgeTPU	`edgetpu_compiler` 16.0.384591198	Coral TFLite/libedgetpu stack. The original Walker2d-5100 host recorded TFLite runtime 2.14.0 with libedgetpu 16.0; later replication and Humanoid hosts recorded TFLite runtime 2.5.0.post1 with PyCoral 2.0.0 or libedgetpu USB labels.
Hailo	`hailo_dfc` 3.33.1	HailoRT 4.20.0 or the corresponding HailoRT Python runtime. Original calibration and task families used the CPU host's automatic level-0 fallback; later row-order and headline-family robustness audits forced levels 0 and 1, and the level-ladder and calibration-budget audits forced levels 0–3 with the quantizer statistics count set to 64 or 256 on that host. An RTX A5000 audit crosses GPU visibility with forced levels 0–3 on a common SDK-emulator panel; a later Hailo task bridge evaluates one level-3 contrast on 512 starts.
TensorRT	TensorRT 10.3 explicit Q/DQ (NVIDIA Corporation, 2024)	TensorRT 10.3 serialized engines on Jetson Orin; engines are platform- and version-specific.

A later attempt on an RTX 5070 Ti reached automatic optimization level 1, but TensorFlow 2.18 failed during model construction with `CUDA_ERROR_INVALID_HANDLE`; it produced no GPU-path artifact or task result. A subsequent Ampere host succeeded and is reported in Appendix D.3.2.

The later Walker2d-5700/5800 EdgeTPU panels change both policy artifact and runtime stack relative to Walker2d-5100, so they are not seed-only replications. To bound the runtime component, we reran the exact compiled Walker2d-5100 model on its fixed 2,048-state panel with TFLite 2.5.0.post1/PyCoral 2.0.0 and compared it with the stored TFLite 2.14.0 output. All $2{,}048 \times 6$ dequantized actions were bit-exact (maximum absolute difference 0). This fixed-model check reduces concern for that artifact and panel, but does not establish runtime invariance for other compiled models.

Reproducibility release bundle and error-field release. The accompanying release bundle contains the source code, run configurations, progress logs, policy specs, actor manifests, normalization metadata, state panels, calibration observations, backend action dumps, derived error-field tables, bootstrap-ready CSV/JSON figure data, and hashes for all compiled artifacts used in the paper. The

error-field dataset records the policy seed, task, backend, state identifier, episode identifier, time index, FP32 action, backend action, raw error, post-clip effective error where applicable, selection-bin labels, and manifest hashes. Vendor compiler binaries and some compiled backend artifacts may be subject to redistribution constraints; the released measured action dumps and hashes are intended to make the paper's analysis reproducible without device access, while exact backend reruns require the compiler/runtime versions listed above. For the row-order audits, the release additionally tracks CAL2's four calibration arrays, four Hailo executables, and complete 2,560-row outcome table, plus CAL3's three fixed cells, 12 executables, and complete 3,840-row outcome table. Included verifiers check every digest and recompute both locked decisions from a clean clone. The CAL8 release directory tracks the compressed 24,576-row matched-level episode table, task summary, primary/secondary summaries, lock/config hashes, and analyzer hashes; no device access is needed to recompute its reported statistics. Private host routing is supplied through an untracked runtime profile rather than embedded in the release configuration.

Registry provenance. "Prospectively locked" means committed to the repository before the corresponding outcome run; it does not denote registration with an external service. The canonical file is `docs/CLAE_ICLR27_experiment_registry.md`, version 0.32-amended through 2026-08-03. Its SHA-256 at this manuscript snapshot is `8fb196a86abc1d11c1b91dd82382fd9b8a4d0ff5bdd0740aa7cd7e885e46dd42`. Git history retains the earlier registry versions. Dated files under `docs/progress/` record the registry version, script hash, inputs, and amendments for each run; for example, the full-episode lock note records both registry and runner hashes before the confirmatory run. The E20 bounded-family, selection, and task-outcome locks are separate dated commits recorded in attempts 367–371; the outcome lock commit is `9003444`. The complete-family E21 outcome was locked in commit `339201d` before its fresh 1,024-seed block was opened. The same-multiset row-order lock was committed as `9aab226` before the fresh 512-seed block was opened; attempts 392–393 retain its lock and result. The CAL3 outcome-blind construction and Stage-B outcome locks were committed as `5f6af56` and `7a94a53`; attempts 395–398 retain its fixed three-cell family and miss. The initial CAL8 Stage-A lock was `d387a04`; before outcome access, the construction-matched 4D amendment and Stage-B outcome lock were committed as `47796e7` and `2c4423a`. Attempts 430–434 retain the superseded lock, amendment reason, outcome-blind gate, fixed two-test Holm family, and result integration. The matched-level release directory records the resulting tables, locks, and analyzer hashes.

I.2 Re-execution audit

We re-executed the committed Walker2d-5100 measured-error injection with unmodified code and compared the result column by column against the stored table. The output is not byte-identical, so we report what does and does not reproduce.

All 42 discrete and label columns match exactly, including every alert and endpoint flag, time-to-boundary, censoring state, and termination reason. All 386 count, fraction, and rate statistics in the run summary match exactly, including the alert fractions that the main panels report. The coherent CLAE columns, open-loop action mean-squared error, and local spectral radius are bit-identical. Among continuous rollout statistics, 197 of 1,862 summary leaves differ, at a maximum relative deviation of 9.4×10^{-7}; `terminal_deviation_norm` is identical in 99.7% of rows with a maximum relative deviation of 3.8×10^{-5}.

One column behaves differently. `min_health_boundary_margin` differs in 97% of rows, with a median relative deviation of 1.2×10^{-4} and a maximum of 47%. It is a minimum over a 20-step chaotic rollout, so the step attaining it moves under floating-point perturbation, and it is an absolute quantity rather than a paired difference. Its sign never changes and no row crosses zero in either execution, so every derived crossing label is unaffected. The closest approach to the health boundary is nevertheless less stable than the stored value suggests, moving from 7.3×10^{-4} to 3.9×10^{-4}; the absence of actual crossings holds in both.

Re-running the health-coordinate endpoint audit of Appendix G.5.5 reproduces every median and every paired fraction exactly, with only two means differing at 3.0×10^{-7} relative. The pattern is consistent: paired differences and order statistics absorb environment numerical drift, while absolute extremal quantities do not. Because the reported endpoints are paired, this supports the design rather than qualifying the results.

Artifact compilation is a second, separate reproducibility limit. Recompiling one unmodified integer TFLite export twice on the same host, under the same name, gives Edge TPU artifacts of identical size that differ in sixteen bytes, in two eight-byte slots holding one value per run; the Hailo compiler similarly varies the element order of a metadata list. The numerics are stable across those runs: Hailo quantization parameters, calibration statistics, and 399 of 400 arrays are identical, and the integer TFLite exports are byte-identical. Agreement is confined to the measured artifact–panel cells: artifacts compiled independently from identical inputs produced action dumps agreeing exactly in all 16 checked artifact–panel cells, and repeated dumps agreed exactly in all 40 checked artifact–panel cells. The later CAL8 construction audit adds an outcome-level bridge: its matched level-0 rebuilds reproduced all 12,288 historical episode rows exactly before level-1 outcomes were opened. Here a cell is one artifact evaluated on one fixed state panel. These historical checks do not establish equality of the controllers on unmeasured states or under other build conditions. The later 200-build audit finds two or three SDK-emulator action fields in 4 of 20 build cells (ten builds per cell), all GPU-visible level 3. This is why the audits fix and ship the executed file rather than a recipe for rebuilding it. The injection runner now records the simulator library versions alongside the backend toolchain versions, so a future re-execution can attribute this kind of residual.

I.3 Timing interruption and recovery provenance

Execution interruption and recovery. The complete assembly contains 104 jobs and 13,312 episodes. The first attempt stopped on SIGTERM after 18 completed jobs; the signal sender remains unknown. Those jobs passed raw simulator/FP32 replay, action-source, RPC and pre/post probe verification before being copied unchanged into an explicitly source-labelled assembly. A separately authorized, session-independent recovery ran 86 jobs, starting the interrupted job again on its original seed block. The partial failed trace was preserved but not spliced into scientific outcomes. There were no further collection failures. All complete rows passed the original full-cohort verifier and fixed analyzer. Reused rows are not additional independent observations; the original failure evidence and its hashes remain in the record.

I.4 Machine-readable result index

Reader-facing descriptions use the analysis names below. Exact repository-relative paths are retained here for reproduction, not as experimental labels. The experiment code glossary in Table 80 preserves the mapping to historical locks; names and paths in those original records have not been rewritten.

1. **Matched-deferral sensitivity:** `docs/iclr27/reviewer_response/2026-09-08/utility/`.
2. **Fresh-policy decision benchmark:** `docs/iclr27/fresh_policy_decision_v1/stage_d_comparison_v1.json`.
3. **Action-source timing intervention and recovery:** `docs/iclr27/action_timing_intervention_v1/cohort_recovery_result_v3.json`; `docs/iclr27/action_timing_intervention_v1/recovery_execution_v3.json`.
4. **Registered and sensitivity analysis roles:** `docs/iclr27/reviewer_response/2026-09-08/qualification_roles/result.json`.
5. **Calibration-budget uncertainty and panel independence:** `docs/iclr27/reviewer_response/2026-09-08/cal10/result.json`.
6. **Source-availability audit:** `docs/iclr27/reviewer_response/2026-09-08/provenance/result.json`.
7. **Ant null-transformation sensitivity:** `docs/iclr27/reviewer_response/2026-09-08/ant/result.json`.

Detailed fixed-panel screening records are indexed under *Screening evidence records* in the separate auxiliary technical record (`paper/iclr_auxiliary_controls.pdf`). Its historical-code glossary also retains definitions belonging only to the separated cost and QNN material. The retained definitions in Table 80 and their original records are unchanged.

The GPU-visibility audit's internal condition IDs `full_original` and `gpu_auto` denote, respectively, the full calibration set in its original order and GPU-visible compilation with automatic optimization-level selection. Compiler/API names such as `calibset_size` retain their technical spelling.

Table 80: Experiment codes and panel identifiers. Main-text claims use descriptive names; CAL distinguishes construction audits from appendix letters.

Code	Meaning	Appendix
E1a	Full-episode online backend versus FP32 lifetime, timing, and return audit	G.2.3
E1b-i / E1b-ii	Post-window FP32 recoverability / early-score versus full-episode outcome audits	G.5.1
E2, S0–S3	Task extension; policy acceptance, export, backend direction screen, and nominal margin/saturation pilot	G.2
E3a / E3b	Paired predictor comparison / norm-matched cross-backend discrimination	G.3
E4	Actuation- and observation-noise robustness audit	G.5.3
E5	Walker2d-5700 precision replication and expanded mechanism controls	G.3.3
E6	Formal propositions, linearization fidelity, and finite-difference sensitivity	G.1, G.5
E8	Prospective within-Hailo calibration-family selection follow-up	G.4.2
E8-R	Independent-episode replication of the fixed E8 selected contrast	G.4.2
E9	Prospective near-boundary ACTUAL endpoint audit	G.5.1
E10	Return-qualified Ant candidate and force-calibration plan; stopped at capability	G.2
E11	Fresh fixed-budget local-alert prioritization audit	G.4.1
E12	Ant-73102 health-normal-preserving endpoint bridge; stopped at precheck	G.5.2
E13	Ant-73102 online compiled-artifact total-effect and actual restricted mean survival time (RMST) audit	G.5.2
E14	Ant-73102 unforced online total-artifact task-return audit	C.5
E15	Ant-73102 fixed-budget return prioritization; stopped before backend outcomes	G.2
E16	Ant-73102 measured-versus-signed-permutation task-return audit	G.6
E17	FetchPush full-episode online total-artifact task-return audit	C.3
E18	FetchReach full-episode online total-artifact task-return replication	C.4
E19	Outcome-blind scalar-match feasibility check on the existing FetchReach artifacts; stopped before outcomes	B.1.2
E20	Fresh bounded FetchReach calibration family and scalar-matched real-artifact task-outcome audit	B.1.2
E21	Complete 16-artifact FetchReach calibration-family audit on one fresh paired seed block	B.1.1
E22	Stored-data calibration-support and early error-tail diagnostic for the E21 family	Technical record
E23	Locked phase-early versus phase-middle outcome replication across eight FetchReach policies	C.1
E24	Locked metric-aware trust-gated CLAE fidelity test on two never-inspected Walker2d policies	G.4.3
E25	Two-stage calibration-support dose response across four FetchReach policies	C.2
E26	Held-out dose-step replication on the remaining four FetchReach policies	C.2
E27–E30	Four locked Ant calibration-transfer attempts: two gate stops, one outcome floor, then one conditioned outcome miss and predictor reversal	C.7
CAL1	Calibration-count control and outcome-blind DFC effective-subset probe	D.2
CAL2 / CAL3	Locked same-multiset DFC task audit and multiset/policy breadth family	D.2
CAL4	Locked TensorRT explicit-Q/DQ calibration family; statistically heterogeneous but practically bounded	B.1

Table 80 (continued)

Code	Meaning	Appendix
CAL6	Locked FetchPush early-versus-middle calibration task-breadth audit	C.3
CAL7	Locked CPU DFC optimization-level by calibration-order factorial control	D.2
CAL8	Outcome-informed matched-level robustness factorial for the complete eight-rule family and eight-policy contrast	D.2
CAL9	Outcome-informed forced level-ladder audit extending CAL8 to the compiler's learned stages at levels 2–3	D.3.1
CAL10	Outcome-informed calibration-budget audit crossing optimization level with the quantizer statistics count	B.2, D.4
CAL11–CAL14	Locked coverage ladder, its geometry separation, and the two-cohort Ant geometry transfer	D.5, C.8, C.9
Alert (registry: SURR) / ACTUAL-TERM	Finite-horizon trajectory separation / actual Gym health-boundary crossing	G.1
P-orig	Eligible windows from the original 2,048-state per-policy collection	A.2
P-4096	Eligible windows from the larger 4,096-state Walker2d episode-count collection	A.2
P-exp	Eligible windows from the 8,192-state Walker2d-5700 precision expansion	A.2

Additional fidelity sensitivities. Each sensitivity keeps its own analysis plan, complete comparison record and generated tables in its own directory. Every one of them is post-outcome with respect to the task results, which were observed before any of these plans was written; each plan was nevertheless fixed, and committed, before its own quantities were computed, and each records that ordering together with its registered predictions and their verdicts. The episode-cluster bootstrap behind Tables 13 and 14 is under `docs/iclr27/fidelity_panel_bootstrap_v1/`; reproduce it with `scripts/analyze_fidelity_panel_bootstrap.py` and regenerate the presentation with `scripts/report_fidelity_panel_bootstrap.py`. Running the analyzer once the record exists writes the seed-stability record of its own amendment instead. The coordinate sensitivity behind Table 15 is under `docs/iclr27/executed_action_fidelity_v1/`, with `scripts/analyze_executed_action_fidelity.py` and `scripts/report_executed_action_fidelity.py`. Its clipping amendment preserves the initial raw diagnostic. Its plan pins the installed simulator package and parses that package's source for the action transform, so rerunning it needs that environment and not the analysis environment alone. The cross-distribution decomposition behind Table 18 is under `docs/iclr27/cross_distribution_fidelity_v1/`, prepared with `scripts/prepare_cross_distribution_panels.py`, executed with `scripts/run_cross_distribution_fidelity.py` on the accelerator and analyzed with `scripts/analyze_cross_distribution_fidelity.py`; it is the only sensitivity here that runs the device. The visited-state comparison behind Tables 16 and 17 is under `docs/iclr27/visitation_fidelity_v1/`, with `scripts/analyze_visitation_fidelity.py` and `scripts/report_visitation_fidelity.py`. It evaluates the full-precision policy, so it needs the policy runtime rather than the analysis environment alone; its active-component amendment was added after the original metric had been reported and leaves that metric unchanged. Of the four, only the panel bootstrap reruns in the analysis environment unaided. Existing result files are not overwritten by the analyzers; regression tests recompute and compare them in memory.