ActionShift: A Benchmark for Hidden Compositional Action-Interface Adaptation in Manipulation

Krishi Attri
krishiattriwork@gmail.com

Abstract

Robot-learning benchmarks that hide *dynamics* parameters—mass, friction, payload—have driven a mature line of adaptation methods: online system identification, meta-RL, and rapid motor adaptation. A distinct, software-level failure mode is equally real and structurally different. The mapping from a policy's action tensor to controller behavior—which channel drives which axis (permutation), each channel's sign and scale, whether the target is a delta or an absolute pose, its reference frame, its pipeline lag, and the gripper convention—can itself be wrong, undeclared, or silently changed, and we are not aware of a prior benchmark that isolates this *action-interface contract* as the object of study while holding task dynamics fixed. We introduce **ActionShift**, a real-simulation benchmark on frozen, competent ManiSkill PPO and Diffusion Policy backbones (PickCube, PushCube, PullCube, StackCube) in which a discrete, compositional interface contract is hidden behind a leakage-safe wrapper across preregistered seen, unseen-composition, and long-lag splits, with a privileged oracle and preregistered promotion gates. A competent policy that is near-perfect under a known interface (up to 1.000) collapses to a 0.000–0.007 floor when it is hidden—identically for RL and imitation policies—and the same policy-agnostic belief adapters restore both. Across a matched-privilege tournament, entropy-guided probing clears our promotion gate over fixed probing on 2 of 4 tasks (+5.8 points, 3-seed paired-significant on PickCube), and our controller, DualABI, matches the champion's success at roughly half the probe cost on all four tasks. Grammar-only belief, with no privileged pool, converts the learned-identifier tier's ∼0.0 into 0.36–0.67 and—via a hold-probe excitation schedule—breaches the all-absolute wall (0.005→0.722, 0.000→0.790). We report the load-bearing negatives in equal detail: passive learned identification fails outright (0.0), every reactive method including the oracle collapses under lag, and a delay-aware backbone solves the long-lag split by planning through delay (∼20×/∼2.7×)—evidence the two adaptation problems are not interchangeable in practice. We make no literature-priority ("first") claims: every novelty statement is an explicit "we are not aware of" backed by a direct literature check.

1 Introduction

Deploying a manipulation policy requires more than a correct model and task: it requires a correct *wire-up* between the numbers the policy emits and what the controller does with them. This **action-interface contract** is a small, discrete, compositional object: which output channel drives which axis (permutation), each channel's sign and physical scale, whether the target is a delta or an absolute pose, its frame, its pipeline lag, and whether the gripper is inverted. Every field is a real, recurring practical axis—exactly what dataset converters (LeRobot [1], Open X-Embodiment/RT-X [15]) must get right by hand—and getting one wrong is not graceful degradation: a permuted or sign-flipped channel drives the robot into a different, often unsafe motion. It does not error out; it acts confidently and wrongly.

Existing adaptation benchmarks do not isolate this. Hidden-parameter adaptation (UP-OSI, RMA, VariBAD) hides *continuous physical dynamics*; compositional-generalization benchmarks (RoboHiMan, ATOM-Bench) vary *skill composition*; robustness suites (LIBERO-Plus, COLOSSEUM) vary *visual/scene* perturbations. None is a controller-convention axis; we checked directly (§2). The closest neighbor, Reflective VLA's calibration-shift, applies a single continuous additive bias—not a compositional discrete grammar, and not a benchmark with splits, oracles, and promotion gates. Even production cross-embodiment systems (GR00T N1) assume the interface is *declared*.

The two problems are not interchangeable, and our data shows it: a UP-OSI-style passive identifier, trained exactly as for a hidden dynamics parameter, collapses to 0.0 on our hidden-interface variable (§5.3), even though the same responses support ∼1.0 for privileged pool-based identification. ActionShift makes this distinction measurable rather than definitional.

Contributions. (1) A benchmark holding dynamics fixed and hiding only the action-interface contract, on four frozen, parity-verified ManiSkill [22] tasks (Peg excluded on a competence floor, not silently dropped). (2) A leakage-safe wrapper with a privileged oracle outside the agent path, reserved-key sanitization, and fail-closed safety checks, plus five content-hashed split manifests. (3) A preregistered promotion-gate methodology applied honestly—only entropy probing (some tasks) and DualABI clear a gate. (4) A matched-privilege adaptation ladder from a 0.000 floor through a grammar-only belief (0.458) to a pool belief (0.987) at the oracle ceiling (1.000), with the load-bearing negative that active probing is not a substitute for hypothesis-space knowledge. (5) **DualABI**, a task-regret-aware probe-selection/early-stopping controller matching the champion at ∼half the probe cost across four tasks—efficiency, not raw success. (6) A precise characterization of the unprivileged information limit, and a hold-probe method breaching the all-absolute wall. (7) A demonstration that brittleness is a property of the interface, not the learning paradigm.

2 Related work

We adopt an absence-of-evidence framing throughout ("we are not aware of X," never claiming to be first in the literature).

Interface/metadata mismatch. **ExecSpec** [21] (our shorthand for this executable-specification work) formalizes an executable policy specification and shows metadata mismatches collapse success, detecting drift *analytically from declared metadata*. ActionShift studies the complementary case: the interface is genuinely hidden and the only evidence is the trajectory. (ExecSpec is an unreviewed single-author cs.CR preprint.)

Hidden-dynamics adaptation. UP-OSI [27], RMA [9], and VariBAD [29] establish history-conditioned identification and belief-conditioned policies for *continuous physical dynamics*. ActionShift's hidden variable is a *discrete compositional software contract*; DualABI's belief is a special case of the BAMDP-via-meta-learning idea, and our contribution is the hidden-variable class plus a task-regret-and-terminal-controllability scoring criterion—not the belief-conditioning idea. The transformer-ICL lineage ([10, 11, 16]) is named in our registry but reported as a structured exclusion (no faithful pinned reproduction), never a renamed proxy.

Active information-seeking. DualABI's probe/stop rule builds on established results [20, 26]; Task-Compatible Active Calibration motivates its terminal-risk term (its DOI `10.64898/2026.06.11.731593` returned HTTP 403 to automated fetch—flagged for manual verification). We claim only the narrow application to hidden action-ABI recovery.

Cross-embodiment / shared action spaces. A distinct group learns or shares an action representation *at training time* (shared-action-space methods such as GR00T N1 [14] and Open X-Embodiment [15]). ActionShift neither learns nor redesigns an action space; it hides an already-fixed contract at deployment and asks whether a policy can *recover*.

Near-miss benchmarks (checked and ruled out). RoboHiMan [2] and ATOM-Bench [25] compose skills; LIBERO-Plus [6] and COLOSSEUM [17] vary visual/scene axes—none is a controller-convention axis. The genuine near-miss is Reflective VLA [12] (one continuous additive bias, not a compositional discrete grammar, not a preregistered split methodology). Based on this direct check: *we are not aware of a prior benchmark that holds task dynamics fixed and varies a compositional, discrete action-interface contract (permutation, sign, scale, target, frame, lag, gripper) as the object of study, with privileged oracles and preregistered promotion gates.*

Delayed-MDP literature (long-lag split). Formal grounding: Katsikopoulos–Engelbrecht [8] (constant-delay reduction our backbone uses), Walsh et al. [23]. Deep-RL: DCAC [18] (the only turnkey baseline), D-TRPO [13]. We are not aware of a prior randomized-action-lag online-PPO ManiSkill result; three near-misses (obs-delay Meta-World [7], offline-RL Adroit [28], teleop Franka [4]) are cited and distinguished.

3 Benchmark design

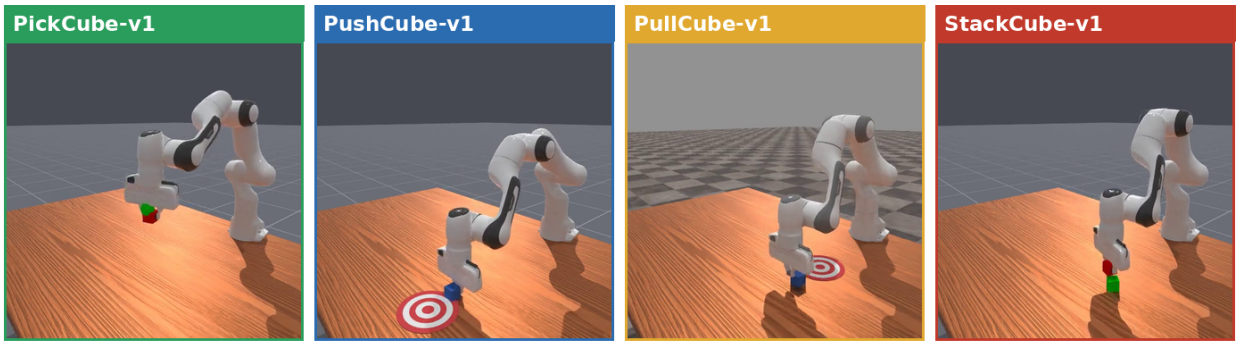

Figure 1: The four frozen-backbone ManiSkill tasks that constitute ActionShift (PegInsertionSide excluded on a Gate-0 competence floor). Task dynamics are held fixed; only the action-interface contract is hidden. Each frame is caught at the *moment of task execution*—the gripper grasping and lifting the cube to the green goal (PickCube), the arm mid-push driving the cube onto the target (PushCube), the gripper pulling the cube toward the target (PullCube), and the red cube being placed onto the green cube (StackCube). *Frames: ManiSkill3 official environment demos (`third_party/maniskill/figures/environment_demos`), regenerable via `media/make_paper_figs.py`; the task set and Gate-0 verdicts are ours (Source: `gate0.md`, `third_task.md`, `fourth_task.md`).*

Hidden variable. A contract is an immutable tuple over seven field families (permutation, sign, scale, target, frame, lag, gripper) supporting differentiable encode/decode, a stateful decoder, and per-environment lag reset. The declared scale grid is $\{0.5, 0.6, 0.75, 1.0, 1.25, 1.5, 2.0\}$, lag $\{0, 1, 2, 4\}$, target {delta, absolute}; a test asserts the grid covers every frozen contract.

Leakage-safe environment. The contract is applied through a ManiSkill wrapper (`pd_ee_delta_pose`, state obs). Reserved keys are sanitized so the agent cannot read the contract; an evaluation-only oracle recorder sits outside the agent path. Hard fail-closed checks (finiteness, bounds, workspace, collision) gate every run; maximum oracle decode error on a real smoke was 4.66×10^{-10}. A unit test asserts unprivileged adapters take no contract argument.

Splits. Five content-hashed manifests: *seen*, *unseen composition* (zero contract- and signature-overlap with training), *long lag* (2–4 steps), and two held for future work.

Frozen backbones (Gate 0). The backbones—PPO [19] and, for the imitation study, Diffusion Policy [3]—are frozen before adaptation and pass wrapper parity (Table 1).

Promotion gates. A method earns a longer run only by clearing $\geq$5 points over a matched baseline, 30% faster recovery, a meaningful unseen/lag gain, or a success-vs-cost Pareto point—without a safety regression. A sub-0.3 one-seed result is not promoted.

Table 1: Gate 0 competence. *Source: gate0.md, third_task.md, fourth_task.md, peg_retry.md*

Task	Floor	Budget	Final success	Parity	Decision
PickCube-v1	0.50	10M	1.00	pass	advance
PushCube-v1	0.50	2M	1.00	pass	advance
PullCube-v1	0.50	2M	1.00	pass	advance
StackCube-v1	0.50	25M[†]	0.98	pass	advance
PegInsertionSide-v1	0.20	$\geq$75M	0.00	—	**exclude**

[†]The 25M StackCube run used num_envs=1024; the true official budget is 50M at num_envs=4096—competence stands, no budget-efficiency claim (§7). Peg is 0/100 at 41M, 71.7M, and 75.7M.

4 Methods

Every method is a policy-agnostic *adapter* between a frozen backbone and the hidden-contract environment (Fig. 2). The privilege ladder: privileged *oracle*; *pool-privileged belief* (exact filter over a declared 9-contract pool with fixed/random/entropy probes and DualABI); *grammar-knowledge belief* (full-grammar factorized, hold-probe, scale corrector); *unprivileged learned identifiers* (passive UP-OSI-style, recurrent, probe-augmented); *delay-aware augmented-state control*; and *no adaptation* (the floor).

DualABI selects probes and stops early by task regret $R(b) = \sum_j b[j] \,\mathrm{mean}_k \|\mathrm{decode}_j(\mathrm{encode}_i(a_k)) - a_k\|$ (i the MAP hypothesis, $b[j]$ the posterior mass on hypothesis j, k indexing sampled task actions): ambiguity that does not change the task action contributes no regret, so it can stop before the belief fully collapses. Its threshold was frozen after a 48-episode pilot (robust $\tau \in [0.4, 1.5]$).

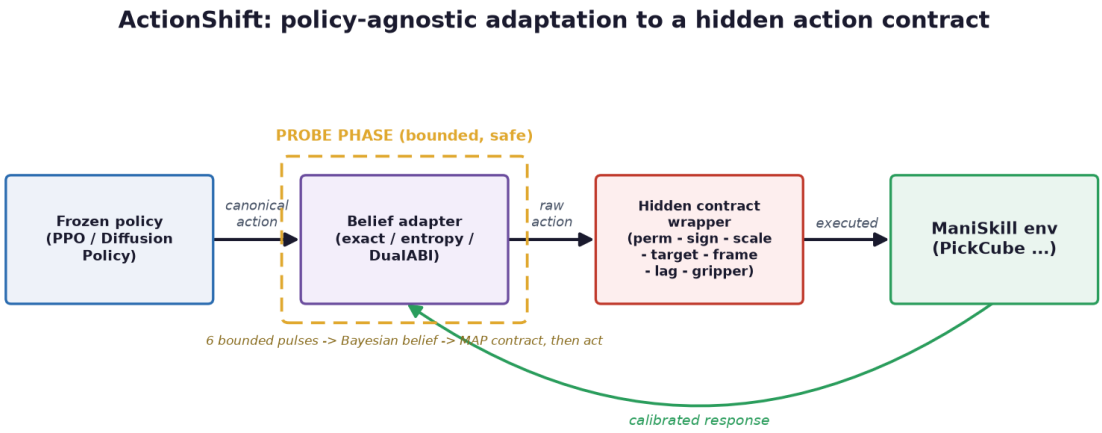

Figure 2: The policy-agnostic adapter architecture. A bounded probe phase (the gripper is never actuated) folds calibrated responses into a Bayesian belief; the adapter acts on the MAP contract.

5 Results

All headline comparisons use three matched seeds with Wilson 95% intervals [24] (success) and paired bootstrap intervals [5] (differences). No broad-novelty or method-superiority claim is made.

5.1 The benchmark premise (Gate 1)

A competent policy is near-perfect with a known ABI and almost always fails without one. **Long lag is a separate falsification**: even the oracle collapses under 2–4-step delay (Pick 2.7%, Push 15.3%, Pull 32.2%, Stack **0.000**). Contract knowledge alone is insufficient.

Table 2: Gate 1: hidden-contract PPO ceiling, 7,200 episodes. *Source: gate1.md*

Task / split	Oracle (Wilson)	No-adapt (Wilson)	Paired diff
Pick / seen	1.000 [.994,1]	0.000 [0,.006]	1.000 [1,1]
Pick / unseen comp.	0.995 [.985,.998]	0.000 [0,.006]	0.995 [.988,1]
Pick / long lag	0.027 [.016,.043]	0.005 [.002,.015]	0.022 [.008,.037]
Push / seen	1.000 [.994,1]	0.003 [.001,.012]	0.997 [.992,1]
Push / unseen comp.	1.000 [.994,1]	0.000 [0,.006]	1.000 [1,1]
Push / long lag	0.153 [.127,.184]	0.000 [0,.006]	0.153 [.125,.182]

5.2 The adaptation ladder

On PickCube/seen the frozen backbone climbs from a 0.000 floor (no-adapt *and* passive learned identifier), through a grammar-only belief (0.458), to a pool belief (0.987) at the oracle ceiling (1.000)—the load-bearing negative being that active probing is not a substitute for hypothesis-space knowledge (Fig. 4).

Entropy probing clears the gate on some tasks, not all. Entropy − fixed = +5.8 [+3.7, +8.2] on Pick/seen (no safety/cost regression; ∼24% less displacement), replicating on Pull/unseen (+5.5 [+3.3, +7.8], a second gate-clear) and in direction on Push/seen (+1.8). It does not generalize: on StackCube/seen entropy is the *lowest* belief method (0.910) and misses the gate on unseen (+2.5 ns), because passive belief already saturates there. *Source: adaptation_tournament.md, third_task.md, fourth_task.md*

5.3 The unprivileged challenge, quantified

Passive UP-OSI-style: 0.0 end-to-end, permutation 0.29, sign at chance; a random-excitation ablation lifts permutation to 0.52 and lag to 0.77 but sign stays at chance and evaluation stays 0 (calibration R^2 0.09–0.34). An episode-length recurrent adapter lifts *discrete* fields with horizon (lag 0.75→0.92, target 0.59→0.69) but permutation plateaus ∼0.35 by step 15 and sign never leaves chance; end-to-end 0/200 and 2/200. A probe-augmented learner caps permutation at 0.39 (< the random ablation's 0.52) and stays 0.000–0.015. Mechanism: 6 basis pulses span ≤6 regressor directions against a 12-dimensional lagged map—a belief can *score* a small hypothesis set to ∼1.0 but a learner cannot *estimate* the continuous map. *Source: adaptation_tournament.md, adaptation_recurrent.md, adaptation_probe_osi.md*

5.4 Grammar-only belief: walls exposed, then breached

Grammar knowledge (no pool) is the earliest belief in our ladder to succeed without a pool: Pick/seen 0.363/0.458, Push/seen 0.633/0.673, reaching 0.990 on the clean sub-cell. It localizes two response-model walls. **Wall 1 (gripper)** is substantially closed by a v2 finger-qpos evidence channel (R^2 0.545): the Pick kill-cell 0.000→0.537, the probed cell 0.458→0.715, no regression. **Wall 2 (absolute+scale)** is breached by *hold-probe excitation* (telescoped window evidence): all-absolute cells 0.005→0.722 (Push) and 0.000→0.790 (Pick), exact permutation recovery, causal controls (probe-only 0.000, flail-then-identity 0.000, flail-then-belief 0.725). Honest residuals: a base-frame absolute contract still executes at 0.44, and a 24-step probe regresses the delta cell to 0.405 (12 steps is the operating point). *Source: adaptation_factorized.md, adaptation_grasp_channel.md, absolute_excitation.md*

5.5 DualABI: a probe-efficiency Pareto win, four times

Aggregated (Table 3): mean probe steps 2.89 vs 6.00 (−52%), displacement 0.0154 vs 0.0267 (−42%), success 0.990 vs 0.987 (a tie—all paired intervals include zero). The Pareto win replicates on PullCube and StackCube (Fig. 5), 4 of 4 tasks. **DualABI does not beat entropy on raw success, and we make no such claim.** The week-one proxy negative (fixed 0.669 > regret-aware 0.519) stays on the record.

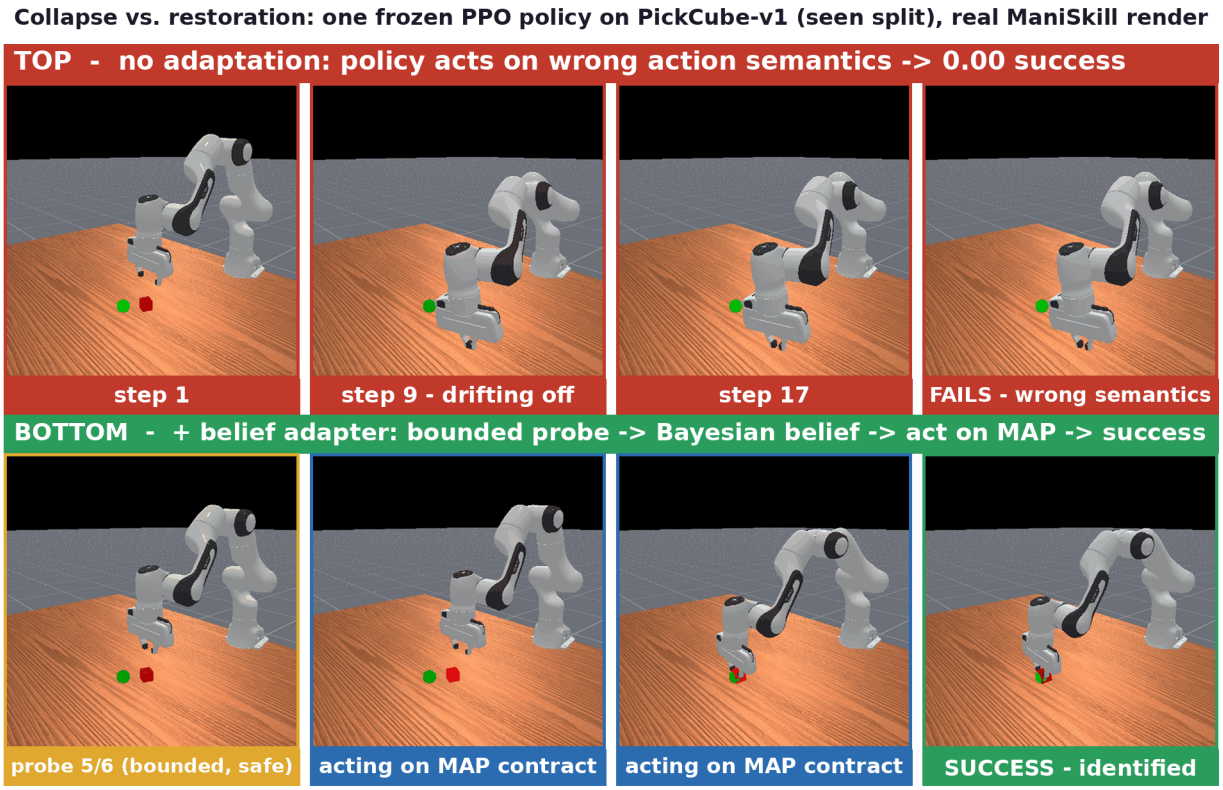

Figure 3: **Collapse vs. restoration, one frozen policy** (two rows × four frames). Real ManiSkill renders of the frozen Gate-1 PPO backbone on PickCube-v1 under one hidden seen-split contract. *Top (red):* no adaptation—the policy acts confidently on wrong action semantics, drifts off the cube, and fails. *Bottom:* the same policy with the policy-agnostic belief adapter—a bounded entropy probe phase (amber, gripper never actuated) folds calibrated responses into a Bayesian belief, the adapter then acts on the MAP contract (blue) and reaches the cube to succeed (green). Frames are cropped from `repo/media/hero_triptych.gif` and composited by `media/make_paper_figs.py`, with the underlying rollout regenerable via `media/make_media.py hero`: frozen Gate-1 PPO checkpoint, entropy adapter (budget 6, amplitude 0.5), seed the first in $[20260718, 20260758)$ with clean-success / no-adapt-fail / entropy-success. Per-cell success rates are in Table 2 (seen: oracle 1.000, no-adapt 0.000). *Source: `gate1.md`, `adaptation_tournament.md`*

5.6 The frame axis, de-degenerated

ActionShift v2 wires a live end-effector rotation into the frame axis (v1: base≡tool), giving a real oracle 1.000 vs no-adapt 0.000 gap on Pick and Push; the pool belief identifies tool-frame end-to-end (0.900–0.997). Disclosed limit: the factorized belief cannot represent tool-frame under real rotation (rotation couples channels), so frame identification needs the pool privilege or a richer evidence model. *Source: `rotation_v2.md`*

5.7 Long lag: solvable by composition, not identification

A delay-aware augmented-state PPO backbone (obs + last-4 canonical actions, randomized lag $\{0, 1, 2, 4\}$) solves the split (Fig. 6).

Identification alone cannot fix delay, delay-aware control alone cannot fix hidden semantics, the combination solves both (Table 4). This is a **new backbone** (context, not a method contest) labeled "delay-aware augmented-state PPO (local)"—the classical state-augmentation reduction, *not* DCAC or D-TRPO. Report both ratios; lag 4 stays hardest (0.30–0.40 oracle-encode); on StackCube even the oracle is 0.000 under lag.

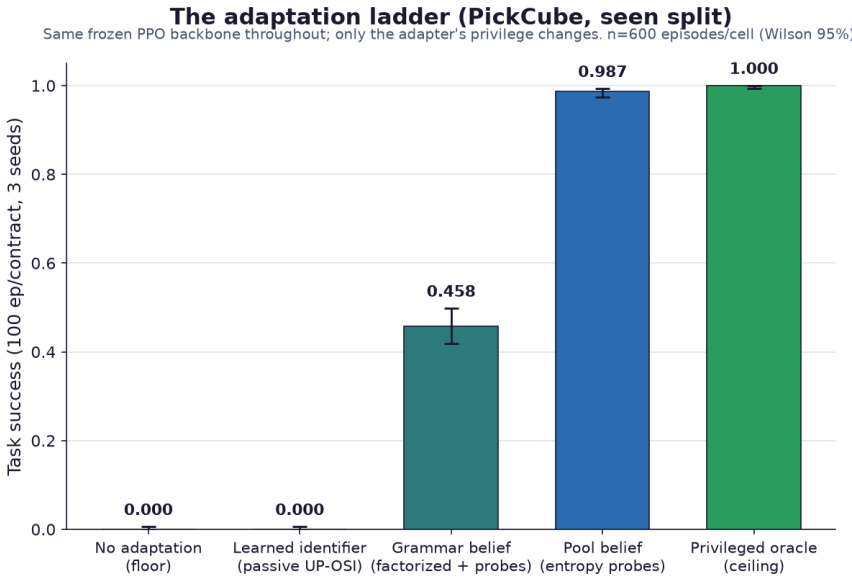

Figure 4: Matched-privilege adaptation ladder on PickCube/seen. *Source: adaptation_tournament.md, adaptation_factorized.md*

Table 3: DualABI vs the entropy champion (seen/unseen cells). *Source: adaptation_dualabi.md*

Task/split	metric	dualabi	entropy	fixed	exact
Pick/seen	success	0.983	0.987	0.928	0.929
	probe steps	**2.78**	6.00	6.00	0.00
Pick/unseen	success	0.983	0.975	0.970	0.958
	probe steps	**3.14**	6.00	6.00	0.00
Push/seen	success	1.000	1.000	0.982	0.983
	probe steps	**2.60**	6.00	6.00	0.00
Push/unseen	success	0.993	0.987	0.980	0.995
	probe steps	**3.03**	6.00	6.00	0.00

5.8 Brittleness is interface-semantic, not paradigm-specific

A frozen Diffusion Policy (clean 0.580/0.670) collapses to ∼0.003 under the same hidden contracts that zero PPO—the same 0.000–0.007 floor—and the same policy-agnostic adapters restore it to its oracle ceiling (0.62–0.735); DualABI's efficiency win transfers (0.0 probe steps on seen) (Fig. 7). Load-bearing quantities are horizon-matched within backbone. *Source: imitation_brittleness.md*

5.9 External anchors, honestly

ActionABI's C++ scoring core runs live in the belief loop with verified parity (float64 max rel diff 5.7×10^{-14}, bit-identical MAP); single-thread C++ wins the CPU regime at every batch size (∼13× at 8 envs, ∼11× at 1024), torch-CUDA wins at scale, transfer-inclusive C++ CUDA loses everywhere. *Source: cpp_fusion.md* On DCAC's constant-delay-5 MuJoCo benchmark, augmented-state SAC matches the augmented-SAC/RTAC *baseline tier* on HalfCheetah (∼1951 vs ∼2000) and Ant (∼716 vs ∼600–700) but reaches only about half on Walker (∼1104 vs ∼2200), while naive SAC collapses; it does *not* reach DC/AC. **No absolute-return claim vs DCAC.** *Source: delayed_rl_sac.md, delayed_rl_external.md*

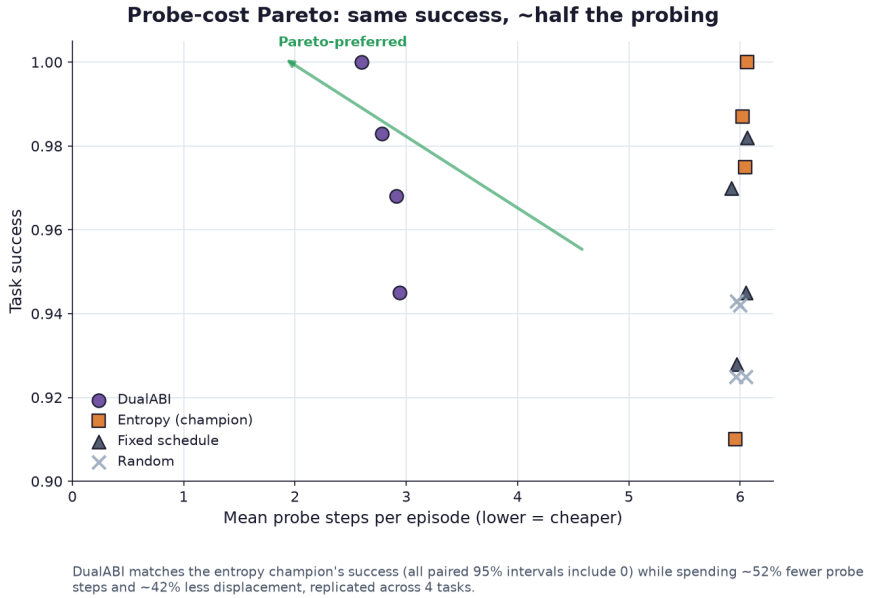

Figure 5: Probe-cost Pareto across four tasks. *Source: adaptation_dualabi.md, third_task.md, fourth_task.md*

Table 4: Long-lag solve, 3 seeds, 600 eps. *Source: adaptation_delay_aware.md*

Cell	delay-aware (Wilson)	frozen ref	ratio
Pick / oracle-encode	0.528 [.488,.568]	0.027	∼20×
Push / oracle-encode	0.415 [.376,.455]	0.153	∼2.7×
Pick / exact-belief	0.360 [.323,.399]	0.022	∼16×
Push / exact-belief	0.387 [.349,.426]	0.117	∼3.3×

6 Honest negatives and self-corrections

Passive learned identification transfers to ∼0 with a dimensional explanation (§5.3)—the strongest empirical signal the two adaptation problems differ. The "pipeline-flush" reading (fixed 0.215 > oracle 0.153 on Push/lag) was *downgraded* to a Push-specific artifact when it failed on PullCube (fixed 0.260 < oracle 0.322); "fixed > entropy under lag" reversed on the delay-aware backbone (a reactive-backbone artifact). Entropy is the worst method on StackCube. The magnitude-dependent gain buys ∼0 R^2; the scale corrector does not move real absolute cells alone. Two budget misreads (Peg 250M→75M, Stack 25M-vs-50M) were caught by our own audits before any claim shipped. The transformer-ICL (Vintix) row is an adjudicated exclusion with a pinned commit. *Source: lag_completions.md, fourth_task.md, adaptation_scale_corrector.md, peg_retry.md, transformer_icl_adjudication.md*

7 Limitations and claim boundary

Matched-privilege comparisons are about information-*seeking*, not unprivileged adaptation. No fully-unprivileged learned method succeeds. Sim-only. Four tasks; Peg excluded. DualABI has no long-lag Pareto claim (it collapses like the belief family, 0.02/0.11). Task-dependent findings (entropy > fixed, pipeline-flush, "> random") are disclosed as such. No literature-priority ("first") claims. **Do not claim**: beating/reproducing DCAC; a Vintix reproduction; any hardware result; matching ManiSkill's official baseline numbers; that the gripper/scale walls are "solved"; that the frame axis is solved without privilege; or that "action-interface shift" is a formally distinct RL problem class (the support is empirical only—a UP-OSI transfer failure).

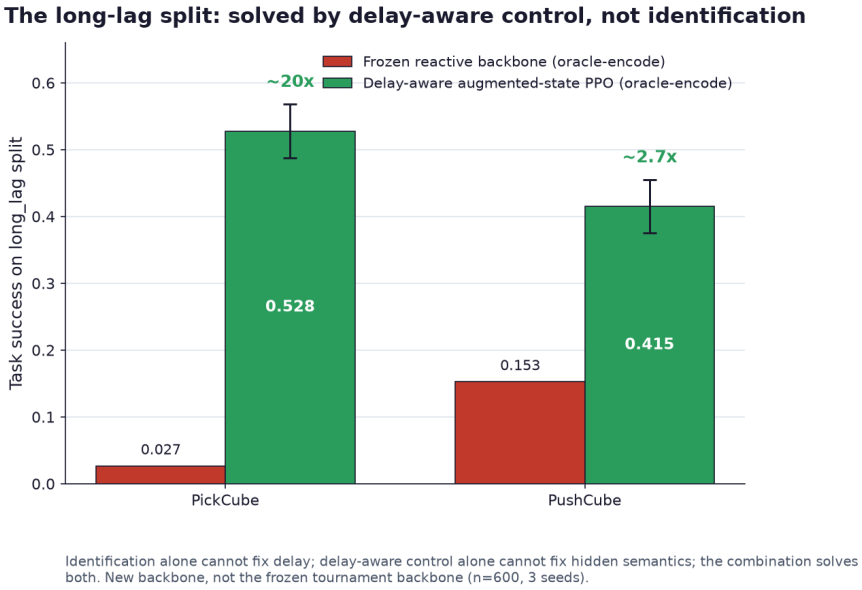

Figure 6: Long-lag solve: report both ratios, not just Pick. *Source: `adaptation_delay_aware.md`*

8 Reproducibility

All episodes are hash-addressed; runs write append-only episode JSONL and summaries with checkpoint SHA-256 per job. Three matched seeds, Wilson and paired-bootstrap intervals; one-seed cells are pruning signals only. Training-contract distributions are hash-rejected against every frozen evaluation contract; a test asserts unprivileged adapters take no contract argument. The C++ backend reproduces torch scores to 5.7×10^{-14}. ruff, strict mypy, and the full test suite pass. Large checkpoints and raw JSONL are local/gitignored; manifests, verdicts, reports, and figures are committed. Simulator/dataset assets retain their own licenses.

A Full adaptation-tournament results

Table 5 reports every populated cell of the matched-privilege adaptation tournament: per-task, per-split, per-method success with Wilson 95% intervals, mean probe steps, and the individual per-seed rates behind each pooled cell. All values are copied verbatim from the committed artifact `artifacts/adaptation/tournament.json` (*Source: `adaptation_tournament.md`*); the oracle and no-adapt reference cells are in Table 2. Seeds are 20260718/19/20; long-lag and StackCube cells are one-seed where a single per-seed value is shown (pruning signals, not promoted comparisons). Every belief-family method shares the same declared 9-contract pool, budget (6), amplitude (0.5), backbone, contracts, and seeds—the comparison is about the value of information-*seeking*, not unprivileged adaptation.

Table 5: Full tournament ladder: success by task × split × method, with per-seed rates. *Source: `adaptation_tournament.json`*

Task	Split	Method	Eps	Succ.	Wilson 95%	Probe	Per-seed (18/19/20)
PickCube	seen	exact-belief	800	0.929	[0.909, 0.945]	0.00	0.930 / 0.920 / 0.930
	seen	entropy	600	0.987	[0.974, 0.993]	6.00	0.985 / 0.990 / 0.985
	seen	fixed	600	0.928	[0.905, 0.946]	6.00	0.950 / 0.920 / 0.915
	seen	random	600	0.925	[0.901, 0.943]	6.00	0.895 / 0.930 / 0.950
	seen	DualABI	600	0.983	[0.970, 0.991]	2.78	0.985 / 0.980 / 0.985
	unseen-comp	exact-belief	600	0.958	[0.939, 0.972]	0.00	0.955 / 0.940 / 0.980
	unseen-comp	entropy	600	0.975	[0.959, 0.985]	6.00	0.990 / 0.960 / 0.975
	unseen-comp	fixed	600	0.970	[0.953, 0.981]	6.00	1.000 / 0.955 / 0.955

Task	Split	Method	Eps	Succ.	Wilson 95%	Probe	Per-seed (18/19/20)
	unseen-comp	random	600	0.980	[0.965, 0.989]	6.00	0.995 / 0.975 / 0.970
	unseen-comp	DualABI	600	0.983	[0.970, 0.991]	3.14	0.990 / 0.985 / 0.975
	long-lag	exact-belief	600	0.022	[0.013, 0.037]	0.00	0.015 / 0.030 / 0.020
	long-lag	entropy	600	0.025	[0.015, 0.041]	6.00	0.045 / 0.010 / 0.020
	long-lag	fixed	600	0.037	[0.024, 0.055]	6.00	0.055 / 0.035 / 0.020
	long-lag	DualABI	200	0.020	[0.008, 0.050]	1.61	0.020
PushCube	seen	exact-belief	600	0.983	[0.970, 0.991]	0.00	0.980 / 0.970 / 1.000
	seen	entropy	600	1.000	[0.994, 1.000]	6.00	1.000 / 1.000 / 1.000
	seen	fixed	600	0.982	[0.967, 0.990]	6.00	0.985 / 0.985 / 0.975
	seen	random	600	0.942	[0.920, 0.958]	6.00	0.920 / 0.945 / 0.960
	seen	DualABI	600	1.000	[0.994, 1.000]	2.60	1.000 / 1.000 / 1.000
	unseen-comp	exact-belief	600	0.995	[0.985, 0.998]	0.00	0.995 / 0.995 / 0.995
	unseen-comp	entropy	600	0.987	[0.974, 0.993]	6.00	0.985 / 0.985 / 0.990
	unseen-comp	fixed	600	0.980	[0.965, 0.989]	6.00	0.970 / 0.990 / 0.980
	unseen-comp	random	600	0.995	[0.985, 0.998]	6.00	0.995 / 0.995 / 0.995
	unseen-comp	DualABI	600	0.993	[0.983, 0.997]	3.03	0.995 / 0.985 / 1.000
	long-lag	exact-belief	600	0.117	[0.093, 0.145]	0.00	0.130 / 0.100 / 0.120
	long-lag	entropy	600	0.100	[0.078, 0.127]	6.00	0.115 / 0.100 / 0.085
	long-lag	fixed	600	0.215	[0.184, 0.250]	6.00	0.205 / 0.260 / 0.180
	long-lag	DualABI	200	0.110	[0.074, 0.161]	1.28	0.110
PullCube	seen	exact-belief	600	0.952	[0.931, 0.966]	0.00	0.930 / 0.965 / 0.960
	seen	entropy	600	0.975	[0.959, 0.985]	6.00	0.970 / 0.975 / 0.980
	seen	fixed	600	0.970	[0.953, 0.981]	6.00	0.970 / 0.965 / 0.975
	seen	random	600	0.943	[0.922, 0.959]	6.00	0.940 / 0.940 / 0.950
	seen	DualABI	600	0.968	[0.951, 0.980]	2.91	0.990 / 0.950 / 0.965
	unseen-comp	exact-belief	600	0.935	[0.912, 0.952]	0.00	0.945 / 0.910 / 0.950
	unseen-comp	entropy	600	0.982	[0.967, 0.990]	6.00	0.990 / 0.990 / 0.965
	unseen-comp	fixed	600	0.927	[0.903, 0.945]	6.00	0.925 / 0.930 / 0.925
	unseen-comp	random	600	0.997	[0.988, 0.999]	6.00	0.995 / 0.995 / 1.000
	unseen-comp	DualABI	600	0.977	[0.961, 0.986]	2.50	0.980 / 0.975 / 0.975
	long-lag	exact-belief	600	0.287	[0.252, 0.324]	0.00	0.275 / 0.330 / 0.255
	long-lag	entropy	600	0.325	[0.289, 0.363]	6.00	0.320 / 0.385 / 0.270
	long-lag	fixed	600	0.260	[0.227, 0.297]	6.00	0.230 / 0.260 / 0.290
StackCube	seen	exact-belief	200	0.945	[0.904, 0.969]	0.00	0.945
	seen	entropy	200	0.910	[0.862, 0.942]	6.00	0.910
	seen	fixed	200	0.945	[0.904, 0.969]	6.00	0.945
	seen	random	200	0.925	[0.880, 0.954]	6.00	0.925
	seen	DualABI	200	0.945	[0.904, 0.969]	2.94	0.945
	unseen-comp	exact-belief	200	0.960	[0.923, 0.980]	0.00	0.960
	unseen-comp	entropy	200	0.940	[0.898, 0.965]	6.00	0.940
	unseen-comp	fixed	200	0.915	[0.868, 0.946]	6.00	0.915
	unseen-comp	random	200	0.930	[0.886, 0.958]	6.00	0.930
	unseen-comp	DualABI	200	0.960	[0.923, 0.980]	3.48	0.960

B Per-round tournament log (rounds 1–13)

The benchmark was built as a preregistered tournament: each round posed a hypothesis and the outcome was recorded whether or not it supported the paper's thesis. Table 6 is the honest log; every row is a one-line summary of a committed report. *Source: adaptation_tournament.md and the per-round reports cited inline*

Table 6: Round-by-round hypothesis → outcome. *Source: adaptation_tournament.md*

#	Hypothesis / probe	Outcome
1	Belief family at protocol strength	**Earliest promotion-gate clear**: entropy > fixed +5.8 pts [+3.7, +8.2] on Pick/seen, 3 seeds, no safety/cost regression. Random probing can *hurt* (−4.2 on Push/seen). *Source: adaptation_tournament.md*
2	DualABI + recurrent identifier	DualABI Pareto win: matches entropy success at 2.89 vs 6.00 probe steps (−52%). Recurrent identifier lifts discrete fields with horizon but permutation plateaus ∼ 0.35; end-to-end ∼ 0. *Source: adaptation_dualabi.md, adaptation_recurrent.md*

#	Hypothesis / probe	Outcome
3	Grammar-only belief (no pool)	Earliest pool-free belief to succeed: Pick/seen 0.458, Push/seen 0.673, 0.990 on the clean sub-cell; unseen collapses—scale non-identifiable under `pd_ee_delta_pose`. *Source: adaptation_factorized.md*
4	Delay-aware augmented-state PPO	Long-lag split solved: Pick 0.528 (∼20×), Push 0.415 (∼2.7×) vs frozen reference. Composition, not identification. *Source: adaptation_delay_aware.md*
5	C++ fusion, gripper channel, PullCube	C++ scoring core parity 5.7×10^{-14}; gripper v2 channel (R^2 0.545) closes wall 1 (0.000→0.537); PullCube gated in, entropy +5.5 on unseen = second gate-clear. *Source: cpp_fusion.md, adaptation_grasp_channel.md, third_task.md*
6	Probe-augmented learned identifier	Honest negative that quantifies the wall: permutation caps 0.39 (< random ablation's 0.52), sign at chance, end-to-end ∼ 0. Active probing is not a substitute for hypothesis-space knowledge. *Source: adaptation_probe_osi.md*
7	Lag-completion self-corrections	Pipeline-flush downgraded to a Push-specific artifact (failed on PullCube); "fixed > entropy under lag" reverses on the delay-aware backbone. *Source: lag_completions.md*
8	PegInsertionSide retry	Exclusion upheld: 0/100 at 41M, 71.7M, 75.7M steps (< 0.20 floor); budget misread 250M→75M corrected by our own audit. *Source: peg_retry.md*
9	Weakness wave: Stack, frame v2, scale corrector	StackCube gated in (task 4; entropy *not* best here—passive saturates); frame axis de-degenerated (oracle 1.000 vs 0.000); scale corrector proven where precondition holds. *Source: fourth_task.md, rotation_v2.md, adaptation_scale_corrector.md*
10	External anchor: DCAC delayed-RL	Augmented-state recipe is delay-robust on DCAC's delay-5 MuJoCo but not competitive on absolute return; Vintix ICL adjudicated unfaithful. No DCAC-beating claim. *Source: delayed_rl_external.md, transformer_icl_adjudication.md*
11	Hold-probe absolute-mode excitation	All-absolute wall breached: 0.005→0.722 (Push), 0.000→0.790 (Pick), exact permutation recovery, causal controls. Residual: one contract still 0.44 after exact ID. *Source: absolute_excitation.md*
12	Frozen Diffusion Policy backbone	Imitation is exactly as brittle and as rescuable: clean 0.58/0.67 → ∼0.003 hidden → 0.62–0.735 restored. Brittleness is interface-, not paradigm-, specific. *Source: imitation_brittleness.md*
13	Off-policy anchor: augmented-state SAC	Reaches DCAC's augmented-SAC/RTAC baseline tier (HalfCheetah 1951 vs ∼2000) while naive SAC collapses; does not reach DC/AC. *Source: delayed_rl_sac.md*

C Promotion-gate audit

A method earns a longer run only by clearing a preregistered gate: ≥ 5 points over a matched baseline, 30% faster recovery, a meaningful unseen/lag gain, or a success-vs-cost Pareto point, all without a safety regression; a sub-0.3 one-seed result is never promoted. Table 7 records every adjudication—cleared *and* failed.

Table 7: Promotion-gate adjudications. *Source: adaptation_tournament.md, third_task.md, fourth_task.md, adaptation_dualabi.md*

Method / cell	Gate criterion	Evidence	Verdict
Entropy vs fixed, Pick/seen	$\geq +5$ pts, no regression	$+5.8$ $[+3.7, +8.2]$, 3-seed, -24% displacement	**CLEAR**
Entropy vs fixed, Pull/unseen	$\geq +5$ pts	$+5.5$ $[+3.3, +7.8]$, 3-seed	**CLEAR**
DualABI, all 4 tasks	success-vs-cost Pareto	tie success, 2.89 vs 6.00 probe steps (-52%)	**CLEAR**
Entropy vs fixed, Push/seen	$\geq +5$ pts	$+1.8$ $[+0.8, +3.0]$ (right direction, below margin)	fail (ns-margin)
Entropy vs fixed, Stack/unseen	$\geq +5$ pts	$+2.5$ ns (passive already saturates)	fail
Entropy, StackCube/seen	top belief method	*lowest* belief method (0.910)	fail
Random vs passive, Push/seen	no safety/cost regression	-4.2 $[-6.3, -2.0]$ (probing hurts)	fail
Passive UP-OSI identifier	$\geq +5$ over no-adapt	0.0 end-to-end (perm 0.29, sign chance)	fail
Recurrent / probe-OSI identifier	any end-to-end success	~ 0; permutation caps 0.35/0.39	fail
DualABI, long-lag	Pareto under lag	collapses with belief family (0.02/0.11)	no claim

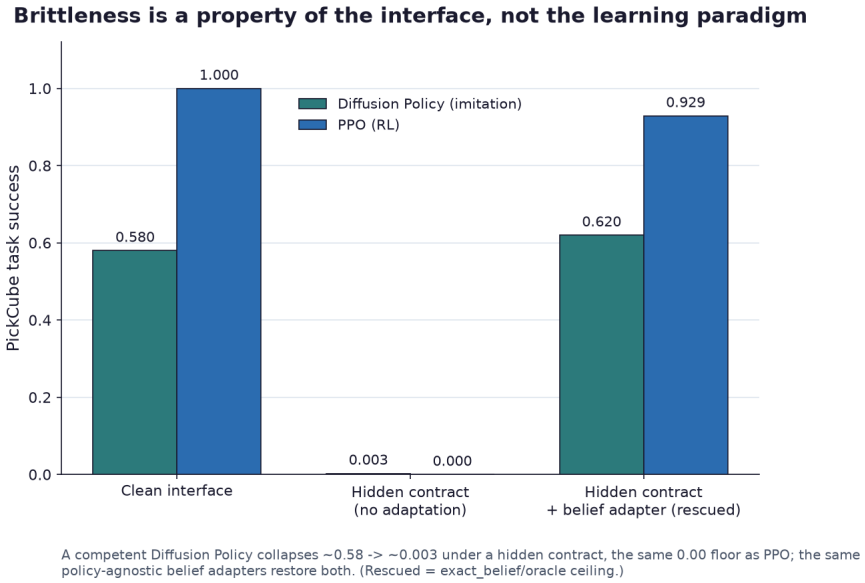

Figure 7: Imitation brittleness and rescue. *Source: `imitation_brittleness.md`*

D Reproducibility details

Budgets, seeds, hardware. Backbones: ManiSkill v3.0.1 pinned at `a4a4f9272ad64b1564035874b605ceb687b63ed8`, state observations, `pd_ee_delta_pose`, on four NVIDIA RTX 5090 GPUs. Gate-0 PPO budgets: Pick 10M, Push 2M, Pull 2M, Stack 25M (num_envs= 1024; true official 50M at 4096—no budget-efficiency claim); Peg $\geq$ 75M and excluded at 0/100. Seeds 20260718, 20260719, 20260720 throughout (the fast controlled gate additionally used 20260721/22). Gate 1 is 7,200 episodes (36 hash-addressed jobs, 200 episodes each); tournament cells are 600 episodes (two contracts $\times$ 100 $\times$ three seeds) except the one-seed cells (DualABI's long-lag Pick/Push and all StackCube cells); the belief-family long-lag cells are full three-seed.

Wrapper parity and safety. Maximum oracle decode error on a real smoke was 4.66×10^{-10}; the ActionABI C++ scoring backend reproduces the torch belief scores to 5.7×10^{-14} (float64) with bit-identical MAP decisions. Pick/Push pass Gate-0 parity under both the two-percentage-point rule and the overlapping-Wilson-interval rule. Hard fail-closed checks (finiteness, bounds, workspace, collision) gate every run; a unit test asserts unprivileged adapters take no contract argument, and training-contract distributions are hash-rejected against every frozen evaluation contract. *Source: `gate0.md`, `gate1.md`, `cpp_fusion.md`*

Model budgets. Feedforward actor-critic 78,863 params vs recurrent 78,208 (0.83% apart, within 10%)—the recurrent robustness baseline is not advantaged by capacity; OSI/RMA adaptation modules add separately reported inference parameters. Probe amplitude 0.5, budget 6 steps (gripper never actuated during probing), DualABI stop threshold frozen after a 48-episode pilot (robust $\tau \in [0.4, 1.5]$). *Source: `baseline_budgets.json`, `adaptation_dualabi.md`*

Provenance. All episodes are hash-addressed; runs write append-only episode JSONL and summaries with checkpoint SHA-256 per job. `ruff`, strict `mypy`, and the full test suite pass. Large checkpoints and raw JSONL are local/gitignored; manifests, verdicts, reports, and figures are committed.

References

[1] Remi Cadene, Simon Alibert, Francesco Capuano, Michel Aractingi, Adil Zouitine, et al. Lerobot: An open-source library for end-to-end robot learning. In *International Conference on Learning Represen-*

tations (ICLR), 2026. arXiv:2602.22818.

[2] Yangtao Chen, Zixuan Chen, Nga Teng Chan, Junting Chen, Junhui Yin, Jieqi Shi, Yang Gao, Yong-Lu Li, and Jing Huo. Robohiman: A hierarchical evaluation paradigm for compositional generalization in long-horizon manipulation. *arXiv preprint arXiv:2510.13149*, 2025.

[3] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems*, 2023. arXiv:2303.04137.

[4] Kaize Deng and Zewen Yang. Residual reinforcement learning for robot teleoperation under stochastic delays. *arXiv preprint arXiv:2605.15480*, 2026.

[5] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

[6] Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. Libero-plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025.

[7] Armin Karamzade, Kyungmin Kim, JB Lanier, Davide Corsi, and Roy Fox. Model-based reinforcement learning under random observation delays. *arXiv preprint arXiv:2509.20869*, 2025.

[8] Konstantinos V. Katsikopoulos and Sascha E. Engelbrecht. Markov decision processes with delays and asynchronous cost collection. *IEEE Transactions on Automatic Control*, 48(4):568–574, 2003.

[9] Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. Rma: Rapid motor adaptation for legged robots. In *Robotics: Science and Systems*, 2021. arXiv:2107.04034.

[10] Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Hansen, Angelos Filos, Ethan Brooks, Maxime Gazeau, Himanshu Sahni, Satinder Singh, and Volodymyr Mnih. In-context reinforcement learning with algorithm distillation. *arXiv preprint arXiv:2210.14215*, 2022.

[11] Jonathan N. Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. Supervised pretraining can learn in-context reinforcement learning. *arXiv preprint arXiv:2306.14892*, 2023.

[12] Qing Lian, Kent Yu, and Lei Zhang. Reflective vla: In-context action consequences make vlas generalize. *arXiv preprint arXiv:2606.25215*, 2026.

[13] Pierre Liotet, Erick Venneri, and Marcello Restelli. Learning a belief representation for delayed reinforcement learning. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021.

[14] NVIDIA. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

[15] Open X-Embodiment Collaboration. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

[16] Andrey Polubarov, Nikita Lyubaykin, Alexander Derevyagin, Ilya Zisman, Denis Tarasov, Alexander Nikulin, and Vladislav Kurenkov. Vintix: Action model via in-context reinforcement learning. *arXiv preprint arXiv:2501.19400*, 2025.

[17] Wilbert Pumacay, Ishika Singh, Jiafei Duan, Ranjay Krishna, Jesse Thomason, and Dieter Fox. The colosseum: A benchmark for evaluating generalization for robotic manipulation. *arXiv preprint arXiv:2402.08191*, 2024.

[18] Simon Ramstedt, Yann Bouteiller, Giovanni Beltrame, Christopher Pal, and Jonathan Binas. Reinforcement learning with random delays. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2010.02966; code github.com/rmst/rlrd.

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

[20] Bhavya Sukhija, Stelian Coros, Andreas Krause, Pieter Abbeel, and Carmelo Sferrazza. Maxinforl: Boosting exploration in reinforcement learning through information gain maximization. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2412.12098.

[21] Jianwei Tai. Same weights, different robot: A deployment safety view of vla policies. *arXiv preprint arXiv:2606.03724*, 2026. Unreviewed single-author cs.CR preprint.

[22] Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, et al. Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai. *arXiv preprint arXiv:2410.00425*, 2024.

[23] Thomas J. Walsh, Ali Nouri, Lihong Li, and Michael L. Littman. Learning and planning in environments with delayed feedback. In *Autonomous Agents and Multi-Agent Systems*, 2009.

[24] Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

[25] Zenan Wu, Bingqing Wei, et al. Atom-bench: A real-world benchmark for atomic skills and compositional generalization in manipulation policies. *arXiv preprint arXiv:2606.16826*, 2026.

[26] Annie Xie, Logan Mondal Bhamidipaty, Evan Zheran Liu, Joey Hong, Sergey Levine, and Chelsea Finn. Learning to explore in pomdps with informational rewards. In *International Conference on Machine Learning (ICML)*, 2024.

[27] Wenhao Yu, Jie Tan, C. Karen Liu, and Greg Turk. Preparing for the unknown: Learning a universal policy with online system identification. In *Robotics: Science and Systems*, 2017. arXiv:1702.02453.

[28] Simon Sinong Zhan, Qingyuan Wu, Philip Wang, Frank Yang, Xiangyu Shi, Chao Huang, and Qi Zhu. Belief-based offline reinforcement learning for delay-robust policy optimization. *arXiv preprint arXiv:2506.00131*, 2025.

[29] Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. Varibad: A very good method for bayes-adaptive deep rl via meta-learning. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:1910.08348.