TARGET-SPECIFIC DISTORTION IN STATE SUPERVISION UNDER POLICY SWAP

Anonymous authors
Paper under double-blind review

ABSTRACT

History encoders for partially observable control are often trained on hidden-state labels logged by a privileged teacher that sees the state, then deployed with a policy that sees only past observations and actions. The labels are correct, but the teacher's actions also reveal what it saw, so training fits a posterior that this policy swap invalidates. Prior analyses treat the hidden state as a whole. We show that the swap distorts a prediction target exactly when, at histories that deployment reaches, the teacher's action likelihood averaged over hidden paths consistent with the history varies with the target's value, and that only the part of the teacher's action information about the target can distort it, although this part need not do so. The exact posterior fitted to the log scores better on teacher histories and worse at deployment, and an exact identity splits a trained predictor's gap into distortion and fitting error. In synthetic channels with fixed histories and action information, coupling the target to the teacher's actions alone reverses the learned risk ordering in every seed, and where this condition and per-target action information disagree, trained predictors follow the condition. On the RockSample benchmark, predictors trained on exact posterior labels show the deployment gap in every seed at every tested level of teacher privilege. At full privilege, fixed controllers acting on them earn more return with deployment-posterior labels, fitting error rather than exact distortion sets the gap's size, and validation on teacher histories favored early checkpoints of the distorted predictors, but only under a loss restricted to targets the history leaves undetermined. When the environment model is known or identifiable from the state-labeled log, relabeling with the deployment posterior removes the distortion. Known teacher propensities allow only a high-variance reweighted evaluation, and otherwise the remedy is history-only data collection.

1 INTRODUCTION

History encoders approximate the belief of a partially observable Markov decision process, POMDP, that is, the posterior over its hidden state, by predicting simulator state from public histories of observations and actions (Smallwood & Sondik, 1973; Kaelbling et al., 1998; Wang et al., 2023). Their state labels often come from a fixed log of a teacher that acted on the hidden state, as in offline driving models (Hu et al., 2022; Shao et al., 2023; Jia et al., 2024) and learned filters, where a recurrent filter fails once the policy changes (Jonschkowski et al., 2018, Sec. V-A.4). The encoder is then deployed with a policy that sees only the history, a replacement we call the *policy swap*. Every label in the log is correct, but the teacher's actions both control the environment and reveal what the teacher saw, and a history-only policy keeps only the control role. The swap therefore moves the correct predictor from the *behavior posterior*, which uses the evidence in the teacher's action choices, to the *do posterior*, which drops it. A student collecting its own histories with a history-only policy, as in on-policy training or dataset aggregation, DAgger, once it stops executing teacher actions (Ross et al., 2011), already trains on its deployment posterior. We study the reuse of a fixed teacher log.

Figure 1a shows why the question is asked per target. A teacher whose action matches a hidden fair bit X with probability 0.9 shifts the posterior of X but leaves an independent fair bit Y untouched. Setting $Y = X$ keeps the teacher, the 0.368 nats of information its action carries about the hidden bits and the deployment entropy of Y of 0.693 nats, yet raises the exact excess deployment log loss on Y from 0 to 0.511 nats. The amount of action information cannot say which target the swap distorts.

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.

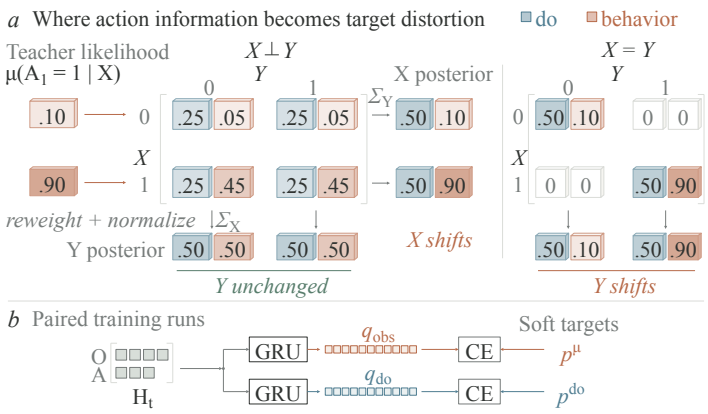


Figure 1: **a** The teacher's action matches the hidden bit X with probability 0.9. Reweighting by its likelihood shifts the posterior of X in both panels and that of Y only when $Y = X$. **b** Matched arms: paired gated recurrent unit networks, GRUs, fit p^{μ} or p^{do} on the same histories by cross-entropy, CE.

Prior work separates the observational and interventional readings of an action for the latent state as a whole (Ortega et al., 2021; Swamy et al., 2022), supplies the conditional importance-weight identity that we project onto a target (Rowland et al., 2020, Prop. 5.1), and decides recoverability under selection bias one query at a time (Bareinboim et al., 2014). Three questions remain open for a state-labeled teacher log: which individual targets the policy swap invalidates, where the teacher's action information goes, and what finite learners and checkpoint rules do with such a log.

We answer them with one mechanism. Reweighting latent paths by the teacher's action likelihoods tilts the do posterior into the behavior posterior, and on a single target the tilt survives exactly when, at histories that deployment reaches, the teacher's mean path weight varies with the target's value, which we call *residual selection*. From exact posteriors to trained networks, we make three contributions.

- **Mechanism.** The residual-selection criterion, Corollary 1, and an exact split of the teacher's action information into a part absorbed by later observations, a part about other path components and an on-target residual. Distortion requires a nonzero on-target residual, a residual does not imply distortion, and per-target action information errs in both directions.
- **Protocol and audit.** Matched-label arms differ only in the label, and an exact identity splits each learned gap into distortion and a signed fitting term. In post hoc tests, a sound per-target certificate that at full privilege needs no numeric action probabilities recovers more undistorted targets than near-zero per-target action information, and at looser KL tolerances only action-information thresholds that clear targets above the tolerance recover more.
- **Evidence.** In controlled channels the label alone reverses learned risk in all ten seeds, and in post hoc variants where per-target action information and PSPD disagree the learned gap follows PSPD. In RockSample the deployment gap appears in every seed across width, learning rate, label type, teacher mixture and number of training layouts, and between trained checkpoints under the all-target loss; fitting error rather than exact distortion sets its size at full privilege. Validation on teacher histories failed in RockSample only under the constrained loss at full privilege, where it picked early checkpoints of the distorted arms in all 15 runs and, with 64 layouts of 256 episodes, observational-soft at or after update 2,000 in every seed. In a post hoc test at full privilege, two fixed controllers earned more return with do-soft than with observational-soft in every seed.

2 Related Work

Observational and interventional semantics. Observational and interventional action semantics separate what an action reveals about hidden state from what it does (Ortega & Braun, 2010; Ortega et al., 2021). Swamy et al. (2022) derive context posteriors from privileged demonstrations that keep the expert-action likelihoods off policy and drop them on policy. Section 4 applies the same change of measure to latent state paths, and Corollary 1 is the conditional importance-weight identity of Rowland et al. (2020, Prop. 5.1) with a state target in place of the return. Recoverability under selection bias is likewise decided per query (Bareinboim et al., 2014); our criterion instantiates this for a target marginal under a policy swap. Causally correct partial models adjust for the behavior

policy (Rezende et al., 2020), and work on causal confusion in imitation (de Haan et al., 2019; Wen et al., 2020; Kumor et al., 2021), confounder shift (Tennenholtz et al., 2022), deconfounded imitation (Vuorio et al., 2024) and identification in confounded POMDPs (Shi et al., 2022; Hong et al., 2024; Gasse et al., 2023) identifies a policy, a value or a model. We ask instead which target marginals of a state-labeled log survive the swap and how learners and checkpoint rules respond.

Belief learning and privileged supervision. Belief learning spans variational, flow, recurrent and imitation methods (Igl et al., 2018; Chen et al., 2022; Ni et al., 2022; Gangwani et al., 2020). Direct belief supervision (Nguyen et al., 2021; Wang et al., 2023) trains on the learner's own rollouts, which is the history-only regime. Privileged information also underlies asymmetric critics and distillation (Pinto et al., 2018; Chen et al., 2020; Baisero et al., 2022; Lambrechts et al., 2025; Luo et al., 2025; Rusu et al., 2016; Cai et al., 2024), imitation-gap analyses (Weihs et al., 2021; Walsman et al., 2023), teachers adapted to the student's information (Warrington et al., 2021; Messikommer et al., 2025) and selectively queried teachers (Kim et al., 2025). The nearest analogues ask which privileged object is a valid critic target (Baisero & Amato, 2022) and when imitation is identifiable from a causal graph (Zhang et al., 2020). None asks per target whether a teacher's state label survives the swap.

3 Setting and Notation

We write S_t for the Markov state, O_t for the public observation, A_t for the action, $H_t = (O_{1:t}, A_{1:t-1})$ for the public history that deployment sees and $Z_t = g(S_t)$ for a prediction target, a specified state component. The teacher μ sees the state. Its trajectory law multiplies $p(s_1)G_1(o_1 \mid s_1)$ by the action factor $\mu(a_t \mid s_t, h_t)$, the transition $T(s_{t+1} \mid s_t, a_t)$ and the observation $G(o_{t+1} \mid s_{t+1}, a_t)$ at every step, and equation 12 in Appendix A.1 writes it out. A deployable policy satisfies $\pi(a_t \mid s_t, h_t) = \pi(a_t \mid h_t)$. The benchmark teachers mix a state-aware policy with a history-only one,

$$\mu_{\lambda,\epsilon}(a \mid s, h) = (1-\epsilon)\big[\lambda\mu_{\text{state}}(a \mid s, h) + (1-\lambda)\pi_{\text{hist}}(a \mid h)\big] + \epsilon U(a), \tag{1}$$

where $\lambda \in [0, 1]$ sets the teacher's privilege, $\epsilon > 0$ gives full support and $\lambda = 0$ makes actions independent of the state given the history. Section 4 holds for any fully supported teacher. The two posteriors are

$$b_t^{\mu}(s \mid h_t) = P^{\mu}(S_t = s \mid H_t = h_t), \tag{2}$$

$$b_t^{\text{do}}(s \mid h_t) = P(S_t = s \mid O_{1:t} = o_{1:t}, \text{do}(A_{1:t-1} = a_{1:t-1}))\,. \tag{3}$$

The behavior posterior keeps the teacher's action likelihoods $\mu(a_k \mid s_k, h_k)$. The do posterior deletes them but keeps the recorded actions in T and G (Pearl, 2009). Every history-only π induces the do posterior, $P^{\pi}(S_t \mid H_t) = b_t^{\text{do}}$, because $\pi(a_k \mid h_k)$ is constant over latent paths and cancels. We write $b_t^{\mu,Z}$ and $b_t^{\text{do},Z}$ for the target marginals of the two posteriors.

Direct state supervision fits a predictor $q_\phi(z_t \mid h_t)$ to the teacher's labels by minimizing

$$\mathcal{L}_\mu(q) = \mathbb{E}_{(Z_t, H_t)\sim P^{\mu}}[-\log q(Z_t \mid H_t)], \tag{4}$$

whose population minimizer is $b_t^{\mu,Z}$ by Lemma 1. The loss rewards every history feature that predicts Z_t under the teacher, including the evidence in the teacher's action choices. After the swap the correct predictor is $b_t^{\text{do},Z}$, although every label stays correct. Table 1 collects the terms used below.

4 Mechanism: How Teacher Actions Distort One Target

4.1 Posterior tilt and the residual-selection criterion

We assume common reference densities, finite displayed expectations and $P^{\pi}(H_t) \ll P^{\mu}(H_t)$; Appendix A.2 covers continuous variables. At a fixed history h_t, reweighting latent paths by the teacher's action likelihoods tilts the do posterior into the behavior posterior,

$$P^{\mu}(ds_{1:t} \mid h_t) = \frac{W_{t-1}^{\mu}}{\mathbb{E}_{\text{do}}[W_{t-1}^{\mu} \mid h_t]} P^{\text{do}}(ds_{1:t} \mid h_t), \qquad W_{t-1}^{\mu} = \prod_{k=1}^{t-1} \mu(a_k \mid s_k, h_k), \tag{5}$$

Table 1: Terms used in this paper.

Term	Meaning
Behavior posterior b^μ	$P^\mu(S_t \mid H_t)$, which keeps the evidence in the teacher's action choices.
Do posterior b^{do}	Drops the teacher's action likelihoods; every history-only π induces it.
Residual selection	At histories that deployment reaches, the teacher's mean path weight varies with the target's value; by Corollary 1 exactly this distorts a target.
PSPD	Deployment mean of $D_{\mathrm{KL}}(b^{\mathrm{do},Z} \| b^{\mu,Z})$, the exact behavior posterior's excess deployment log loss.
Observational-soft, do-soft, DSS	GRUs trained on the same teacher histories with b^μ marginals, b^{do} marginals or realized hard labels; DSS is direct state supervision.
Unresolved, constrained loss	A coordinate is unresolved while the public history leaves it open; the constrained loss scores only these, the all-target loss every coordinate.
$D_{\mathrm{marg}}, E_{\mathrm{do}}, B_{\mathrm{obs}}, G$	Exact distortion, do-side and signed observational fitting terms, learned gap.
Selected, final checkpoint	Lowest validation loss against the arm's own training target, or update 20,000.
Behavior-label, deployment-label validation	Choosing by realized-label negative log-likelihood, NLL, on teacher histories of the validation layouts, or on history-only rollouts from unseen-A layouts.

where W^μ_{t-1} is the path weight and $\mathbb{E}_{\mathrm{do}}$ the do expectation. We measure the cost on the target marginal. The policy-swap posterior divergence, PSPD, is the excess deployment log risk of the exact behavior posterior, Proposition 1, in Kullback–Leibler divergence D_{KL},

$$\mathrm{PSPD}_t(\mu \to \pi) = \mathbb{E}_{H_t \sim P^\pi} D_{\mathrm{KL}}\Big(b_t^{\mathrm{do},Z}(\cdot \mid H_t) \| b_t^{\mu,Z}(\cdot \mid H_t)\Big), \tag{6}$$

averaged over the scoring times. Projecting the tilt onto the target decides when PSPD is zero.

Corollary 1 (Residual selection; Rowland et al., 2020, Prop. 5.1)**.** *Assume mutually absolutely continuous conditional posteriors and* $0 < \mathbb{E}_{\mathrm{do}}[W^\mu_{t-1} \mid h] < \infty$*. Marginalizing equation 5 gives the Radon–Nikodym derivative*

$$\frac{db_t^{\mu,Z}(\cdot \mid h)}{db_t^{\mathrm{do},Z}(\cdot \mid h)}(z) = \frac{\mathbb{E}_{\mathrm{do}}[W^\mu_{t-1} \mid Z_t = z, h]}{\mathbb{E}_{\mathrm{do}}[W^\mu_{t-1} \mid h]}. \tag{7}$$

Thus $\mathrm{PSPD}_t = 0$ *exactly when this ratio is one* $P^\pi(dh)b_t^{\mathrm{do},Z}(dz \mid h)$*-almost surely. If* $Z_t = f_t(H_t)$ *under the do law on target support, both posteriors equal the same point mass and* $\mathrm{PSPD}_t = 0$*.*

The swap therefore moves a target marginal exactly when the teacher's mean path weight at a history varies with the target's value on a set of positive deployment probability, which is residual selection. The condition concerns conditional means rather than pathwise invariance, so it can hold for one state component and fail for another under the same teacher. If W^μ_{t-1} depends on the latent path only through components U with $Z_t \perp U \mid H_t$ under the do law, the ratio is one, as for the independent bit of Figure 1a; by Appendix A.6 this is not necessary.

4.2 Risk reversal for exact posteriors

The time-t risk of q under policy ν is $\mathcal{R}_{\nu,t}(q) = \mathbb{E}_{(Z_t,H_t)\sim P^\nu}[-\log q(Z_t \mid H_t)]$.

Proposition 1 (Behavior advantage and deployment disadvantage)**.** *Let* π *be a history-only policy with* $P^\pi(H_t) \ll P^\mu(H_t)$*, and suppose* $b_t^{\mu,Z}$ *and* $b_t^{\mathrm{do},Z}$ *are mutually absolutely continuous on the histories where they are evaluated. Then*

$$\mathcal{R}_{\mu,t}(b^{\mathrm{do},Z}) - \mathcal{R}_{\mu,t}(b^{\mu,Z}) = \mathbb{E}_{P^\mu(H_t)} D_{\mathrm{KL}}(b_t^{\mu,Z} \| b_t^{\mathrm{do},Z}) \geq 0, \tag{8}$$

$$\mathcal{R}_{\pi,t}(b^{\mu,Z}) - \mathcal{R}_{\pi,t}(b^{\mathrm{do},Z}) = \mathbb{E}_{P^\pi(H_t)} D_{\mathrm{KL}}(b_t^{\mathrm{do},Z} \| b_t^{\mu,Z}) \geq 0. \tag{9}$$

The exact behavior posterior thus scores better on teacher histories and worse after the swap, so validation on teacher histories rewards exactly the posterior that the swap invalidates.

4.3 Where the teacher's action information goes

The teacher's action information up to time $t-1$, $\mathrm{TAL}_{t-1}(\mu) = \sum_{k<t} I_\mu(S_k; A_k \mid H_k)$, is a trajectory KL divergence between the teacher's law and a history-only surrogate, Proposition 2. The KL chain rule, applied to this divergence and to equation 7, splits it exactly, Corollary 2:

$$\mathrm{TAL}_{t-1}(\mu) = \mathcal{I}_t^{\mathrm{hist}} + \mathcal{I}_t^{\mathrm{off}} + \mathcal{I}_t^{\mathrm{on}}, \qquad \mathcal{I}_t^{\mathrm{on}} = \mathbb{E}_{H_t \sim P^\mu} D_{\mathrm{KL}}\big(b_t^{\mu,Z}(\cdot \mid H_t) \| b_t^{\mathrm{do},Z}(\cdot \mid H_t)\big). \tag{10}$$

Later observations carry $\mathcal{I}_t^{\text{hist}}$ into the public history, $\mathcal{I}_t^{\text{off}}$ concerns path components other than the target once target and history are fixed, and the on-target residual $\mathcal{I}_t^{\text{on}}$ is the behavior advantage in equation 8. Only the residual measures how far the target marginal moves on teacher histories, so distortion requires $\mathcal{I}_t^{\text{on}} > 0$. The residual is necessary but not sufficient, because PSPD counts only the distorted histories that deployment reaches. In the support example of Table 5 in Appendix A.12, a residual of 0.128 nats leaves PSPD at 0 because deployment avoids the revealing actions.

The split also shows why per-target action information, Table 5, does not replace the criterion. Per-step action information about the target component, $\iota_t^{\text{act}} = \sum_{k<t} I_\mu(g(S_k); A_k \mid H_k)$, errs in both directions: ι_t^{act} is 0.368 nats with PSPD 0 when a later observation reveals the target, because the evidence moves into $\mathcal{I}_t^{\text{hist}}$, and 0 with PSPD 0.511 nats when the dynamics move the bit the teacher acted on into the target. Per-step action information about the scored target, $\iota_t^{\text{fwd}} = \sum_{k<t} I_\mu(Z_t; A_k \mid H_k)$, is 0.368 nats with PSPD 0 when a history-only teacher's actions merely control the target. Given all observations and without control effects, the joint action information about the scored target, ι_t^{obs}, equals the residual. Targets also disagree within one history: in 40.4% of RockSample deployment histories with two or more unresolved rocks under a fully privileged teacher, one such rock with do entropy of at least 0.05 nats keeps its marginal within 10^{-6} nats of KL while another is distorted by over 1 nat.

4.4 Exact distortion versus fitting error

A fitted predictor need not reach its population target, so a paired gap mixes the exact gap of Proposition 1 with fitting error. Take target coordinates j, the eleven rock marginals in RockSample, reference marginals $p_j = b_j^{\text{do}}$ and $b_j = b_j^{\mu}$, and paired predictors $q_{\text{obs},j}$ and $q_{\text{do},j}$. We average over a deployment history measure d with public target weights $u_j(h) \in [0, 1]$, which in RockSample mark the unresolved rocks. The gap G in marginal do-reference risk satisfies

$$G = D_{\text{marg}} - E_{\text{do}} + B_{\text{obs}}, \qquad B_{\text{obs}} = \int \sum_j u_j \mathbb{E}_{p_j} \log \frac{b_j}{q_{\text{obs},j}} \, d(dh), \tag{11}$$

where the exact distortion D_{marg} sums $D_{\text{KL}}(p_j \| b_j)$ and the do-side fitting error E_{do} sums $D_{\text{KL}}(p_j \| q_{\text{do},j})$, both with weights u_j under d. The term B_{obs} is signed: it averages under p_j, whereas the observational predictor's own fitting error E_{obs} averages under b_j, so fitting can move the gap either way. A negative B_{obs} places G below the exact distortion, and only a positive one can place it above. Appendix A.4 gives coverage conditions, counterexamples and a sufficient bound.

5 Experiments

5.1 Setup

Environments. *Controlled channels* extend Figure 1a to eleven hidden pairs (X_j, Y_j). The parity of the teacher's action equals a queried X_j with probability 0.9, actions control both channels, and the targets are the Y_j. The coupling $\rho \in \{0, 0.5, 1\}$ is the probability that Y_j shares the innovation of X_j, and a public cue bit keeps its value with probability 0.9, or 0.6 with shifted cues. By Proposition 4, the three couplings share their public trajectories, the teacher's action information at every history and the recursion of target uncertainty under deployment, so only the coupling changes. *RockSample* 11×11 has controlled sensing and 2,048 rock configurations (Smith & Simmons, 2004); teachers follow equation 1 with $\epsilon = 0.05$. A rock is *unresolved* until the public history determines its quality, and the *Public* baseline outputs determined qualities and $1/2$ otherwise. Teacher histories come from eight training, four validation and four held-out behavior layouts. Deployment histories come from the training layouts and two unseen groups of four layouts, unseen-A for deployment-label validation and unseen-B held out, under four history-only policies: random, history-greedy, history-cautious and different-seed, the last three acting on exact do marginals. Two deployment measures weight the four policies equally: the *three-group measure* of Tables 2A and 4A–B weights the training, unseen-A and unseen-B layouts by $1/2$, $1/4$ and $1/4$, and the *deployment pool* of Tables 3 and 4C weights the training and unseen-B layouts equally. Deployment numbers in the text use the measure of the table or appendix they cite.

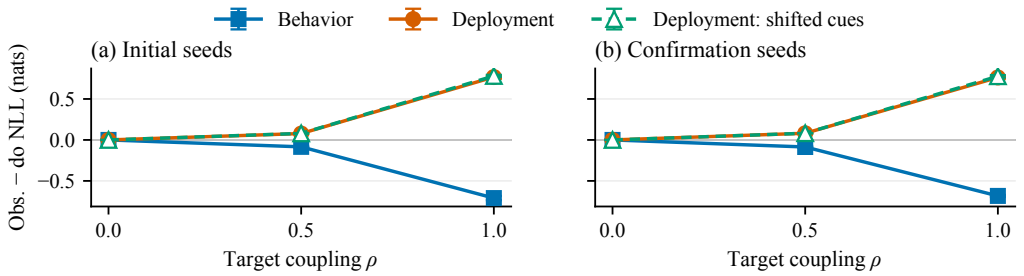


Figure 2: Controlled channels: coupling alone reverses learned risk. Observational-soft minus do-soft realized NLL at selected checkpoints, two groups of five seeds; deployment averages two policies at cue retention 0.9, dashed 0.6. Paired 95% intervals lie within the markers.

Arms and studies. All arms are 128-unit GRUs with eleven Bernoulli heads. Within a seed they share initialization, minibatch order and a 20,000-update budget and differ only in the label of the same teacher histories. Four studies are prespecified. Two were fixed before collection: the channel protocol with two groups of five seeds, and the RockSample training study, whose forty models cross five seeds, both soft arms, both losses and $\lambda \in \{0, 1\}$. The sensitivity study retrains the constrained recipe at $\lambda = 1$ with 256 units or learning rate 10^{-4}, and its hard-label extension adds DSS and a $\lambda = 0$ control; both were fixed before training. The teacher-mixing study trains the default recipe with the same seeds on logs collected at $\lambda = 0.5$, its primary mixture, 0.25 and 0.75, under a protocol fixed before collection and after the $\lambda \in \{0, 1\}$ results were known. The decompositions, the heterogeneity count, the CartPole increments, the action-information probe and Appendices B.5, C.5, C.8 and C.9 are post hoc.

Metrics and checkpoints. Realized NLL scores the realized labels, reported as a difference between arms, and against-do KL $D_{\mathrm{KL}}(b^{\mathrm{do}} \| q)$ is the excess deployment log loss over the do posterior. RockSample losses sum over unresolved coordinates and frames and divide by 60 steps, and at evaluation every predictor's outputs for determined rocks take their public values. Checkpoints are saved every 250 updates, and Table 1 defines the selection rules.

5.2 Controlled channels

At $\rho = 1$, observational-soft lowers behavior NLL and raises deployment NLL in all ten seeds, Figure 2. Its deployment excess is $+0.768$ nats in the initial group, 95% interval $+0.747$ to $+0.787$, and $+0.763$ in the confirmation group. At $\rho = 0.5$ both directions repeat at about a tenth of the size, -0.085 on behavior and $+0.080$ at deployment over ten seeds, and shifted cues keep the reversal. At $\rho = 0$ both soft arms fit the same target and coincide, a structural control. A post hoc DSS arm trails do-soft by $+0.191$ nats at $\rho = 0$ and its label noise cancels the behavior advantage at $\rho = 1$, but coupling still moves it in all ten seeds, by -0.173 nats on behavior and $+0.206$ at deployment.

A post hoc paired test, Appendix B.5, sets PSPD against per-target action information in channel variants without control. When a pad revealed one frame later masks the teacher's parity, per-step action information about every target is zero, yet observational-soft wins on teacher histories and loses at deployment by 0.773 nats, 95% interval 0.765 to 0.781, in every seed. Under uniform deployment, where it has the labels and PSPD of an unmasked copy, its gap is 0.980 times the copy's. Moving the acted-on bit into the target gives 0.634 nats, and both variants meet the prespecified rule. A target revealed only through probes gives 0.213 when deployment probes and 0.001, inside the prespecified band of ± 0.02 nats, when it never does.

5.3 RockSample main result

The constrained loss makes the do-posterior predictor accurate. In Table 2A it lowers do-soft deployment KL from 1.232 to 0.146 nats between final checkpoints, against 1.967 for Public. Observational-soft never approaches its own target. Own-target selection stops it at update 250 or 500 in every $\lambda = 1$ run, because later checkpoints fit the training layouts better but the validation layouts

Table 2: RockSample main result at $\lambda = 1$: the constrained loss makes do-soft accurate, and the reversal persists. Bold: constrained do-soft in A and selected checkpoints in B.

A. Unresolved-target deployment KL loss mass (nats; lower is better)

Supervision	All-target	Constrained	Constrained − all-target
Selected checkpoints: two primary effects, 97.5% intervals adjusted over both (Appendix C.1)			
Observational-soft	3.052	2.549	$-0.504\,(-0.644, -0.407)$
Do-soft (all-target selected = initialization)	1.987	**0.147**	$\mathbf{-1.841}\,(-1.856, -1.824)$
Final checkpoints (trained versus trained): secondary effects, unadjusted 95% intervals			
Observational-soft	14.844	11.865	$-2.979\,(-3.642, -2.316)$
Do-soft	1.232	**0.146**	$\mathbf{-1.086}\,(-1.196, -0.975)$
Public reference	1.967	1.967	–

B. Realized NLL, constrained loss: observational-soft − do-soft (nats)

Checkpoint	Training behavior	Held-out behavior	Training deployment
Selected	$\mathbf{-0.480}\,(-0.505, -0.451)$	$\mathbf{-0.228}\,(-0.260, -0.205)$	$\mathbf{+2.533}\,(+2.457, +2.602)$
Final	$-1.018\,(-1.042, -0.994)$	$+2.095\,(+1.860, +2.289)$	$+11.127\,(+10.917, +11.337)$

Loss masses on unresolved coordinates. A: against-do KL on the three-group measure of Section 5.1; the do posterior scores 0. B: realized NLL on new histories, deployment averaging the four policies; Table 3 gives the all-target loss. Paired seed-bootstrap percentile intervals.

Table 3: Ablations of the RockSample reversal at own-target-selected checkpoints, with $\lambda = 1$ unless a row states otherwise. Cells give realized-label NLL differences in nats, observational-soft minus do-soft unless the row names DSS, as the five-seed mean and the number of seeds sharing its sign. Training and Held-out: behavior histories of the training and held-out layouts. Deployment: the deployment pool of Section 5.1. Picks: seeds in which behavior-label validation, ranking every checkpoint of the compared arm and do-soft, selects the compared arm. All-target own-target selection keeps do-soft at its initialization. Bold: deployment gaps. The $\lambda = 0.25, 0.5, 0.75$ rows train the default recipe on behavior logs collected at that mixture.

Variant	Training	Held-out	Deployment	Picks
Default: constrained loss, 128 units, lr 10^{-3}	-0.480 (5/5)	-0.228 (5/5)	$\mathbf{+2.388}$ (5/5)	5/5
All-target loss, do-soft at initialization	-0.682 (5/5)	-0.379 (5/5)	$\mathbf{+1.052}$ (5/5)	0/5
256 units	-0.559 (5/5)	-0.200 (5/5)	$\mathbf{+2.722}$ (5/5)	4/5
Learning rate 10^{-4}	-0.336 (5/5)	-0.324 (5/5)	$\mathbf{+1.079}$ (5/5)	5/5
Hard labels (DSS − do-soft)	-0.408 (5/5)	-0.241 (5/5)	$\mathbf{+2.647}$ (5/5)	5/5
Hard labels, $\lambda = 0$ control (DSS − do-soft)	$+0.347$ (5/5)	$+0.296$ (5/5)	$\mathbf{+0.998}$ (5/5)	0/5
$\lambda = 0.25$, soft labels	$+0.100$ (5/5)	$+0.175$ (5/5)	$\mathbf{+0.498}$ (5/5)	0/5
$\lambda = 0.5$, soft labels	-0.016 (3/5)	$+0.108$ (5/5)	$\mathbf{+0.731}$ (5/5)	0/5
$\lambda = 0.75$, soft labels	-0.299 (5/5)	-0.013 (3/5)	$\mathbf{+1.652}$ (5/5)	0/5

worse, Appendix C.3. In the default recipe its own-target KL there on new training-layout teacher histories is 0.168 nats per unresolved rock, not a loss mass, against 0.0007 for do-soft, Table 16.

Table 2B shows that the reversal persists: at selected checkpoints observational-soft wins on teacher histories of the training and held-out layouts and loses on training-layout deployment histories by 2.533 nats, in every seed; final checkpoints lose the held-out behavior advantage.

5.4 Ablations

Loss. Under the all-target loss, final observational-soft has deployment KL 14.844 nats against 1.232 for do-soft, Table 2A, higher in every seed, and behavior-label validation picks do-soft in every seed.

Capacity, learning rate and layouts. With 256 units or learning rate 10^{-4} every sign holds in every seed, and observational-soft stays at an own-target KL of 0.153 and 0.210 nats per unresolved rock. In a post hoc study with 64 layouts of 256 episodes, Appendix C.5, its KL on held-out layouts falls from 0.215 to 0.158 in every seed, meeting the fit rule, yet G stays at 13.8% of D_{marg} against 13.2%, as B_{obs} moves only from -15.666 to -15.605 nats, and from -11.464 to -11.090 on held-out teacher histories, where no swap occurs. Eight times the episodes on eight layouts change neither, 32

Table 4: Decomposition of the deployment gap by equation 11 on common histories, post hoc. Marginal loss masses in nats. Bold: $|B_{\text{obs}}| > 1$ nat.

A. Study and predictors	D_{marg}	E_{obs}	E_{do}	B_{obs}	G
Channels, $\rho = 0.5$, ten seeds	0.082	0.003	0.0004	-0.002	0.080
Channels, $\rho = 1$, ten seeds	0.765	0.003	0.0004	-0.006	0.759
RockSample, constrained, selected	18.083	2.504	0.147	$\mathbf{-15.535}$	2.402
RockSample, constrained, final	18.083	4.407	0.146	$\mathbf{-6.218}$	11.719

B. Policy	D_{marg}	B_{obs}	G
Random	1.583	$\mathbf{+2.498}$	3.768
History-greedy	22.999	$\mathbf{-20.955}$	1.952
History-cautious	23.192	$\mathbf{-21.103}$	1.997
Different-seed	24.560	$\mathbf{-22.578}$	1.891

C. Mixing	D_{marg}	E_{do}	B_{obs}	G	G/D_{marg}
$\lambda = 0.25$	0.527	0.090	$+0.065$	0.503	95.3%
$\lambda = 0.5$	1.318	0.089	-0.490	0.739	56.0%
$\lambda = 0.75$	3.448	0.098	$\mathbf{-1.693}$	1.656	48.0%
$\lambda = 1$	18.213	0.151	$\mathbf{-15.666}$	2.397	13.2%

Channels: selected checkpoints, cue retention 0.9, two deployment policies, both five-seed groups. A and B: RockSample at $\lambda = 1$ on unresolved coordinates of the three-group measure; B uses constrained selected models and each policy's layouts. C: default constrained recipe at selected checkpoints on the deployment pool, so its $\lambda = 1$ row differs from A. B_{obs} is computed directly and checked against the identity; E_{obs}, the observational arm's fitting error against b^{μ}, does not enter it.

episodes on each of 64 layouts lower the share in every seed, to 8.8%, and 256 layouts reach a KL of 0.131, against 0.001 for do-soft, and a share of 15.1%.

Teacher mixing. Logs collected at $\lambda = 0.25$, 0.5 and 0.75, Appendix C.6, keep the deployment gap positive in every seed, and the gap grows with λ. The behavior-side advantage shrinks as λ falls: observational-soft is worse than do-soft on every history set at $\lambda = 0.25$ and on held-out layouts at 0.5, and better on them in only 3 of 5 seeds at 0.75. As b^{μ} is Bayes-optimal on teacher histories, a positive behavior gap means only that observational-soft fits b^{μ} worse than do-soft. Both validation streams pick do-soft in every seed at every tested $\lambda < 1$.

5.5 What sets the size of the gap

In the channels, Table 4A, the fitting terms of equation 11 stay below 0.01 nats and the learned gap tracks the exact distortion, 0.080 against 0.082 nats at $\rho = 0.5$ and 0.759 against 0.765 at $\rho = 1$.

In RockSample, panel C follows one deployment pool across teacher mixtures. At $\lambda = 0.25$ the learned gap is 95.3% of the exact distortion, and E_{do} and B_{obs} stay below 0.1 nats although E_{obs} is 0.741. As λ grows, the exact distortion rises from 0.527 to 18.213 nats, but B_{obs} turns negative in every seed and cancels a growing share of it: G is 56.0%, 48.0% and 13.2% of D_{marg} at $\lambda = 0.5$, 0.75 and 1. At full privilege, fitting error rather than exact distortion therefore sets the size of the gap. In panel B the random policy has the smallest exact distortion but a positive B_{obs} in every seed, so its gap exceeds D_{marg}, while each history-driven policy has a large negative one. On the same histories, and so with the same D_{marg}, panel A shows that final checkpoints grow the gap through B_{obs} as observational-soft overfits the teacher histories. For imperfect fits the identity fixes neither the sign of G nor its trend, so these signs are empirical.

5.6 Hard labels and the $\lambda = 0$ control

DSS trains on realized labels, whose population target is also b^{μ}. At $\lambda = 1$ it beats do-soft on teacher histories and loses at deployment in all 15 setting-seed cases of the sensitivity study. The observational-soft minus do-soft gap accounts for 90.2–99.8% of the DSS minus do-soft gap across the three setting means, and 79.1–119.1% per seed, so exact soft b^{μ} labels reproduce most of it. Early stopping is not separated, as own-target selection stops DSS after 250–750 updates.

On history-only $\lambda = 0$ logs, which carry no distortion, the same seeds still give a DSS minus do-soft gap of $+0.998$ nats on the deployment pool, from label noise and early stopping. At $\lambda = 1$ it is larger by $+1.650$ nats, 95% interval $+1.480$ to $+1.844$, in every seed. This descriptive increment is only 9.1% of D_{marg}. In Position-Only CartPole, Appendix E.2, whose ideal PSPD is zero, target NLL rises from $\lambda = 0$ to $\lambda = 1$ by $+15.496$ nats for DSS and by $+6.021$ for distillation targets that ignore the teacher, in every seed, so changed histories alone can produce a large increment.

5.7 SELECTION, REMEDIES AND A TARGET CERTIFICATE

Selection. Behavior-label validation preferred observational-soft to do-soft in all ten channel seeds. In RockSample it failed only under the constrained loss at $\lambda = 1$, where among three arms it chose DSS in 11 cases and observational-soft in 4, always at update 250 or 500. Restricted post hoc to updates from 2,000 on, Appendix C.4, it chose do-soft in 10 of 15 cases but still failed in all five seeds at learning rate 10^{-4}. With 64 layouts of 256 episodes it chose observational-soft in all five seeds at updates 2,000–2,750. Deployment-label validation chose do-soft in all 15 cases yet did no better than the last do-soft checkpoint, and within the observational-soft and DSS arms it returned the untrained initialization in every eight-layout $\lambda = 1$ run, as distortion alone predicts: on the deployment pool the initialization's against-do KL is 1.990 nats at 128 units, the exact behavior posterior's 18.21.

Remedies. If the transition and observation models are known, or are identified from the state-labeled log by Proposition 3 of Appendix A.9, which needs coverage but not the teacher's propensities, do relabeling removes the distortion. If only the propensities are known, weights π/μ on teacher prefixes score a predictor at deployment from the log, Appendix A.10, but at $\lambda = 1$ with uniform π each step multiplies their second moment by 9.32. Otherwise the remedy is history-only collection, which in RockSample costs do-soft nothing: trained on $\lambda = 0$ instead of $\lambda = 1$ logs, its NLL on the deployment pool changes by -0.058 nats, negative in every seed.

Audit. Under RockSample's product law, Proposition 5, Appendix F, certifies a target when flipping its bit leaves every path weight unchanged, a sufficient pathwise form of Corollary 1. In a post hoc probe, Appendix C.7, it recovers 73.9% of undistorted unresolved targets at $\lambda = 1$ and 83.7% at $\lambda = 0.5$, against 41.1% and 47.2% for cumulative per-target action information below 10^{-20} nats, which needs the exact behavior posterior. At KL tolerances from 10^{-6} to 0.1 nats, Appendix C.8, the certificate clears no target above the tolerance and beats every action-information threshold that clears none, in every seed, whereas a threshold of 10^{-4} nats recovers 4.8–11.1 points more but clears targets above the tolerance, 2 and 7 of them by over 0.1 nats, though all it clears carry at most 4.5×10^{-4}% of D_{marg}. At $\lambda = 1$ the certificate uses only the public history, the layout and the teacher's preference map, holds for every positive ratio of the two action probabilities and stayed exact under every misspecified analyst model, under which that threshold and a plug-in KL below 10^{-3} nats cleared targets with true KL up to 16.3 and 24.4 nats; at $\lambda = 0.5$ it also uses the student's preference map and stayed exact only while that map came from the log. Building its training targets took 87.1–362.8 times as long as do relabeling, Appendix F.4, so it diagnoses a log rather than replacing do relabeling.

6 LIMITATIONS

The criterion needs the do path law and the teacher's propensities, and the certificate assumes independent rock qualities and the two-level action family of Appendix F. In RockSample observational-soft does not fit its own target even with 256 training layouts, so learnability and distortion are not separated there. The criterion concerns posteriors at fixed histories, not occupancy shift, which was small for do-soft in RockSample but not for distillation in CartPole. RockSample studies use five seeds and percentile intervals without guaranteed small-sample coverage. Control was tested post hoc with two fixed heuristic controllers, Appendix C.9: at $\lambda = 1$ do-soft minus observational-soft return was $+25.69$ at selected and $+11.83$ at final checkpoints, in every seed, but at $\lambda = 0.25$ and 0.5 the exact behavior posterior earned more than the exact do marginals, so there the learned gains reflect fitting error. Neural teachers, high-dimensional observations, continuous actions and joint learning are untested.

7 CONCLUSION

A policy swap distorts exactly the targets whose value moves the teacher's mean path weight at histories that deployment reaches, which neither total nor per-target action information identifies. Trained predictors follow this criterion in controlled channels and carry the gap at every tested RockSample mixture. A known or identifiable model allows do relabeling; otherwise the remedy is history-only collection.

AI USE STATEMENT

In this work, we used Claude (Anthropic), a generative AI tool, to polish the language of the manuscript and to organize parts of its typesetting and layout. The authors reviewed all AI-assisted text and layout changes, checked all proofs by hand, and checked the reported numbers against the recorded artifacts. We take responsibility for the final content of this work.

ETHICS STATEMENT

The study uses simulated control environments and contains no human-subject, personal, or sensitive data. Its practical message is that privileged information can create misleading confidence when the data-generating policy is not modeled. Reporting includes the paired comparisons behind every main-text claim, including negative results.

REPRODUCIBILITY STATEMENT

Code, protocols and recorded outputs are provided in the supplementary material, and we will release them publicly upon publication. The submitted archive is a core export with path-configurable scoring code, protocols, input and model manifests, and the weights and histories needed to rescore Table 2 and the controlled-channel results. Appendix C.10 separates recorded-statistic replay, frozen-model rescoring and paired retraining; replay and full frozen-model scoring were executed in clean directories with explicit inputs, and software versions, commands, tolerances and input checks accompany the export. `r30_tables.py`, `r34_tables.py` and `r35_policy_breakdown.py` regenerate the tables of the sensitivity and hard-label studies byte for byte from recorded curves, selections and context-level scores. Their candidate weights are not packaged; `r30_fit_selection.py` and `r34_hard_label.py` retrain them on a CUDA GPU from inputs expanded by `prepare_inputs.py`. The teacher-mixing study of Appendix C.6 is trained by `r38_mid_mixture.py` under a protocol fixed before training and summarized by `r38_mid_mixture_tables.py`. `r38_ablation_table.py` and `r38_mixture_decomposition.py` render Table 3 and panel C of Table 4 from recorded summaries, `r38_recorded_numbers.py` recomputes further statistics from recorded outputs, `r38_late_selection.py` applies behavior-label validation to later checkpoints of the recorded candidate curves, `r38_target_action_info.py` computes the exact per-target action information and the certificate comparison of Appendix C.7, `r39_recorded_tolerance_grid.py` tabulates its recorded tolerance grid, and `r38_per_target_examples.py` evaluates the finite examples of Appendix A.12 in exact arithmetic. `r36_target_heterogeneity.py` and `r36_is_second_moment.py` recompute the within-history heterogeneity and the importance-weight moment. The post hoc studies of Appendices B.5, C.5, C.8 and C.9 are run by `r39_per_target_learning.py`, `r39_more_layouts.py`, `r39_tolerance.py` and `r39_closed_loop.py` under protocols frozen before their data or results; `r39_per_target_report.py`, `r39_more_layouts_tables.py` and the `render` and `tables` steps of the other two regenerate their tables and macros byte for byte from recorded outputs that contain no weights. Proofs appear in Appendix A. The appendix guide summarizes further studies that are not reported in full here, and the archive does not include the historical Original/Fresh RockSample models of Appendix E.3 or the CartPole, historical closed-loop, factor and certificate studies.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. URL https://arxiv.org/abs/2107.07511.

Andrea Baisero and Christopher Amato. Unbiased asymmetric reinforcement learning under partial observability. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 44–52. International Foundation for Autonomous Agents and Multiagent Systems, 2022. URL https://www.ifaamas.org/Proceedings/aamas2022/pdfs/p44.pdf.

Andrea Baisero, Brett Daley, and Christopher Amato. Asymmetric DQN for partially observable reinforcement learning. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial*

Intelligence, volume 180 of *Proceedings of Machine Learning Research*, pp. 107–117. PMLR, 2022. URL https://proceedings.mlr.press/v180/baisero22a.html.

Elias Bareinboim, Jin Tian, and Judea Pearl. Recovering from selection bias in causal and statistical inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, pp. 2410–2416, 2014. doi: 10.1609/aaai.v28i1.9074. URL https://ojs.aaai.org/index.php/AAAI/article/view/9074.

Yang Cai, Xiangyu Liu, Argyris Oikonomou, and Kaiqing Zhang. Provable partially observable reinforcement learning with privileged information. In *Advances in Neural Information Processing Systems*, volume 37, pp. 63790–63857, 2024. doi: 10.52202/079017-2038. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/74d188c51d97fcfbc0269f584d6a53b7-Abstract-Conference.html.

Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating. In *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pp. 66–75. PMLR, 2020. URL https://proceedings.mlr.press/v100/chen20a.html.

Xiaoyu Chen, Yao Mark Mu, Ping Luo, Shengbo Li, and Jianyu Chen. Flow-based recurrent belief state learning for POMDPs. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 3444–3468. PMLR, 2022. URL https://proceedings.mlr.press/v162/chen22q.html.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/947018640bf36a2bb609d3557a285329-Abstract.html.

Tanmay Gangwani, Joel Lehman, Qiang Liu, and Jian Peng. Learning belief representations for imitation learning in POMDPs. In *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 1061–1071. PMLR, 2020. URL https://proceedings.mlr.press/v115/gangwani20a.html.

Maxime Gasse, Damien Grasset, Guillaume Gaudron, and Pierre-Yves Oudeyer. Using confounded data in latent model-based reinforcement learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=nFWRuJXPkU.

Mao Hong, Zhengling Qi, and Yanxun Xu. Model-based reinforcement learning for confounded POMDPs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 18668–18710. PMLR, 2024. URL https://proceedings.mlr.press/v235/hong24d.html.

Anthony Hu, Gianluca Corrado, Nicolas Griffiths, Zachary Murez, Corina Gurau, Hudson Yeo, Alex Kendall, Roberto Cipolla, and Jamie Shotton. Model-based imitation learning for urban driving. In *Advances in Neural Information Processing Systems*, volume 35, pp. 20703–20716, 2022. doi: 10.52202/068431-1505. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/827cb489449ea216e4a257c47e407d18-Abstract-Conference.html.

Edward S. Hu, James Springer, Oleh Rybkin, and Dinesh Jayaraman. Privileged sensing scaffolds reinforcement learning. In *International Conference on Learning Representations*, pp. 2608–2631, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0b334c6419042e761051d82fc621679c-Abstract-Conference.html.

Maximilian Igl, Luisa Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. Deep variational reinforcement learning for POMDPs. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2117–2126. PMLR, 2018. URL https://proceedings.mlr.press/v80/igl18a.html.

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2Drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 37, pp. 819–844, 2024. doi: 10.52202/079017-0025.

Rico Jonschkowski, Divyam Rastogi, and Oliver Brock. Differentiable particle filters: End-to-end learning with algorithmic priors. In *Proceedings of Robotics: Science and Systems*, Pittsburgh, Pennsylvania, June 2018. doi: 10.15607/RSS.2018.XIV.001. URL https://www.roboticsproceedings.org/rss14/p01.html.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X. URL https://doi.org/10.1016/S0004-3702(98)00023-X.

Yujin Kim, Nathaniel Chin, Arnav Vasudev, and Sanjiban Choudhury. Distilling realizable students from unrealizable teachers. *arXiv preprint arXiv:2505.09546*, 2025. URL https://arxiv.org/abs/2505.09546.

Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. RMA: Rapid motor adaptation for legged robots. In *Proceedings of Robotics: Science and Systems*, July 2021. doi: 10.15607/RSS.2021.XVII.011. URL https://www.roboticsproceedings.org/rss17/p011.html.

Daniel Kumor, Junzhe Zhang, and Elias Bareinboim. Sequential causal imitation learning with unobserved confounders. In *Advances in Neural Information Processing Systems*, volume 34, pp. 14669–14680, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/7b670d553471ad0fd7491c75bad587ff-Abstract.html.

Gaspard Lambrechts, Damien Ernst, and Aditya Mahajan. A theoretical justification for asymmetric actor-critic algorithms. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 32375–32405. PMLR, 2025. URL https://proceedings.mlr.press/v267/lambrechts25a.html.

Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. *Science Robotics*, 5(47):eabc5986, 2020. doi: 10.1126/scirobotics.abc5986. URL https://doi.org/10.1126/scirobotics.abc5986.

Yueheng Li, Guangming Xie, and Zongqing Lu. Guided policy optimization under partial observability. In *International Conference on Learning Representations*, pp. 34041–34066, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/398753ddd2e10ce39d0c620e85fab922-Abstract-Conference.html.

Antonio Loquercio, Elia Kaufmann, René Ranftl, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. Learning high-speed flight in the wild. *Science Robotics*, 6(59):eabg5810, 2021. doi: 10.1126/scirobotics.abg5810. URL https://doi.org/10.1126/scirobotics.abg5810.

Fan-Ming Luo, Lei Yuan, and Yang Yu. Uncertainty-sensitive privileged learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 79634–79669. Curran Associates, Inc., 2025. doi: 10.52202/085713-2662. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/724711fccb09d4519cbbb6d245d3675d-Paper-Conference.pdf.

Nico Messikommer, Jiaxu Xing, Elie Aljalbout, and Davide Scaramuzza. Student-informed teacher training. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2412.09149.

Takahiro Miki, Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science Robotics*, 7(62):eabk2822, 2022. doi: 10.1126/scirobotics.abk2822. URL https://doi.org/10.1126/scirobotics.abk2822.

Steven Morad, Ryan Kortvelesy, Matteo Bettini, Stephan Liwicki, and Amanda Prorok. POPGym: Benchmarking partially observable reinforcement learning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=chDrutUTs0K.

Hai Nguyen, Brett Daley, Xinchao Song, Christopher Amato, and Robert Platt. Belief-grounded networks for accelerated robot learning under partial observability. In *Proceedings of the 2020 Conference on Robot Learning*, volume 155 of *Proceedings of Machine Learning Research*, pp. 1640–1653. PMLR, 2021. URL https://proceedings.mlr.press/v155/nguyen21a.html.

Tianwei Ni, Benjamin Eysenbach, and Ruslan Salakhutdinov. Recurrent model-free RL can be a strong baseline for many POMDPs. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16691–16723. PMLR, 2022. URL https://proceedings.mlr.press/v162/ni22a.html.

Pedro A. Ortega and Daniel A. Braun. A minimum relative entropy principle for learning and acting. *Journal of Artificial Intelligence Research*, 38:475–511, 2010. doi: 10.1613/jair.3062. URL https://doi.org/10.1613/jair.3062.

Pedro A. Ortega, Markus Kunesch, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Joel Veness, Jonas Buchli, Jonas Degrave, Bilal Piot, Julien Perolat, Tom Everitt, Corentin Tallec, Emilio Parisotto, Tom Erez, Yutian Chen, Scott Reed, Marcus Hutter, Nando de Freitas, and Shane Legg. Shaking the foundations: Delusions in sequence models for interaction and control. *arXiv preprint arXiv:2110.10819*, 2021. URL https://arxiv.org/abs/2110.10819.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2 edition, 2009. ISBN 9780521895606. doi: 10.1017/CBO9780511803161. URL https://doi.org/10.1017/CBO9780511803161.

Lerrel Pinto, Marcin Andrychowicz, Peter Welinder, Wojciech Zaremba, and Pieter Abbeel. Asymmetric actor critic for image-based robot learning. In *Proceedings of Robotics: Science and Systems*, Pittsburgh, Pennsylvania, June 2018. doi: 10.15607/RSS.2018.XIV.008. URL https://www.roboticsproceedings.org/rss14/p08.html.

Danilo J. Rezende, Ivo Danihelka, George Papamakarios, Nan Rosemary Ke, Ray Jiang, Theophane Weber, Karol Gregor, Hamza Merzic, Fabio Viola, Jane Wang, Jovana Mitrovic, Frederic Besse, Ioannis Antonoglou, and Lars Buesing. Causally correct partial models for reinforcement learning. *arXiv preprint arXiv:2002.02836*, 2020. URL https://arxiv.org/abs/2002.02836.

Stéphane Ross, Geoffrey Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635. PMLR, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Mark Rowland, Anna Harutyunyan, Hado van Hasselt, Diana Borsa, Tom Schaul, Remi Munos, and Will Dabney. Conditional importance sampling for off-policy learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 45–55. PMLR, 2020. URL https://proceedings.mlr.press/v108/rowland20b.html.

Andrei A. Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.06295.

Hao Shao, Letian Wang, Ruobing Chen, Steven L. Waslander, Hongsheng Li, and Yu Liu. ReasonNet: End-to-end driving with temporal and global reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13723–13733, 2023. doi: 10.1109/CVPR52729.2023.01319.

Chengchun Shi, Masatoshi Uehara, Jiawei Huang, and Nan Jiang. A minimax learning approach to off-policy evaluation in confounded partially observable markov decision processes. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20057–20094. PMLR, 2022. URL https://proceedings.mlr.press/v162/shi22f.html.

Richard D. Smallwood and Edward J. Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973. doi: 10.1287/opre.21.5.1071. URL https://doi.org/10.1287/opre.21.5.1071.

Trey Smith and Reid Simmons. Heuristic search value iteration for POMDPs. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp. 520–527, July 2004. URL https://publications.ri.cmu.edu/heuristic-search-value-iteration-for-pomdps.

Gokul Swamy, Sanjiban Choudhury, J. Andrew Bagnell, and Steven Wu. Sequence model imitation learning with unobserved contexts. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17665–17676, 2022. doi: 10.52202/068431-1284. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/708e58b0b99e3e62d42022b4564bad7a-Abstract.html.

Ruo Yu Tao, Kaicheng Guo, Cameron Allen, and George Konidaris. Benchmarking partial observability in reinforcement learning with a suite of memory-improvable domains. *Reinforcement Learning Journal*, 6:1412–1439, 2025. URL https://rlj.cs.umass.edu/2025/papers/Paper153.html.

Guy Tennenholtz, Assaf Hallak, Gal Dalal, Shie Mannor, Gal Chechik, and Uri Shalit. On covariate shift of latent confounders in imitation and reinforcement learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=w01vBAcewNX.

Risto Vuorio, Pim de Haan, Johann Brehmer, Hanno Ackermann, Daniel Dijkman, and Taco Cohen. Deconfounding imitation learning with variational inference. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=3FsVtsISHW.

Aaron Walsman, Muru Zhang, Sanjiban Choudhury, Ali Farhadi, and Dieter Fox. Impossibly good experts and how to follow them. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sciA_xgYofB.

Andrew Wang, Andrew C. Li, Toryn Q. Klassen, Rodrigo Toro Icarte, and Sheila A. McIlraith. Learning belief representations for partially observable deep RL. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35970–35988. PMLR, 2023. URL https://proceedings.mlr.press/v202/wang23p.html.

Andrew Warrington, Jonathan W. Lavington, Adam Scibior, Mark Schmidt, and Frank Wood. Robust asymmetric learning in POMDPs. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11013–11023. PMLR, 2021. URL https://proceedings.mlr.press/v139/warrington21a.html.

Luca Weihs, Unnat Jain, Iou-Jen Liu, Jordi Salvador, Svetlana Lazebnik, Aniruddha Kembhavi, and Alexander Schwing. Bridging the imitation gap by adaptive insubordination. In *Advances in Neural Information Processing Systems*, volume 34, pp. 19134–19146, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/9fc664916bce863561527f06a96f5ff3-Abstract.html.

Chuan Wen, Jierui Lin, Trevor Darrell, Dinesh Jayaraman, and Yang Gao. Fighting copycat agents in behavioral cloning from observation histories. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2564–2575, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/1b113258af3968aaf3969ca67e744ff8-Abstract.html.

Junzhe Zhang, Daniel Kumor, and Elias Bareinboim. Causal imitation learning with unobserved confounders. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12263–12274, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/8fdd149fcaa7058caccc9c4ad5b0d89a-Abstract.html.

GUIDE TO THE APPENDIX

Appendix A proves the posterior and risk results, including finite fitting errors (Appendix A.4) and the split of the teacher's action information (Appendix A.12). Appendix B constructs the controlled channels, reports their complete results and adds a paired learning test of per-target action information (Appendix B.5). Appendix C holds the RockSample learning evidence: the constrained-loss study, the common-measure decomposition with the exact mixture sweep, training fit, the sensitivity, hard-label and checkpoint-selection study, more training layouts (Appendix C.5), the prespecified intermediate-mixture study (Appendix C.6), the per-target action-information analysis (Appendix C.7) with its tolerance and misspecification follow-up (Appendix C.8), closed-loop control (Appendix C.9) and the reproduction paths. Appendix D fixes protocols and metrics (Table 27). Appendix E adds the exact synthetic grid, the binary residual-selection control, CartPole and a summary of the historical Original/Fresh RockSample study. Appendix F states the target-wise certificate and measures its coverage and cost.

Studies summarized but not reported in full. The following further studies are summarized here; their full protocols and results are not part of this submission. Several are negative or mixed. A decision certificate for the frozen historical GRUs, which adds a split-calibrated prediction-error radius (Angelopoulos & Bates, 2021) to the structural interval, lacks verified exchangeability and policy-shift coverage, retained only 0.032–0.135% of unresolved sampling opportunities and still retained 19 incorrect actions. Closed-loop control with the sixty historical GRUs is summarized with Appendix C.9, whose constrained-loss predictors reverse its full-mixing sign. A paired single-action intervention under common exact-do continuation gave a full-mixing gap of $+0.304$ $(0.047, 0.588)$, with Fresh-group and final-15-step intervals including zero, so the historical return gaps have mixed signs. In a one-step sampling-regret analysis of the historical GRUs, the random-policy contrast was negative at full mixing in both groups and both splits, whereas the other three policies had positive contrasts. Decision certificates on exact posteriors retained 17.29–19.48% of full-mixing sampling opportunities, against 8.09–10.65% for cumulative equality and 76.14–78.50% for an oracle, at 8.6–11.8 times the cost of full do filtering. Filters whose initial, sensing and teacher factors were estimated from the training logs kept the reversal, with $\lambda = 1$ target NLL gaps of 17.2–18.1 nats. The estimated factors entered only exact filters, and no GRU was retrained with them. A layout-by-policy cross-evaluation of the twenty historical $\lambda = 1$ GRUs kept both reversal signs on training layouts, gave observational-soft no behavior advantage on unseen layouts, and showed clamped do-soft recovering only 3.6–4.8% of exact do's gain over Public (Appendix E.3). Time-varying teacher ratios at amplitude 0.30 gave targets certified only by cumulative counts probability errors up to 0.269, with 0.8–3.1% of their mass above 0.01, while the factor rule stayed exact. Without a dependence budget, the product certificate's retained-coordinate error reached 0.193 under correlated initial qualities. After public clamping, historical do-soft improved on the Public baseline by 0.061 and 0.059 nats in the Original splits, with intervals including zero, and by 0.138 and 0.143 nats in the Fresh splits, whereas exact do scored 4.257–4.413 nats against Public's 6.423–6.508. Also omitted are the historical fixed-final, capacity, optimization, policy-wise, own-target-fit and sensing-model analyses and the certificate attribution, unknown-ratio and mechanism analyses.

A PROOFS

A.1 TRAJECTORY LAW AND FILTERING RECURSIONS

This subsection records the displays that Section 3 states in words. The teacher μ collected the training histories and could act on S_t; its trajectory law is

$$p^{\mu}(s_{1:T}, o_{1:T}, a_{1:T-1}) = p(s_1)G_1(o_1 \mid s_1) \prod_{t=1}^{T-1} \mu(a_t \mid s_t, h_t)T(s_{t+1} \mid s_t, a_t)G(o_{t+1} \mid s_{t+1}, a_t). \tag{12}$$

Action-dependent G covers sensing, and a deployable policy is history-only: $\pi(a_t \mid s_t, h_t) = \pi(a_t \mid h_t)$.

Actions have two roles: selection favors states consistent with the teacher's action, and control moves the next state and observation. Only the behavior recursion selects:

$$\widetilde{b}_t(s) \propto \mu(a_t \mid s, h_t) b_t^{\mathrm{obs}}(s), \tag{13}$$

$$b_{t+1}^{\mathrm{obs}}(s') \propto G(o_{t+1} \mid s', a_t) \sum_s T(s' \mid s, a_t) \widetilde{b}_t(s). \tag{14}$$

$$b_{t+1}^{\mathrm{do}}(s') \propto G(o_{t+1} \mid s', a_t) \sum_s T(s' \mid s, a_t) b_t^{\mathrm{do}}(s). \tag{15}$$

Both recursions keep the action's control effects in T and G; the separate ablation in Appendix A.8 removes the action input itself.

The do-soft objective of Section 5.1, with $q_{\phi,j}(H_t)$ the output of head j and $p_{t,j}^{\mathrm{do}}$ the do-posterior marginal of target j, is

$$\mathcal{L}_{\mathrm{do}}(\phi) = \mathbb{E}_{H_t \sim P^\mu} \left[\sum_{j=1}^{11} \mathrm{CE}\big(p_{t,j}^{\mathrm{do}}, q_{\phi,j}(H_t)\big) \right]; \tag{16}$$

observational-soft substitutes the behavior-posterior marginal $p_{t,j}^{\mu}$.

A.2 Continuous variables and support

Take the state, observation, and action spaces to be standard Borel spaces, so regular conditional laws exist. The factorization in equation 12 uses probability kernels. Dominated kernels can be written as densities. For continuous actions, μ, π, and $\bar{\mu}$ denote densities relative to one common σ-finite action reference measure. State and action sums in the filtering recursions become integrals. Deterministic transitions and point-mass posteriors are represented as kernels and measures rather than Lebesgue densities.

At each evaluated finite prefix, retain the stated target-history and posterior absolute-continuity conditions, positive finite normalization $0 < \mathbb{E}_{\mathrm{do}}[W_{t-1}^{\mu} \mid h] < \infty$, and finite log-risk expectations where used. The path weight is then a product of action densities. An equivalent action reference measure multiplies this product by a history-dependent factor that cancels in equation 5. The conditional cross-entropy, Radon–Nikodym, and trajectory-KL arguments below therefore apply to both discrete and continuous variables.

The finite-action mixture in equation 1 supplies the required support. For continuous actions, a sufficient coverage condition is domination of target action kernels by behavior kernels on the required state–history support. A positive exploration density with respect to Lebesgue measure does not dominate a deterministic target action. Such targets require appropriate atomic support or an additional identification argument. Density positivity also gives no uniform control of finite-sample importance-weight variance.

A.3 Direct supervision and the risk reversal

Lemma 1 (Population target of direct state supervision)**.** *For any predictor q with finite risk,*

$$\mathcal{R}_{\mu,t}(q) - \mathcal{R}_{\mu,t}(b^{\mu,Z}) = \mathbb{E}_{H_t \sim P^\mu} D_{\mathrm{KL}}\Big(b_t^{\mu,Z} \| q(\cdot \mid H_t)\Big). \tag{17}$$

Hence, if the predictor class contains $P^\mu(Z_t \mid H_t)$ on behavior support, the population minimizer of equation 4 is $b_t^{\mu,Z}$. Under squared error with finite second moment, the target is $\mathbb{E}_\mu[Z_t \mid H_t]$.

A fixed representation $R_t = f(H_t)$ instead selects $P^\mu(Z_t \mid R_t)$, which recovers the full-history conditional exactly when R_t is sufficient for Z_t under P^μ.

For any policy ν and predictor q, condition on $H_t = h$:

$$\mathbb{E}_\nu[-\log q(Z_t \mid h) \mid h] = H_\nu(Z_t \mid h) + D_{\mathrm{KL}}(P^\nu(Z_t \mid h) \| q(\cdot \mid h)). \tag{18}$$

Averaging equation 18 under $P^\mu(H_t)$ proves Lemma 1 and identifies $b_t^{\mu,Z}$ as the population minimizer. For squared error, write

$$\mathbb{E}[\|Z - q(H)\|^2 \mid H] = \mathrm{tr}\,\mathrm{Var}(Z \mid H) + \|\mathbb{E}[Z \mid H] - q(H)\|^2. \tag{19}$$

Each side of Proposition 1 is a difference of cross-entropies of the true conditional against another distribution on the same histories. It remains to identify the target conditional. Under a history-only policy, the target trajectory factorization replaces $\mu(a_k \mid s_k, h_k)$ in equation 12 by $\pi(a_k \mid h_k)$. Once $H_t = h_t$ is fixed, every factor $\pi(a_k \mid h_k)$ is constant with respect to the latent path. Those factors cancel from the normalized conditional over $S_{1:t}$, leaving exactly the truncated-factorization posterior. Therefore $P^\pi(Z_t \mid H_t) = b_t^{\mathrm{do},Z}$. Applying equation 18 first under $P^\mu(H_t)$ with $q = b_t^{\mathrm{do},Z}$ and then under $P^\pi(H_t)$ with $q = b_t^{\mu,Z}$ proves both equations in Proposition 1. Its target-risk identity is

$$\mathrm{PSPD}_t(\mu \to \pi) = \mathcal{R}_{\pi,t}(b^{\mu,Z}) - \mathcal{R}_{\pi,t}(b^{\mathrm{do},Z}). \tag{20}$$

A.4 Finite fitting errors and deployment ordering

Write $p_j(\cdot \mid h) = b_j^{\mathrm{do}}(\cdot \mid h)$ and $b_j(\cdot \mid h) = b_j^{\mu}(\cdot \mid h)$ for J target marginals. Let d be a finite measure over public histories, including the scoring time, with $d(\mathcal{H}) \le 1$, and let $u_j(h) \in [0,1]$ be a public-history-measurable target weight. All quantities below use this same measure and these same weights:

$$D = \int \sum_{j=1}^{J} u_j(h) D_{\mathrm{KL}}(p_j \| b_j)\, d(dh), \tag{21}$$

$$E_{\mathrm{obs}} = \int \sum_{j=1}^{J} u_j(h) D_{\mathrm{KL}}(b_j \| q_{\mathrm{obs},j})\, d(dh), \tag{22}$$

$$E_{\mathrm{do}} = \int \sum_{j=1}^{J} u_j(h) D_{\mathrm{KL}}(p_j \| q_{\mathrm{do},j})\, d(dh). \tag{23}$$

The do-reference expected log-score contrast is

$$G = \int \sum_{j} u_j(h) \mathbb{E}_{p_j}\left[\log \frac{q_{\mathrm{do},j}}{q_{\mathrm{obs},j}}\right] d(dh) \tag{24}$$

$$= D - E_{\mathrm{do}} + \int \sum_{j} u_j(h) \mathbb{E}_{p_j}\left[\log \frac{b_j}{q_{\mathrm{obs},j}}\right] d(dh). \tag{25}$$

For a history-only deployment law, p_j is the true conditional target. The three KL summaries leave a signed term undetermined. These are sums of marginal scores, not joint-state KL divergences.

Equal KL summaries can give opposite rankings. For one Bernoulli target, set $p = 3/5$, $b = q_{\mathrm{do}} = 1/2$, and compare $q_{\mathrm{obs}} = 11/20$ with $q_{\mathrm{obs}} = 9/20$. Both choices have exactly the same $D = E_{\mathrm{do}} = D_{\mathrm{KL}}(\mathrm{Ber}(3/5) \| \mathrm{Ber}(1/2))$ and $E_{\mathrm{obs}} = \frac{1}{2}\log(100/99)$. Yet their respective contrasts are

$$G_+ = \tfrac{1}{5}\log\frac{100000}{107811} < 0, \qquad G_- = \tfrac{1}{5}\log\frac{100000}{88209} > 0. \tag{26}$$

All target and predictor probabilities are strictly between zero and one. The two posteriors arise from a valid two-step model: a static state $Y \sim \mathrm{Ber}(3/5)$, constant observations, behavior action probabilities $\mu(A_1 = 1 \mid Y = 1) = 2/5$ and $\mu(A_1 = 1 \mid Y = 0) = 3/5$, and a history-only target policy that takes $A_1 = 1$. At the next prediction, the behavior posterior is $1/2$ and the do posterior is $3/5$. This is an analytic example, not a RockSample measurement. It concerns deployment ordering; behavior-side superiority would require a separate comparison under the behavior history law.

There is also no uniform guarantee from vanishing own-target errors alone. For $0 < \eta < 1/2$, take $p = \eta$, $b = e^{-1/\eta}$, $q_{\mathrm{obs}} = p$ and $q_{\mathrm{do}} = 2\eta$. As $\eta \downarrow 0$, $D \to 1$ and both own-target errors tend to zero, yet $G = -E_{\mathrm{do}} < 0$. This again has a positive-support realization: use prior p, behavior action-one probabilities cb/p and $c(1-b)/(1-p)$ for the two states, with $c = (1-p)/2$, and deploy action one. The posterior ratio $p/b = \eta e^{1/\eta}$ diverges, identifying the missing control in the following sufficient condition.

A sufficient condition with a bounded posterior ratio. Suppose all displayed expectations are finite, the conditional laws have common dominating measures, and, on the weighted history–target

support, $p_j \ll b_j$, $b_j \ll q_{\mathrm{obs},j}$, and $p_j \ll q_{\mathrm{do},j}$. Assume additionally

$$\frac{dp_j}{db_j}(z \mid h) \le M < \infty \quad \text{for every weighted coordinate, almost surely.} \tag{27}$$

Let $A = \int \sum_j u_j(h)\, d(dh) \le J$. For a single coordinate, using $\log x \le x-1$ only where $q_{\mathrm{obs},j} > b_j$ gives

$$\mathbb{E}_{p_j} \log \frac{b_j}{q_{\mathrm{obs},j}} \ge -M \int_{q_{\mathrm{obs},j} > b_j > 0} b_j \log \frac{q_{\mathrm{obs},j}}{b_j} \tag{28}$$

$$\ge -M \ \mathrm{TV}(b_j, q_{\mathrm{obs},j}) \tag{29}$$

$$\ge -M \sqrt{\tfrac{1}{2} D_{\mathrm{KL}}(b_j \| q_{\mathrm{obs},j})}. \tag{30}$$

Cauchy–Schwarz over the weighted coordinates and histories therefore yields

$$G \ge D - E_{\mathrm{do}} - M\sqrt{AE_{\mathrm{obs}}/2} \ge D - E_{\mathrm{do}} - M\sqrt{JE_{\mathrm{obs}}/2}. \tag{31}$$

This is a sufficient log-score bound obtained from Pinsker's inequality. A positive right-hand side ensures $G > 0$; a nonpositive value leaves its sign unresolved.

For the fixed-60 analysis, use $d(\{t\} \times dh) = 60^{-1} P^{\pi}(H_t \in dh, t < L)$ and $u_j = 1 - K_{tj}$ for undetermined targets. Then A is the expected active undetermined-coordinate mass, including the zero contributions after termination. Conditional per-coordinate scores divide $D, E_{\mathrm{obs}}, E_{\mathrm{do}}, G$ by this same $A > 0$; the resulting square-root term is $M\sqrt{(E_{\mathrm{obs}}/A)/2}$. The original $1/L$ weighting can depend on future outcomes and does not automatically identify these scores with a fixed-time population risk.

The errors in equation 31 are evaluated on deployment histories. Replacing them with population errors on a training-history measure requires an additional domination condition $d \le C d_{\mathrm{train}}$ with the same measurable weights, giving $G \ge D - CE_{\mathrm{do,train}} - M\sqrt{ACE_{\mathrm{obs,train}}/2}$. Empirical training or selected-validation errors do not establish that condition or the required population error bounds. The posterior ratio M and history ratio C are different assumptions; either can be large or unavailable. We do not infer a fitted-GRU guarantee or a two-sided behavior/deployment reversal from the three observed KL summaries.

A.5 Posterior path tilt

For a fixed h_t, the do and behavior path densities share the initial-state, transition, and observation factors. Their unnormalized density ratio is

$$W^{\mu}_{t-1}(s_{1:t-1}; h_t) = \prod_{k=1}^{t-1} \mu(a_k \mid s_k, h_k). \tag{32}$$

Normalizing the behavior path measure gives equation 5. Its S_t marginal is

$$b^{\mu}_t(s \mid h_t) = b^{\mathrm{do}}_t(s \mid h_t) \frac{\mathbb{E}_{\mathrm{do}}[W^{\mu}_{t-1} \mid S_t = s, h_t]}{\mathbb{E}_{\mathrm{do}}[W^{\mu}_{t-1} \mid h_t]}. \tag{33}$$

If $\mu(a_k \mid s_k, h_k)$ is independent of s_k for every k, the weight is constant on latent paths and the ratio is one.

A.6 Residual selection and sufficient histories

Averaging the path tilt equation 5 over the latent paths that share a target value leaves the ratio of the target-conditional to the unconditional mean path weight. Conditional expectation of the path density ratio given (Z_t, H_t) gives the Radon–Nikodym derivative in equation 7, with no Lebesgue-density assumption. For the finite-action mixture in equation 1, every action has probability at least $\epsilon/|\mathcal{A}|$. At each finite horizon, this ensures target-history absolute continuity and positive path weights on the relevant support. KL divergence vanishes exactly when its two probability measures agree. Since it is nonnegative, its target-history expectation vanishes exactly when they agree on almost every target history; under common support this is equivalent to the conditional weight ratio being one.

If $Z_t = f_t(H_t)$ under the do law, conditioning additionally on Z_t cannot change the conditional expectation of W^μ_{t-1}. The behavior posterior has the same point mass by absolute continuity. This proves Corollary 1, including deterministic targets for which a Lebesgue-density log risk need not exist.

This does not require the teacher-action information to be zero. For example, let $S_1 \sim \mathrm{Bernoulli}(1/2)$, O_1 be constant, and let a supported teacher choose $A_1 = S_1$ with probability $r \in (1/2, 1)$. Then $I(S_1; A_1 \mid O_1) = \log 2 - h(r) > 0$, where h is binary entropy in nats. If $S_2 = S_1$ and $O_2 = S_2$, both time-2 posteriors are the observed point mass under every supported history-only target policy, so $\mathrm{PSPD}_2 = 0$. Teacher information and residual posterior distortion can therefore differ even in a fully specified finite process.

Marginal cancellation without bit-flip invariance. A finite analytic example separates the necessary-and-sufficient criterion from the structural certificate. Let X, Y be independent fair bits, let $S_1 = S_2 = (X, Y)$, and let observations be constant. Actions have no effect on the state. For $\delta = 1/2$, choose

$$\mu(A_1 = 1 \mid x, y) = \frac{1 + \delta(2x - 1)(2y - 1)}{2}, \qquad \pi(A_1 = 1) = \frac{1}{2}. \tag{34}$$

Behavior action probabilities are $1/4$ or $3/4$, so support is shared. At either realized action and for either value of target $Z_2 = X$, $\mathbb{E}_{\mathrm{do}}[W^\mu_1 \mid X, h_2] = 1/2 = \mathbb{E}_{\mathrm{do}}[W^\mu_1 \mid h_2]$. Both target posteriors are therefore Bernoulli$(1/2)$. Nevertheless, flipping X at any fixed Y changes the action likelihood, so the pointwise bit-flip test rejects every pair. Here W^μ_1 depends on both state components and X is not independent of (X, Y), so the component condition after Corollary 1 fails although the ratio is one; it would hold for the path function $X \oplus Y$, which is independent of X. Cancellation across the unobserved Y values preserves the marginal even though the action selects their joint configuration. This analytic example does not estimate how often such cancellation occurs in RockSample; Appendix F.3 measures the certificate's coverage numerically.

Position-Only CartPole. For the deterministic Euler equations with step $\tau = 0.02$,

$$\dot{x}_{t-1} = \frac{x_t - x_{t-1}}{\tau}, \qquad \dot{\theta}_{t-1} = \frac{\theta_t - \theta_{t-1}}{\tau}. \tag{35}$$

Both quantities use only two public observations. Together with A_{t-1}, the recovered state determines the accelerations and hence $(\dot{x}_t, \dot{\theta}_t)$ through the known transition equations. Consequently, from $t = 2$ onward, the velocity target is a function of the full history in the ideal real-valued environment. At $t = 1$ no past action factor is present. Both posterior laws thus agree at every scoring time. Float32 observation quantization introduces residual uncertainty in the implemented environment; the reconstruction diagnostic measures velocity error and does not bound the corresponding KL divergence.

A.7 Action-channel KL and TAL

Proposition 2 (Action-channel decomposition)**.** *Let $\bar{\mu}(a \mid h) = P^\mu(A_t = a \mid H_t = h)$, and let $\mathbb{P}^\mu_K$ denote the full latent-augmented trajectory law of $(S_{1:K+1}, O_{1:K+1}, A_{1:K})$. Let $\mathbb{P}^{\bar{\mu}}_K$ replace only its action factors by $\bar{\mu}$. For common support,*

$$D_{\mathrm{KL}}(\mathbb{P}^\mu_K \| \mathbb{P}^{\bar{\mu}}_K) = \sum_{t=1}^{K} I_\mu(S_t; A_t \mid H_t) = \mathrm{TAL}_K(\mu). \tag{36}$$

The full latent-augmented laws $\mathbb{P}^\mu_K$ and $\mathbb{P}^{\bar{\mu}}_K$ in Proposition 2 share every environment factor. Their log-density ratio is

$$\log \frac{d\mathbb{P}^\mu_K}{d\mathbb{P}^{\bar{\mu}}_K} = \sum_{t=1}^{K} \log \frac{\mu(A_t \mid S_t, H_t)}{\bar{\mu}(A_t \mid H_t)}. \tag{37}$$

Taking expectation under $\mathbb{P}^\mu_K$ and conditioning each summand on H_t yields $I_\mu(S_t; A_t \mid H_t)$. Summing proves equation 36.

A.8 EXACT COST OF DISCARDING THE CONTROL

The recursive no-action ablation marginalizes each realized control:

$$b_{t+1}^{\mathrm{noact}}(s') \propto \sum_a \pi(a \mid h_t) G(o_{t+1} \mid s', a) \sum_s T(s' \mid s, a) b_t^{\mathrm{noact}}(s). \tag{38}$$

Its weights still use h_t, so it is generally neither the full-history nor the observation-only posterior. Together with equation 14 and equation 15, this recursion defines the matched three-way filter comparison.

Starting from the same target belief as the do-update, the one-step omitted action posterior is

$$b_{t+1}^{\mathrm{omit}}(s') \propto \sum_a \pi(a \mid h_t) G(o_{t+1} \mid s', a) \sum_s T(s' \mid s, a) b_t^{\mathrm{do}}(s). \tag{39}$$

Under P^π, set $C_t = (H_t, O_{t+1})$. The entropy form of the one-step Bayes log risks is

$$\mathbb{E}_\pi[-\log b_{t+1}^{\mathrm{omit},Z}] = H_\pi(Z_{t+1} \mid C_t), \tag{40}$$

$$\mathbb{E}_\pi[-\log b_{t+1}^{\mathrm{do},Z}] = H_\pi(Z_{t+1} \mid C_t, A_t). \tag{41}$$

Their difference is $I_\pi(Z_{t+1}; A_t \mid C_t)$, which proves

$$\mathbb{E}_\pi[-\log b_{t+1}^{\mathrm{omit},Z}] - \mathbb{E}_\pi[-\log b_{t+1}^{\mathrm{do},Z}] = I_\pi(Z_{t+1}; A_t \mid H_t, O_{t+1}) \geq 0. \tag{42}$$

Equation 39 starts from b_t^{do} and omits only A_t. The recursive filter in equation 38 instead feeds its previous no-action belief into the next step, so it is the cumulative action-removal ablation used in the experiments, not the oracle in this identity. Thus equation 42 gives the one-step Bayes cost of discarding control, not the recursive filter's cumulative penalty.

A.9 PASSIVE IDENTIFICATION

Proposition 3 (Passive identification of the do-filter)**.** *Assume: (i) consistency of logged and intervened transitions; (ii) policy replacement leaves $p(s_1)$, T, G_1, and G invariant; (iii) target state–action occupancy is covered by the behavior data; and (iv) within the stated model class, state-labeled trajectories make $p(s_1)$, $G_1(o \mid s)$, $T(s' \mid s, a)$, and $G(o' \mid s', a)$ recoverable on the required support. Then the target-support do-filter in equation 15 is identified from passive privileged logs. The factor $\mu(a \mid s, h)$ is not required by its recursion.*

The state-labeled behavior law directly identifies $p(s_1)$ and $G_1(o \mid s)$. By consistency and state–action coverage, it also identifies

$$T(s' \mid s, a) = P^\mu(S_{t+1} = s' \mid S_t = s, A_t = a), \tag{43}$$

$$G(o' \mid s', a) = P^\mu(O_{t+1} = o' \mid S_{t+1} = s', A_t = a) \tag{44}$$

on every pair required by the target occupancy. Environment modularity makes these conditionals invariant to policy replacement. The identified initial law and G_1 establish the base belief. If b_t^{do} is identified, substitution of T, the realized action, and G into equation 15 identifies b_{t+1}^{do}. Induction proves Proposition 3. Its content is modest: once the factors are recovered, the do recursion follows by substitution, whereas the behavior recursion additionally needs the propensity. Finite coverage and labeled-log diagnostics are treated in Appendices D.1 and A.10; the factor-estimation study is listed in the guide to the appendix.

A.10 DIAGNOSTICS FROM LABELED BEHAVIOR LOGS

For a specified history-only target policy, identified environment factors permit target-history simulation. Comparing the do and observational posteriors on those histories also requires the behavior propensity, either known or recovered from labeled logs. Target risk can alternatively be estimated by weighting held-out labeled behavior prefixes. For fixed predictors, define

$$L_t = \log \frac{q_{\mathrm{do}}(Z_t \mid H_t)}{q_{\mathrm{obs}}(Z_t \mid H_t)}, \qquad w_{t-1} = \prod_{k=1}^{t-1} \frac{\pi(A_k \mid H_k)}{\mu(A_k \mid S_k, H_k)}. \tag{45}$$

Assume that the target latent-augmented prefix law is absolutely continuous with respect to the behavior law and that the weighted loss is integrable. The shared environment factors cancel from their density ratio, yielding

$$\mathbb{E}_\mu[w_{t-1}L_t] = \mathbb{E}_\pi[L_t]. \tag{46}$$

With exact posteriors this is $\mathrm{PSPD}_t(\mu \to \pi)$. With fitted predictors it is the prediction-risk contrast in Section 4. Applying the identity requires known propensities or their estimation from the logs. It also requires evaluating $\pi(A_k \mid H_k)$ on the logged histories. This needs no kernels when π is a given function of the history, but the three history-driven RockSample deployment policies act on exact do marginals, so evaluating them runs the do filter and needs the kernels. Small behavior propensities can produce high-variance weights, and estimated kernels or propensities introduce additional error. Neither the population identity nor TAL supplies a small-sample certificate of residual posterior distortion.

With only observation–action logs, identifying the same do-posterior requires additional structure relating those logs to the latent target Z_t, such as an identifiable latent model. These assumptions go beyond Proposition 3. Agreement of fitted posteriors across history-only policies alone does not identify PSPD: it neither recovers the behavior posterior nor specifies the target-history weighting. Conversely, disagreement can arise from finite representation or estimation error even when the exact do-posterior is policy invariant.

A.11 Non-identification without coverage

Let $S \sim \mathrm{Bernoulli}(1/2)$, let the public observation be constant, and let the behavior policy choose $A = S$. Environment E_1 sets $S' = S$ for both actions. Environment E_2 sets

$$S' = \begin{cases} S, & A = S, \\ 1 - S, & A \neq S. \end{cases} \tag{47}$$

Behavior data visit only $(S, A) = (0, 0)$ and $(1, 1)$, on which both environments produce $S' = S$. Their complete passive distributions over (S, A, S') are identical. The deployable fixed policy $A = 0$ yields

$$P_{E_1}(S' = 1 \mid \mathrm{do}(A = 0)) = \tfrac{1}{2}, \qquad P_{E_2}(S' = 1 \mid \mathrm{do}(A = 0)) = 0. \tag{48}$$

The target interventional law is therefore not identified without the missing state–action support.

A.12 Where the teacher's action information goes

Proposition 2 writes the teacher's action information as a trajectory KL divergence, and Corollary 1 decides whether a target marginal moves. Combining the two with the chain rule for KL divergence splits the action information accumulated before a scoring time into three nonnegative parts, and only one of them concerns the target marginal. The split is a corollary of these two results. It places TAL relative to the residual-selection criterion and relates the criterion to per-target measures of action information.

Notation. Fix a scoring time $t \geq 2$ and a target $Z_t = g(S_t)$. Take $\bar{\mu}$ and $\mathbb{P}^{\bar{\mu}}_{t-1}$ from Proposition 2 with $K = t - 1$. Both $\mathbb{P}^{\mu}_{t-1}$ and $\mathbb{P}^{\bar{\mu}}_{t-1}$ are laws of $(S_{1:t}, H_t)$. With densities relative to the common reference measures of Appendix A.2, write the environment factor of equation 12 as

$$e_t(s_{1:t}, h_t) = p(s_1)G_1(o_1 \mid s_1) \prod_{k=1}^{t-1} T(s_{k+1} \mid s_k, a_k)G(o_{k+1} \mid s_{k+1}, a_k), \tag{49}$$

so that the two joint densities are $e_t W^{\mu}_{t-1}$ and $e_t \bar{W}_{t-1}$, with W^{μ}_{t-1} the path weight of equation 5 and $\bar{W}_{t-1}(h_t) = \prod_{k=1}^{t-1} \bar{\mu}(a_k \mid h_k)$. The do path posterior $P^{\mathrm{do}}(ds_{1:t} \mid h_t)$ is proportional to e_t, and $P^{\mathrm{do}}(ds_{1:t} \mid Z_t = z, h_t)$ is its conditional given the target value. Let $r_t(z, h)$ denote the ratio in equation 7. The three parts are

$$\mathcal{I}^{\mathrm{hist}}_t = D_{\mathrm{KL}}\big(P^{\mu}(H_t \in \cdot) \,\|\, P^{\bar{\mu}}(H_t \in \cdot)\big), \tag{50}$$

$$\mathcal{I}^{\mathrm{off}}_t = \mathbb{E}_{(Z_t, H_t) \sim P^{\mu}} D_{\mathrm{KL}}\big(P^{\mu}(S_{1:t} \in \cdot \mid Z_t, H_t) \,\|\, P^{\mathrm{do}}(S_{1:t} \in \cdot \mid Z_t, H_t)\big), \tag{51}$$

$$\mathcal{I}^{\mathrm{on}}_t = \mathbb{E}_{H_t \sim P^{\mu}} D_{\mathrm{KL}}\big(b^{\mu,Z}_t(\cdot \mid H_t) \,\|\, b^{\mathrm{do},Z}_t(\cdot \mid H_t)\big). \tag{52}$$

The on-target residual $\mathcal{I}_t^{\mathrm{on}}$ is the behavior advantage equation 8. Let $D_t = \{h : D_{\mathrm{KL}}(b_t^{\mu,Z}(\cdot \mid h) \| b_t^{\mathrm{do},Z}(\cdot \mid h)) > 0\}$ be the histories at which the two target marginals differ. For comparison, three per-target measures of action information are

$$\iota_t^{\mathrm{act}} = \sum_{k=1}^{t-1} I_\mu(g(S_k); A_k \mid H_k), \qquad \iota_t^{\mathrm{fwd}} = \sum_{k=1}^{t-1} I_\mu(Z_t; A_k \mid H_k),$$
$$\iota_t^{\mathrm{obs}} = I_\mu(Z_t; A_{1:t-1} \mid O_{1:t}). \tag{53}$$

Here ι_t^{act} measures information about the target component when the teacher acts, ι_t^{fwd} information about the scored target given the history at each action, and ι_t^{obs} information about the scored target given all observations. Finally, $\chi^2(P\|Q) = \int (dP/dQ - 1)^2\, dQ$ for $P \ll Q$.

Corollary 2 (Action-information split)**.** *Assume the conditions of Proposition 2 with $K = t-1$ and $\mathrm{TAL}_{t-1}(\mu) < \infty$, those of Corollary 1 at time t, and $P^\pi(H_t) \ll P^\mu(H_t)$. Then:*

(i) Split. *The three parts are nonnegative and finite, and*

$$\mathrm{TAL}_{t-1}(\mu) = \mathcal{I}_t^{\mathrm{hist}} + \mathcal{I}_t^{\mathrm{off}} + \mathcal{I}_t^{\mathrm{on}}. \tag{54}$$

The first part is carried by later observations:

$$\mathcal{I}_t^{\mathrm{hist}} = \sum_{k=1}^{t-1} \mathbb{E}_{P^\mu} D_{\mathrm{KL}}\big(P^\mu(O_{k+1} \in \cdot \mid H_k, A_k) \,\|\, P^{\mathrm{do}}(O_{k+1} \in \cdot \mid H_k, A_k)\big), \tag{55}$$

where $P^{\mathrm{do}}(O_{k+1} \in \cdot \mid h_k, a_k) = \iint G(\cdot \mid s', a_k) T(ds' \mid s, a_k)\, b_k^{\mathrm{do}}(ds \mid h_k)$ is the one-step observation predictive of the do filter, and the behavior predictive applies the same kernels to the selection-updated belief of equation 13.

(ii) Residual and distortion. *$\mathcal{I}_t^{\mathrm{on}} > 0$ exactly when $P^\mu(H_t \in D_t) > 0$, and $\mathrm{PSPD}_t > 0$ exactly when $P^\pi(H_t \in D_t) > 0$. Hence $\mathrm{PSPD}_t > 0$ implies $0 < \mathcal{I}_t^{\mathrm{on}} \le \mathrm{TAL}_{t-1}(\mu)$, while $\mathcal{I}_t^{\mathrm{on}} > 0$ leaves $\mathrm{PSPD}_t = 0$ whenever $P^\pi(H_t \in D_t) = 0$.*

(iii) Cumulative information without control. *If T and G do not depend on the action, then $\iota_t^{\mathrm{obs}} = \mathcal{I}_t^{\mathrm{on}}$.*

(iv) Ratio variance. *At every history h,*

$$\mathrm{Var}_{Z \sim b_t^{\mathrm{do},Z}(\cdot|h)}\big[r_t(Z,h)\big] = \chi^2\big(b_t^{\mu,Z}(\cdot \mid h) \,\|\, b_t^{\mathrm{do},Z}(\cdot \mid h)\big), \tag{56}$$

and its average under $P^\pi(H_t)$ is zero exactly when $\mathrm{PSPD}_t = 0$. If $0 < m \le r_t(\cdot, h) \le M$ on the support of $b_t^{\mathrm{do},Z}(\cdot \mid h)$, then

$$\frac{\chi^2(b_t^{\mu,Z} \| b_t^{\mathrm{do},Z})}{2M^2} \le D_{\mathrm{KL}}\big(b_t^{\mathrm{do},Z} \| b_t^{\mu,Z}\big) \le \frac{\chi^2(b_t^{\mu,Z} \| b_t^{\mathrm{do},Z})}{2m^2} \quad \text{at } h. \tag{57}$$

Proof. All statements below hold P^μ-almost surely. The weights W_{t-1}^μ and $\bar{W}_{t-1}$, the do likelihood $\int e_t(s_{1:t}, H_t)\, ds_{1:t}$, the normalizer $\mathbb{E}_{\mathrm{do}}[W_{t-1}^\mu \mid H_t]$ and the conditional mean $\mathbb{E}_{\mathrm{do}}[W_{t-1}^\mu \mid Z_t, H_t] = r_t(Z_t, H_t)\, \mathbb{E}_{\mathrm{do}}[W_{t-1}^\mu \mid H_t]$ are positive and finite, the last because the density $r_t = db_t^{\mu,Z}/db_t^{\mathrm{do},Z}$ is positive and finite $b_t^{\mu,Z}$-almost surely.

Three density ratios. (a) Integrating the joint densities over $s_{1:t}$ gives $p^\mu(h) = \mathbb{E}_{\mathrm{do}}[W_{t-1}^\mu \mid h] \int e_t(s,h)\, ds$ and $p^{\bar{\mu}}(h) = \bar{W}_{t-1}(h) \int e_t(s,h)\, ds$, so

$$\frac{dP^\mu(H_t \in \cdot)}{dP^{\bar{\mu}}(H_t \in \cdot)}(h) = \frac{\mathbb{E}_{\mathrm{do}}[W_{t-1}^\mu \mid h]}{\bar{W}_{t-1}(h)}. \tag{58}$$

(b) By Corollary 1, $r_t = db_t^{\mu,Z}/db_t^{\mathrm{do},Z}$. (c) For measurable sets B of paths and C of target values, equation 5 and (b) give

$$P^\mu(S_{1:t} \in B, Z_t \in C \mid h) = \frac{\mathbb{E}_{\mathrm{do}}[W_{t-1}^\mu \mathbf{1}_B \mathbf{1}_C(Z_t) \mid h]}{\mathbb{E}_{\mathrm{do}}[W_{t-1}^\mu \mid h]} = \int_C \frac{\mathbb{E}_{\mathrm{do}}[W_{t-1}^\mu \mathbf{1}_B \mid z, h]}{\mathbb{E}_{\mathrm{do}}[W_{t-1}^\mu \mid h]}\, b_t^{\mathrm{do},Z}(dz \mid h)$$
$$= \int_C \frac{\mathbb{E}_{\mathrm{do}}[W_{t-1}^\mu \mathbf{1}_B \mid z, h]}{\mathbb{E}_{\mathrm{do}}[W_{t-1}^\mu \mid z, h]}\, b_t^{\mu,Z}(dz \mid h).$$

Hence, for $b_t^{\mu,Z}$-almost every z,

$$\frac{dP^{\mu}(S_{1:t} \in \cdot \mid z, h)}{dP^{\mathrm{do}}(S_{1:t} \in \cdot \mid z, h)} = \frac{W_{t-1}^{\mu}}{\mathbb{E}_{\mathrm{do}}[W_{t-1}^{\mu} \mid z, h]}. \tag{59}$$

Telescoping. The environment factor cancels from $d\mathbb{P}_{t-1}^{\mu}/d\mathbb{P}_{t-1}^{\bar{\mu}} = W_{t-1}^{\mu}/\bar{W}_{t-1}$, and

$$\log \frac{W_{t-1}^{\mu}}{\bar{W}_{t-1}} = \log \frac{\mathbb{E}_{\mathrm{do}}[W_{t-1}^{\mu} \mid H_t]}{\bar{W}_{t-1}} + \log r_t(Z_t, H_t) + \log \frac{W_{t-1}^{\mu}}{\mathbb{E}_{\mathrm{do}}[W_{t-1}^{\mu} \mid Z_t, H_t]}. \tag{60}$$

The left side has P^{μ}-expectation $\mathrm{TAL}_{t-1}(\mu)$ by Proposition 2. By equation 58, the first term on the right has expectation $\mathcal{I}_t^{\mathrm{hist}}$. By (b) and the tower property over H_t, the second has expectation $\mathcal{I}_t^{\mathrm{on}}$. By equation 59 and the tower property over (Z_t, H_t), the third has expectation $\mathcal{I}_t^{\mathrm{off}}$. Each term is the logarithm of a (conditional) density ratio integrated against its numerator measure, so its negative part has expectation at most $\sup_{0<x<1} x \log(1/x) = e^{-1}$, and its expectation is a (conditional) KL divergence in $[0, \infty]$. Expectation therefore distributes over the sum, and the three nonnegative expectations add up to the finite $\mathrm{TAL}_{t-1}(\mu)$, which proves equation 54 and finiteness.

Observation form. Order H_t as $(O_1, A_1, O_2, \ldots, A_{t-1}, O_t)$ and apply the chain rule for KL divergence to equation 50. Both laws give O_1 the density $\int p(s_1)G_1(o_1 \mid s_1)\, ds_1$. Both give A_k the conditional law $\bar{\mu}(\cdot \mid H_k)$, under P^{μ} by the definition of $\bar{\mu}$. Under $P^{\bar{\mu}}$ the action is independent of S_k given H_k, and $P^{\bar{\mu}}(S_k \in \cdot \mid H_k) = b_k^{\mathrm{do}}$ because $\bar{\mu}$ is history-only (Section 3), so O_{k+1} has the do predictive. Under P^{μ}, S_k given (H_k, A_k) follows the selection-updated belief of equation 13. Only the observation terms remain, which is equation 55.

Part (ii). The integrand of $\mathcal{I}_t^{\mathrm{on}}$, $D_{\mathrm{KL}}(b_t^{\mu,Z} \| b_t^{\mathrm{do},Z})(h)$, and that of PSPD_t, $D_{\mathrm{KL}}(b_t^{\mathrm{do},Z} \| b_t^{\mu,Z})(h)$, are nonnegative and vanish exactly when the two marginals coincide, that is, off D_t. A nonnegative function has a positive integral exactly when it is positive on a set of positive measure, which gives both equivalences. Since $P^{\pi}(H_t) \ll P^{\mu}(H_t)$, $P^{\pi}(H_t \in D_t) > 0$ implies $P^{\mu}(H_t \in D_t) > 0$, and $\mathcal{I}_t^{\mathrm{on}} \le \mathrm{TAL}_{t-1}(\mu)$ follows from equation 54.

Part (iii). When T and G do not depend on the action, e_t does not depend on $a_{1:t-1}$. Integrating the actions out of $e_t W_{t-1}^{\mu}$ in the order $a_{t-1}, \ldots, a_1$ removes one normalized factor $\mu(a_k \mid s_k, h_k)$ at a time, so the law of $(S_{1:t}, O_{1:t})$ under P^{μ} has density e_t. Hence $P^{\mu}(Z_t \in \cdot \mid O_{1:t})$ is the target marginal of the posterior proportional to e_t, which is $b_t^{\mathrm{do},Z}(\cdot \mid H_t)$. Writing conditional mutual information as an expected KL divergence, $\iota_t^{\mathrm{obs}} = \mathbb{E}_{P^{\mu}} D_{\mathrm{KL}}\big(P^{\mu}(Z_t \in \cdot \mid H_t) \| P^{\mu}(Z_t \in \cdot \mid O_{1:t})\big) = \mathcal{I}_t^{\mathrm{on}}$.

Part (iv). By (b), $r_t(\cdot, h)$ is the density of $b_t^{\mu,Z}$ with respect to $b_t^{\mathrm{do},Z}$, so its $b_t^{\mathrm{do},Z}$-mean is one and its variance is $\mathbb{E}_{b_t^{\mathrm{do},Z}}[(r_t - 1)^2] = \chi^2(b_t^{\mu,Z} \| b_t^{\mathrm{do},Z})$. The variance vanishes exactly when $r_t = 1$ almost surely, that is, off D_t, so part (ii) gives the zero set. For the bounds, the zero mean of $r_t - 1$ gives $D_{\mathrm{KL}}(b_t^{\mathrm{do},Z} \| b_t^{\mu,Z}) = \mathbb{E}_{b_t^{\mathrm{do},Z}}[-\log r_t] = \mathbb{E}_{b_t^{\mathrm{do},Z}}[\phi(r_t)]$ with $\phi(x) = x - 1 - \log x$. Since $\phi(1) = \phi'(1) = 0$ and $\phi''(x) = x^{-2}$, Taylor's theorem gives $\phi(x) = (x-1)^2/(2\xi^2)$ for some ξ between 1 and x. A mean of one forces $m \le 1 \le M$, so $\xi \in [m, M]$, and taking expectations of $(r_t - 1)^2/(2M^2) \le \phi(r_t) \le (r_t - 1)^2/(2m^2)$ proves equation 57. □

Reading the split. Only the on-target residual measures how far the target marginal moves on teacher histories. The parts of the action information carried into later observations or spent on other path components do not enter it, however large they are. The residual is necessary but not sufficient for distortion after the swap: it averages over teacher histories, whereas PSPD_t counts only the distorted histories that the deployment policy reaches. The split bounds the residual by TAL_{t-1} but does not bound PSPD_t, which is a reverse divergence under another history law. Part (iii) identifies the cumulative per-target information given observations with the residual when actions do not control the environment, so it inherits the same blind spot for deployment support. Part (iv) applies the standard identity between the variance of a likelihood ratio and the χ^2 divergence: the variance of the Corollary 1 ratio is a χ^2 form of the same criterion, with the same zero set and bounds that compare its size with PSPD_t.

Exact examples. Table 5 evaluates the split and the three measures of equation 53 on 10 finite examples. Each has one teacher action, a constant O_1 and the scoring time $t = 2$, so $H_2 =$

(O_1, A_1, O_2); a teacher that copies a bit does so with probability 0.9, and deployment takes each action with probability $1/2$ unless stated otherwise.

- *Coupled bits*, the example of Appendix E.1 and Figure 1a: $S_1 = (X, Y)$ with $P(X = Y) = (1 + \rho)/2$ and uniform marginals, the teacher copies X, $S_2 = S_1$, O_2 is constant and $Z_2 = Y$, for $\rho \in \{0, 1/2, 1\}$. The rows $\rho = 0$ and $\rho = 1$ are the independent and copied bits of Figure 1a.
- *Cancellation*: the example equation 34 with $\delta = 0.5$ and $Z_2 = X$.
- *Revealed state*: the example of Appendix A.6, with a fair bit S_1 copied by the teacher, $S_2 = S_1$, $O_2 = S_2$ and $Z_2 = S_2$. *Noisy reveal* replaces O_2 by a copy of S_2 that is correct with probability 0.8.
- *Explaining away*: independent fair bits X, Y, a teacher that copies $X \oplus Y$, $S_2 = S_1$, $O_2 = Y_2$ and $Z_2 = X_2$.
- *Transport*: independent fair bits X, Y, a teacher that copies X, $S_2 = (Y_1, X_1)$, constant O_2 and $Z_2 = Y_2 = X_1$.
- *Support*: a fair bit S_1, three actions with $\mu(\cdot \mid s) = \left(\frac{1}{3}, \frac{1}{3} + c(2s - 1), \frac{1}{3} - c(2s - 1)\right)$ for $c = 0.2$, $S_2 = S_1$, constant O_2, $Z_2 = S_2$, and a deployment policy that always takes the first action, whose teacher probability $1/3$ does not depend on the state.
- *Control*: $P(S_1 = 1) = 0.9$, a history-only teacher ($\lambda = 0$) with uniform actions, $S_2 = S_1 \oplus A_1$, constant O_2 and $Z_2 = S_2$.

Probabilities are exact rationals and logarithms use 60-digit decimal arithmetic. The split equation 54, its observation form equation 55, the trajectory KL of Proposition 2, both identities of Proposition 1, part (iii) where it applies, the bounds equation 57 and the available closed forms hold to within 10^{-55}. Zeros and the distortion flag are decided in rational arithmetic, and the Figure 1a rows reproduce $\mathrm{TAL}_1 = 0.368$ nats with PSPD_2 equal to 0 and 0.511 nats (`r38_per_target_examples.py`).

Table 5: Exact action-information split on ten finite examples, in nats. Each example has one teacher action and scores $Z_2 = g(S_2)$ at time 2. In every row $\mathrm{TAL}_1 = \mathcal{I}_2^{\mathrm{hist}} + \mathcal{I}_2^{\mathrm{off}} + \mathcal{I}_2^{\mathrm{on}}$ (Corollary 2). Per-target action information: $\iota^{\mathrm{act}} = I_\mu(g(S_1); A_1 \mid H_1)$, $\iota^{\mathrm{fwd}} = I_\mu(Z_2; A_1 \mid H_1)$ and $\iota^{\mathrm{obs}} = I_\mu(Z_2; A_1 \mid O_{1:2})$. χ^2 is the deployment mean of the variance of the Corollary 1 ratio under the do posterior. Dist. records whether $b_2^{\mu,Z} \neq b_2^{\mathrm{do},Z}$ at some deployment history. Zeros and the distortion flag are decided in exact rational arithmetic; Appendix A.12 defines the examples.

		Per-target			Split of TAL_1			Deployment		
Example	TAL_1	ι^{act}	ι^{fwd}	ι^{obs}	$\mathcal{I}_2^{\mathrm{hist}}$	$\mathcal{I}_2^{\mathrm{off}}$	$\mathcal{I}_2^{\mathrm{on}}$	PSPD_2	χ^2	Dist.
Independent bit, $\rho = 0$	0.368	0	0	0	0	0.368	0	0	0	no
Partial coupling, $\rho = 1/2$	0.368	0.082	0.082	0.082	0	0.286	0.082	0.087	0.160	yes
Copied bit, $\rho = 1$	0.368	0.368	0.368	0.368	0	0	0.368	0.511	0.640	yes
Cancellation	0.131	0	0	0	0	0.131	0	0	0	no
Revealed state	0.368	0.368	0.368	0	0.368	0	0	0	0	no
Noisy reveal	0.368	0.368	0.368	0.248	0.120	0	0.248	0.380	0.851	yes
Explaining away	0.368	0	0	0.368	0	0	0.368	0.511	0.640	yes
Transport	0.368	0	0.368	0.368	0	0	0.368	0.511	0.640	yes
Support	0.128	0.128	0.128	0.128	0	0	0.128	0	0	no
Control	0	0	0.368	0.368	0	0	0	0	0	no

The rows separate the measures. In the independent-bit and cancellation rows all of TAL_1 is off-target and no target is distorted. In the revealed-state row O_2 carries all 0.368 nats, so $\mathrm{PSPD}_2 = 0$ although $\iota^{\mathrm{act}} = 0.368$ and $\iota^{\mathrm{fwd}} = 0.368$. With a noisy reveal, O_2 carries 0.120 nats and the residual of 0.248 nats yields $\mathrm{PSPD}_2 = 0.380$. Explaining away has $\iota^{\mathrm{act}} = 0$, $\iota^{\mathrm{fwd}} = 0$ and $\mathrm{PSPD}_2 = 0.511$ nats, because observing Y turns the action's evidence about $X \oplus Y$ into evidence about X. Transport has $\iota^{\mathrm{act}} = 0$ and $\mathrm{PSPD}_2 = 0.511$ nats, because the dynamics carry the revealed bit into the target. In the support row, a residual of 0.128 nats and positive values of all three measures leave $\mathrm{PSPD}_2 = 0$, since deployment never takes the revealing actions. The residual and the three measures average over teacher histories without deployment weights, so each of them misses this case. In the control row, $\mathrm{TAL}_1 = 0$ while ι^{fwd} and ι^{obs} equal 0.368 nats, because they also count the action's control of the target. In the copied-bit row, $\mathrm{PSPD}_2 = 0.511$ exceeds $\mathrm{TAL}_1 = 0.368$. In the three coupled-bit rows, observations are constant and actions have no control effect, so all three measures equal the residual and order the rows as PSPD_2 does. This matched design alone does not separate per-target action information from the criterion, and the rows that do rely on informative observations, dynamics, deployment support or control. Appendix B.5 builds sequential channels with the explaining-away

structure, called masked parity there, and the transport and support structures and trains paired learners on them; there the learned gap follows PSPD, not per-target action information, in both directions.

B CONTROLLED LEARNING WITH MATCHED TARGET UNCERTAINTY

This experiment tests whether target alignment changes learned policy-swap risk when public histories, teacher information, and target uncertainty are held fixed. The Controlled Coupled Markov Channels task extends the two-time example in Appendix E.1 to recurrent prediction with state transitions and action-dependent control. It is a purpose-built synthetic mechanism study. Its contribution is the separation of these factors in a finite-learning comparison.

B.1 COUPLED CHANNELS AND EXACT FILTERING

An episode has 60 prediction frames, indexed locally by $t = 0, \dots, 59$, and eleven hidden pairs $(X_{tj}, Y_{tj}) \in \{0,1\}^2, j = 0, \dots, 10$. The prediction target is always the vector Y_t. Public observations contain eleven cue bits C_t and an independent uniform query $Q_t \in \{0, \dots, 10\}$. Cue bits start independently fair and each retains its value with probability η at the next step, otherwise flipping. The history is $H_t = (C_{0:t}, Q_{0:t}, A_{0:t-1})$. The full Markov state augments the 22 hidden bits with the public cues and query; these observed components enter the transition kernel.

Each action $a \in \{0, \dots, 15\}$ controls every pair through the bit $m_j(a) = \lfloor a/2^{j \bmod 4} \rfloor \bmod 2$. Both channels have the same marginal transition probability

$$f_j(z, a, C_t) = 0.85\big(z \oplus m_j(a)\big) + 0.15(0.1 + 0.8C_{tj}). \tag{61}$$

Correlation $\rho \in \{0, 0.5, 1\}$ changes the dependence between the two transition innovations. For current pair (x, y), write $u = f_j(x, a, C_t)$ and $v = f_j(y, a, C_t)$. Its four-state transition kernel is

$$\begin{aligned} T_\rho(11 \mid x, y, a, C_t) &= (1-\rho)uv + \rho \min(u, v), \\ T_\rho(10 \mid x, y, a, C_t) &= u - T_\rho(11 \mid x, y, a, C_t), \\ T_\rho(01 \mid x, y, a, C_t) &= v - T_\rho(11 \mid x, y, a, C_t), \\ T_\rho(00 \mid x, y, a, C_t) &= 1 - u - v + T_\rho(11 \mid x, y, a, C_t). \end{aligned} \tag{62}$$

Equivalently, $X_{t+1,j}$ uses a uniform innovation U^X_{tj}, while $Y_{t+1,j}$ uses that same uniform with probability ρ and an independent uniform otherwise. Initial pairs have mass $(1-\rho)/4 + (\rho/2)\mathbf{1}\{x = y\}$ on (x, y). Pairs have independent innovations conditional on the public controls. The correlation changes hidden joint occupancy, while preserving each channel's marginal transition and the action's control role.

The teacher's action parity is a noisy copy of the queried X bit:

$$\mu(a \mid H_t, X_t, Y_t) = \frac{1}{8} \begin{cases} 0.9, & a \bmod 2 = X_{t,Q_t}, \\ 0.1, & a \bmod 2 \neq X_{t,Q_t}. \end{cases} \tag{63}$$

The other three action bits are independent and uniform. Deployment uses either a uniform action or a public-cue policy whose probability of odd parity is $0.2 + 0.6C_{t,Q_t}$, again with uniform remaining bits. All policies have full action support.

Exact filtering maintains eleven four-state factors. At each frame we record the current target, do marginal p_{tj} and behavior marginal $q^{Y,\rho}_{tj}$ (the behavior marginal of the target Y, which depends on ρ) before choosing A_t. The behavior filter weights the queried factor by equation 63. Both filters then apply equation 62 using C_t and A_t. New cues and queries are generated afterward. On deployment histories, $q^{Y,\rho}$ evaluates the original teacher law, whereas p is the actual history-only policy posterior. The GRU receives the chosen action at the next frame through its previous-action input.

B.2 WHAT THE COUPLING COMPARISON HOLDS FIXED

Proposition 4 (Matched histories, uncertainty, and teacher information)**.** *For this task, the three coupling conditions admit identical public trajectories under each specified policy. At every such history,*

$$\begin{aligned} P^{\mathrm{do}}_\rho(Y_{tj} = 1 \mid H_t) = P^{\mathrm{do}}_\rho(X_{tj} = 1 \mid H_t) &= p_{tj}, \\ p_{t+1,j} &= 0.85\big((1 - m_j(A_t))p_{tj} + m_j(A_t)(1 - p_{tj})\big) + 0.15(0.1 + 0.8C_{tj}), \end{aligned} \tag{64}$$

with $p_{0j} = 1/2$, independently of ρ. Writing h for binary entropy, the do-conditional entropy of Y_t is $\sum_j h(p_{tj})$ for every coupling. The teacher's pointwise conditional information is also invariant:

$$\iota_t(H_t) = h\big(0.1 + 0.8 q^X_{t,Q_t}\big) - h(0.1), \qquad q^X_{tj} = P^\mu_\rho(X_{tj} = 1 \mid H_t), \tag{65}$$

where q^X does not depend on ρ. Consequently, $\mathbb{E}_{H_t \sim P^\mu} \iota_t(H_t)$ and its cumulative sum TAL are identical across coupling conditions. At $\rho = 0$, $q^{Y,0}_{tj} = p_{tj}$; at $\rho = 1$, $X_t = Y_t$ almost surely.

Proof. Use common X innovations, cue and query streams, and action random numbers across coupling conditions. The evolution of X and each policy's action ignore Y, so their public histories coincide pathwise. Marginalizing equation 62 gives the same transition for X and Y. Applying it to their initially equal do marginals yields equation 64. Initial pair independence, factorwise transitions, and a teacher likelihood involving just the queried pair preserve factorization across pairs. Thus the do posterior of Y_t is a product of Bernoulli laws, giving the entropy expression.

Marginalizing the queried pair's likelihood update over Y gives a binary update involving X alone. Its subsequent marginal transition is independent of ρ, so induction gives a common q^X. The conditional entropy of the teacher action is $h(0.1) + \log 8$ given the full state and $h(0.1 + 0.8 q^X_{t,Q_t}) + \log 8$ given the history. Their difference gives equation 65. At $\rho = 0$, the independent Y channel receives no teacher likelihood update and follows its do filter. At $\rho = 1$, equal initial bits and shared innovations preserve $X = Y$. □

Equation 65, evaluated on deployment histories, describes the original teacher's counterfactual information. The actual history-only deployment policy has zero conditional state-action information. The matching proposition controls the population comparison. Finite GRU approximation and optimization can still change its learned risk ordering.

B.3 Learning protocol and estimands

The protocol specifies ten independent data/training seeds, partitioned in advance into an initial group (101–105) and a confirmation group (106–110). Each seed has 2,048 behavior-training episodes and 256 behavior-validation episodes at $\eta = 0.9$. Separate test streams contain 256 episodes for each of the three policies and each cue-retention value $\eta \in \{0.9, 0.6\}$. Within a stream, all coupling conditions use the same public trajectories and paired innovations. Both seed groups use the same parameters and estimands.

The model follows the one-layer GRU training template in Appendix D.3, with 128 hidden units and eleven Bernoulli logits. Its 38 inputs are the eleven current cues, eleven query one-hot channels, and sixteen previous-action one-hot channels, initially zero. It has 65,931 parameters. Hidden bits, reference posteriors, current actions, and ρ are absent from these inputs. For each seed, six models use identical initial parameters, public training histories, and ordered minibatches: do-soft, observational-soft at $\rho = 0.5$ and 1, and hard state supervision at all three correlations. Observational-soft at $\rho = 0$ and do-soft share the same targets and model. The hard-label arms assess supervision noise.

The training budget is 20,000 Adam updates per model, learning rate 10^{-3}, batch size 64, and gradient-norm bound 1. Loss sums eleven Bernoulli cross-entropies and averages equally over episodes and all 60 frames. Models use Float32 without mixed precision or TF32. The selected checkpoint minimizes its own behavior-validation objective over updates $0, 250, \ldots, 20000$, taking the earliest exact tie. Training always continues to update 20,000, and both selected and final weights are retained.

Let $\widehat{R}_{s,\pi,\rho}(g)$ be the mean test-episode negative log likelihood of the eleven realized Y labels, summed within each frame and averaged across its 60 frames. At cue retention 0.9, define

$$D_s(\rho) = \frac{1}{2} \sum_{\pi \in \{\text{uniform},\text{cue}\}} \left[\widehat{R}_{s,\pi,\rho}(g^{\text{obs}}_{s,\rho}) - \widehat{R}_{s,\pi,\rho}(g^{\text{do}}_s) \right]. \tag{66}$$

The primary estimand is the mean of $D_s(1) - D_s(0)$, assessed separately in the two predefined seed groups. Here $D_s(0) = 0$ because the soft predictors coincide. A separately declared companion measures the observational-soft minus do-soft realized NLL on fresh behavior test episodes at $\rho = 1$. The hypotheses are positive deployment and negative behavior contrasts. They concern fitted predictors, not the exact KL risk identity.

Secondary analyses cover intermediate coupling, cue retention 0.6, individual policies, hard-label training, final checkpoints, exact posterior distortion, prediction error, and immediate classification regret. These secondary analyses, including the cue-retention 0.6 generalization test, were listed as secondary in the protocol frozen before collection (`protocol/target_alignment_learning.json`) and are reported with unadjusted intervals. Model-selection comparisons choose between the two selected soft predictors using a common behavior-validation label loss or a common do-reference cross-entropy, then evaluate the frozen choice on test episodes. The latter rule requires exact do references. Its test regret relative to the better of the two candidates is an evaluation quantity. The study measures prediction and selection under known dynamics. Control returns and computational acceleration are outside its scope.

Unadjusted 95% percentile intervals use 10,000 paired seed bootstrap resamples, with five seeds per predefined group and ten when pooled. Each seed first averages episodes and then equally weights the two target policies for the primary endpoint. Secondary comparisons are exploratory. Exact posterior quantities diagnose the mechanism, whereas realized-label losses test its expression in finite recurrent learners.

B.4 Results and interpretation

Table 6 reports every coupling value in both predefined groups. At $\rho = 1$, the deployment contrasts are $+0.768$ $(+0.747, +0.787)$ and $+0.763$ $(+0.750, +0.775)$ nats. The corresponding behavior contrasts are -0.710 $(-0.736, -0.684)$ and -0.683 $(-0.697, -0.671)$. Every seed has both directions. At $\rho = 0.5$, both groups have smaller positive deployment and negative behavior contrasts. The two target policies and shifted public cues preserve these directions. Zero coupling gives exact equality of the soft objectives and shared parameters by construction; its zero contrast is not an independent training confirmation.

At unit coupling and default cues, behavior-label validation selects observational-soft in all ten seeds, while exact-do-reference validation selects do-soft. The resulting difference in test selection regret equals the soft-predictor contrast. This result identifies a failure of behavior validation under the specified policy swap. The reference-assisted rule requires the known filter, and always choosing do-soft achieves the same test loss here; the experiment does not establish a superior adaptive selector.

For the pooled selected models at unit coupling, exact marginal PSPD is 0.765 nats and observational-soft's do-relative KL is 0.760 nats, versus 0.00044 for do-soft. In this controlled task, the learned contrast is close to the exact distortion, unlike RockSample's larger approximation effects. At $\rho = 1$ each target bit equals the channel bit the teacher reads when it is queried ($Y_t = X_t$), and the fitting terms of equation 11 are near zero (Table 4A), so this study isolates the coupling mechanism at matched information rather than testing robustness to fitting error. At $\rho = 0.5$, where Y_t and X_t can differ, the learned expected gap G is 0.080 nats against an exact distortion D_{marg} of 0.082 nats. Full-budget final soft models retain the reversal: deployment contrasts are 0.769 (0.748, 0.788) and 0.763 (0.750, 0.775).

Hard-label supervision has pooled deployment NLL 7.744, between observational-soft's 8.113 and do-soft's 7.348. A post hoc analysis of the recorded hard-label arms shows that finite hard-label training is not a numerical stand-in for observational-soft here. At $\rho = 1$ the selected hard-label predictor has behavior-history NLL 7.358, worse than do-soft's 7.340 in 10 of 10 seeds, whereas observational-soft reaches 6.643. Its KL from the behavior marginals, 0.710 nats, is no smaller than do-soft's 0.702. At $\rho = 0$, where the two soft targets coincide, hard-label minus do-soft NLL is $+0.191$ nats on both behavior and deployment histories, positive in 10 of 10 seeds. Relative to this penalty, coupling moves the hard-label arm in the direction of the soft reversal. The difference-in-differences, hard-minus-do-soft NLL at $\rho = 1$ minus the same contrast at $\rho = 0$, is -0.173 $(-0.178, -0.168)$ on behavior histories and $+0.206$ $(+0.182, +0.230)$ at deployment, with that sign in 10 of 10 seeds (pooled ten-seed intervals). The hard-label arms reach their own validation minimum at 750–1,000 of the 20,000 updates and were not entered in validation selection, which compares only the two soft predictors. Final hard-label models fit worse out of sample, with hard-minus-do gaps of 3.468 (3.414, 3.529) and 3.517 (3.438, 3.614), and their behavior-side difference-in-differences reverses sign: $+0.146$, positive in 8 of 10 seeds. The released seed tables include all hard-label, classification, policy-specific and final-model results.

The study uses 38,400 base history episodes shared across the three couplings and 1,200,000 optimizer updates across 60 models; training takes 2,575.1 summed seed-job seconds on one NVIDIA RTX PRO 6000 Blackwell GPU, a sum over concurrently running jobs rather than elapsed wall time. An

independent CPU replay of the selected-model validation sequences and the primary soft models on three test policies takes 4.55 seconds for 30,720 model–episode sequences. Independent checks reconstruct four-state filtering on a complete trajectory from each of 80 data files, all 3,888 reported paired contrasts, 360 model choices and 54 choice-rule intervals, with filter differences below 3.4×10^{-15} and contrast differences below 1.1×10^{-14}. CPU/GPU forward replay agrees within 2.0×10^{-5} nats per episode.

Table 6: Controlled Markov-channel learning. Entries are observational-soft minus do-soft realized-label NLL (sum over eleven targets), with paired five-seed 95% bootstrap intervals. Within each collection policy, public histories, do target marginals and teacher information match across ρ. At $\rho = 0$, both soft objectives share the same trained predictor. Deployment averages the uniform and public-cue policies at cue retention 0.9; the shifted-cue column repeats deployment at cue retention 0.6.

Group	ρ	Behavior	Deployment	Deployment, shifted cues
Initial	0	$+0.000$ ($+0.000$, $+0.000$)	$+0.000$ ($+0.000$, $+0.000$)	$+0.000$ ($+0.000$, $+0.000$)
Initial	0.5	-0.084 (-0.089, -0.079)	$+0.078$ ($+0.070$, $+0.086$)	$+0.080$ ($+0.075$, $+0.083$)
Initial	1	-0.710 (-0.736, -0.684)	$+0.768$ ($+0.747$, $+0.787$)	$+0.781$ ($+0.769$, $+0.795$)
Confirmation	0	$+0.000$ ($+0.000$, $+0.000$)	$+0.000$ ($+0.000$, $+0.000$)	$+0.000$ ($+0.000$, $+0.000$)
Confirmation	0.5	-0.086 (-0.092, -0.079)	$+0.081$ ($+0.078$, $+0.086$)	$+0.082$ ($+0.080$, $+0.084$)
Confirmation	1	-0.683 (-0.697, -0.671)	$+0.763$ ($+0.750$, $+0.775$)	$+0.777$ ($+0.762$, $+0.792$)

B.5 A PAIRED LEARNING TEST OF PER-TARGET ACTION INFORMATION

The finite examples of Table 5 separate per-target action information from the criterion in exact arithmetic. This study asks which of the two a finite recurrent learner follows when they disagree. It was designed after the channel and RockSample results were known and is therefore post hoc. Its environments, estimands and decision rules were fixed in a protocol before any of its data were collected, and it is run by `r39_per_target_learning.py` and summarized by `r39_per_target_report.py`.

Variants. Four variants of the channels above keep 11 pairs, 60 frames, sixteen actions and a teacher whose parity bit is correct with probability 0.9 at the queried pair, but no action affects any hidden bit, cue, query or pad. A persistent bit X_j is kept with probability 0.85 and otherwise redrawn from a cue-dependent Bernoulli law, and every frame draws a fair pad bit that becomes public at the next frame. In *copy* the parity copies X_j at the query and the targets are the X_j. In *masked parity* the parity is the same copy evidence combined by exclusive or with the pad. At the time of the action it is independent of every target, so per-step action information about every target is zero. Once the pad is revealed, the history carries the same evidence about the target as in copy. In *transport* a fresh bit drawn at each frame is copied by the parity and, with probability 0.8, overwrites the target register of the queried pair, so the action carries no information about the current target while the forward information ι_t^{fwd} is positive. In *support* the parity copies the target only when the teacher sets a probe bit, with probability 0.25, and is fair otherwise. Copy, masked parity and transport are deployed under a uniform and a cue-driven history-only policy; under the uniform policy copy and masked parity have identical labels and identical PSPD. Support is deployed under a policy that never probes, with PSPD zero, and one that probes at the teacher's rate.

Protocol. The 10 seeds form an initial group, 101–105, and a confirmation group, 106–110. Each seed has 2,048 training, 256 validation and 256 test episodes per policy and cue cell. Per seed, eight 128-unit GRUs, do-soft and observational-soft for each variant, share one initialization and one minibatch order, and their inputs are those of the channels above plus the revealed pad. Training uses Adam with learning rate 10^{-3}, 20,000 updates and batches of 64, and the selected checkpoint minimizes the arm's own validation loss over candidates every 250 updates. The primary endpoints are masked parity and transport: the deployment gap of equation 66 at cue retention 0.9 and the behavior gap on fresh teacher histories. A variant meets the rule when, in both groups, the deployment interval lies above zero and the behavior interval below zero, with at least nine of ten seeds positive at deployment. Two secondary readings were fixed with it. The ratio of the masked-parity gap to the copy gap under uniform deployment reads as consistent with PSPD if its lower bound exceeds 0.5, whereas per-target action information predicts a ratio near zero. Support is confirmed if the never-probe gap lies within ± 0.02 nats and the probe-rate gap above zero. Before the freeze, an

exhaustive enumeration of a two-pair, three-frame instance checked both filters, per-step and forward action information, the copy and masked-parity identity and the zero never-probe PSPD to within 10^{-10}. The report script was written after the freeze, while the seeds trained, and after the first report the ratio was printed to three instead of two decimals. No estimate or decision changed. Intervals are paired seed-bootstrap percentile intervals with 10,000 resamples.

Table 7: Paired learning test of per-target action information on four controlled channel variants. Nats per frame, eleven targets summed, cue retention 0.9 unless stated. Copy, masked parity and transport average the uniform and cue deployment policies equally. Parentheses: paired seed-bootstrap 95% intervals.

A. Exact measures	ι^{act} per step	ι_t^{act}	ι_t^{fwd}	$\mathcal{I}_t^{\mathrm{on}}$	PSPD
Copy	0.254	7.764	0.574	0.616	0.751
Masked parity	0.000	0.000	0.000	0.616	0.789
Transport	0.000	0.000	0.483	0.567	0.637
Support, no probe	0.069	2.107	0.160	0.170	0.000
Support, probes	0.069	2.107	0.160	0.170	0.223

B. Learned gap, 10 seeds	Deployment	> 0	Behavior	Final	Cue 0.6
Copy	0.752 (0.743, 0.759)	10/10	-0.615 (-0.628, -0.602)	0.752	0.831
Masked parity	0.773 (0.765, 0.781)	10/10	-0.604 (-0.616, -0.592)	0.773	0.836
Transport	0.634 (0.628, 0.640)	10/10	-0.574 (-0.581, -0.567)	0.634	0.573
Support, no probe	0.001 (0.000, 0.002)	9/10	-0.173 (-0.177, -0.168)	0.001	0.002
Support, probes	0.213 (0.207, 0.220)	10/10	-0.173 (-0.177, -0.168)	0.214	0.231

C. Decomposition	D_{marg}	E_{do}	E_{obs}	B_{obs}	G
Copy	0.751	0.000	0.005	-0.004	0.747 (0.744, 0.749)
Masked parity	0.789	0.000	0.016	-0.020	0.768 (0.767, 0.770)
Transport	0.637	0.000	0.001	-0.002	0.635 (0.634, 0.636)
Support, no probe	0.000	0.000	0.001	0.001	0.001 (0.001, 0.001)
Support, probes	0.223	0.000	0.003	-0.004	0.218 (0.217, 0.220)

	Initial group		Confirmation group		
D. Frozen rule	Deployment	Behavior	Deployment	Behavior	Rule
Copy (reference)	(0.747, 0.763)	(-0.637, -0.601)	(0.734, 0.761)	(-0.629, -0.596)	met
Masked parity	(0.756, 0.786)	(-0.626, -0.590)	(0.769, 0.780)	(-0.616, -0.588)	met
Transport	(0.623, 0.645)	(-0.576, -0.560)	(0.628, 0.640)	(-0.589, -0.569)	met

A: teacher-side measures on teacher test histories; PSPD is the exact D_{marg} on deployment histories. ι_t^{act} and ι_t^{fwd} are the cumulative terms of Section 4.3, averaged over 60 frames. B: observational-soft minus do-soft realized-label NLL on deployment and fresh teacher histories, selected checkpoints unless stated. C: selected checkpoints, seed means; G is the do-reference gap with its interval, and $G = D_{\mathrm{marg}} - E_{\mathrm{do}} + B_{\mathrm{obs}}$ holds to within 10^{-15}. D: a variant meets the frozen rule when, in both five-seed groups, the deployment interval lies above zero and the behavior interval below zero, and at least 9 of 10 seeds have a positive deployment gap. Under uniform deployment, where copy and masked parity share labels and PSPD, the ratio of their mean gaps is 0.980 (0.976, 0.983). Support: the no-probe gap 0.001 (0.000, 0.002) is within the frozen band of ± 0.02 nats, and the probe-rate interval lies above zero.

Results. Table 7 reports every variant. Both primary endpoints and the copy reference meet the rule. In masked parity, observational-soft loses at deployment by 0.773 nats (0.765, 0.781), positive in 10 of 10 seeds, while its behavior gap on teacher histories is -0.604 nats. In transport the two gaps are 0.634 and -0.574. Under uniform deployment, where copy and masked parity share labels and a PSPD of 0.787 nats, the ratio of their gaps is 0.980 (0.976, 0.983), although copy's per-step action information is 0.254 nats per frame and masked parity's is zero. Support evaluates the same trained predictors, with the same per-target action information, under both deployment policies, yet its gap is 0.213 (0.207, 0.220) when deployment probes and 0.001 (0.000, 0.002) when it never probes. The latter is inside the prespecified band but not zero: it is positive in 9 of 10 seeds and comes from a signed fitting term B_{obs} of 0.001 nats. In every variant E_{do} is 0.000 to three decimals and the learned gap G tracks D_{marg}, panel C; the largest shortfall is masked parity's, 0.768 against 0.789 nats. Selected checkpoints fall between updates 18,000 and 20,000, and the observational-soft arms of masked parity, transport and support are selected at update 20,000 in the median seed, so they

were still improving slightly when training stopped. Masked parity's E_{obs} is 0.016 against copy's 0.005. The variants have no control effect and use a teacher that acts on one queried pair, so the study tests which quantity finite learners follow, not the size of the per-target measures in RockSample, where Appendix C.7 compares them in exact arithmetic.

C CORE ROCKSAMPLE LEARNING EVIDENCE

This section collects the public-value training study, its frozen-model diagnostics, the sensitivity and selection study, a study with more training layouts, the intermediate-mixture study, the per-target action-information analysis with its tolerance and misspecification follow-up, a closed-loop evaluation and the reproduction paths. The frozen-model diagnostics reuse the trained models and histories and leave the two primary selected-checkpoint comparisons and all prespecified secondary analyses unchanged.

C.1 TRAINING WITH COMMON PUBLIC-VALUE CONSTRAINTS

In the historical Original models at full mixing, about 61% of do-soft's raw training own-target error lay on publicly determined coordinates: public clamping lowered it from 1.366 to 0.535 nats. We therefore remove those coordinates from both supervision losses and test unresolved prediction on five new seeds, excluding earlier histories and models.

Loss and public information. Let $V_{it} = \mathbf{1}\{t < L_i\}$, and let the public-prefix mask K_{itj} indicate that rock j's current quality is determined before action t: earlier sampling sets it to zero, and a check at the unsampled rock's own location reveals it exactly. Noisy checks from other locations, current actions, future readings and hidden qualities do not enter K. With $U_{itj} = 1 - K_{itj}$, the losses for either target b are

$$\mathcal{L}_{\text{all}} = \frac{\sum_{it} V_{it} \sum_j \text{BCE}(b_{itj}, q_{itj})}{\sum_{it} V_{it}}, \tag{67}$$

$$\mathcal{L}_{\text{constrained}} = \frac{\sum_{it} V_{it} \sum_j U_{itj} \text{BCE}(b_{itj}, q_{itj})}{\sum_{it} V_{it}}. \tag{68}$$

Finite-logit BCE gives zero constrained-loss gradient on determined coordinates, without rescaling unresolved terms or changing the denominator. Common public clamping at evaluation leaves unresolved predictions unchanged.

Models, data and selection. Five new seeds, $\lambda \in \{0, 1\}$, two supervision targets and two losses give forty GRUs with the template of Appendix D.3 (20,000 updates at learning rate 0.001). All four arms share initialization and ordered minibatches within seed and mixture; at zero mixing, both supervision arms run under each recipe despite coincident targets.

The 51,200 new episodes span 24 new layouts: eight for training and four each for behavior validation, independent behavior evaluation and two unseen deployment sets. Independent histories on the eight training layouts add seen behavior and deployment evaluation. Four public policies share histories across models and mixtures; split-specific streams separate training, selection and evaluation. Behavior retains the original support mixture.

Both recipes minimize own-target, common-clamped validation BCE pooled over valid frames, using 81 candidates $0, 250, \ldots, 20000$ and the earliest exact tie. All-target-loss arms are baselines under this rule. Selected and final checkpoints are retained. Training uses Float32; scoring uses Float64 with exact target probabilities. Independent evaluation follows all forty training runs and selections.

This study and the sensitivity study (Appendix C.4) cover only $\lambda \in \{0, 1\}$, with $\epsilon = 0.05$, the support floor used in every study. Appendix C.6 retrains the constrained recipe at $\lambda \in \{0.25, 0.5, 0.75\}$ with the same seeds; the historical study adds $\lambda = 0.5$ runs with a different recipe, where the Original and Fresh estimates differ (Appendix E.3). Exact distortion across λ and ϵ on this study's deployment histories is in Table 14.

Primary quality comparison. For each deployment episode, define unresolved against-do loss mass

$$Z_i(q) = \frac{1}{60} \sum_{t,j} V_{it} U_{itj} D_{\text{KL}}\big(\text{Ber}(b^{\text{do}}_{itj}) \| \, \text{Ber}(q_{itj})\big) \, . \tag{69}$$

This eleven-coordinate sum is zero after termination. We average episodes, policies and layouts, weighting seen deployment by $1/2$ and each unseen set by $1/4$, equally weighting all sixteen layouts. The two primary $\lambda = 1$ contrasts, separately by supervision target, are $\Delta = Z(q_{\text{constrained}}) - Z(q_{\text{all}})$ for selected checkpoints. Negative values indicate improved prediction. The observational-minus-do gap is a separate outcome.

Ten thousand with-replacement paired seed-bootstrap draws provide descriptive 95% percentile intervals and 97.5% Bonferroni sensitivity intervals for the two primary contrasts. The intervals summarize variation across these seeds conditional on the fixed layouts. With five seeds, nominal and family-wise coverage are unverified; the wider percentile limits serve as sensitivity summaries. Other intervals are secondary. Table 11 reports the five individual effects for both primary contrasts and the fixed-final do-soft comparison. Conditional unresolved risks divide pooled loss by pooled coordinate mass, recomputed in every draw and undefined at zero mass. Original $1/L_i$ scores, raw/clamped losses, reference entropies, Public and exact-do scores are retained.

Table 8: Training public-value constraints: unresolved deployment KL loss mass at full mixing. All models share architecture, initialization, batches, budget and public-clamped own-target selection. Scores divide valid-frame loss by 60, then equally weight seen and unseen layouts and four policies. Differences are constrained minus all-target; intervals are paired five-seed bootstrap 97.5% sensitivity intervals for the two primary comparisons.

Supervision	All-target	Constrained	Difference [97.5%]
Observational-soft	3.052	2.549	-0.504 (-0.644, -0.407)
Do-soft	1.987	0.147	-1.841 (-1.856, -1.824)

Absolute prediction improves most for do-soft. Both selected arms improve in every seed (Table 8). Observational-soft's unresolved deployment KL mass falls from 3.0522 to 2.5485 nats, whereas do-soft falls from 1.9871 to 0.1466. All-target-loss do-soft selects update zero in all five seeds after completing the training budget; these selected models retain their initialization. Constrained do-soft selects updates 19,250–20,000. The fixed-budget do comparison also improves: KL mass falls from 1.2319 to 0.1462, with paired difference -1.0857 ($-1.1962, -0.9751$) (secondary 95% interval). Public has KL mass 1.9669 on the same histories. Constrained do-soft removes 92.5% of this Public-to-exact-do expected-loss gap, a descriptive ratio of aggregate means. Observational-soft improves but remains worse than Public.

Own-target KL distinguishes the learners' remaining fit. On independent seen behavior histories, the constrained selected models have unresolved own-target KL 0.16802 for observational-soft and 0.000685 for do-soft, in nats per rock. Their corresponding values are 0.28292 and 0.01273 on seen deployment, and 0.31304 and 0.02217 on unseen deployment. Each KL subtracts its own target entropy; different-target BCE values are not comparable fitting errors. Do-soft closely approximates its reference, while observational-soft retains appreciable own-target error. Under behavior histories, exact do is only a common reference and the true posterior is b^{μ}; under public deployment it is b^{do}.

Table 9: Same-layout risk reversal after the training intervention. Common-clamped realized NLL uses fixed-60 weights on fresh behavior and four pooled deployment policies. Differences are observational-soft minus do-soft; parentheses are secondary unadjusted paired-seed 95% intervals. Reversal requires a negative behavior difference and a positive deployment difference.

Checkpoint	Training loss	Behavior difference	Deployment difference
Selected	All-target	-0.682 (-0.708, -0.651)	+1.161 (+1.084, +1.260)
Selected	Constrained	-0.480 (-0.505, -0.451)	+2.533 (+2.457, +2.602)
Final	All-target	-0.982 (-1.008, -0.946)	+12.955 (+12.830, +13.049)
Final	Constrained	-1.018 (-1.042, -0.994)	+11.127 (+10.917, +11.337)

Reversal and continued training. For each recipe, reversal requires both a negative observational-minus-do realized NLL under seen behavior histories and a positive difference under seen deployment. Both selected recipes satisfy these directions (Table 9). The constrained contrasts are -0.480 and $+2.533$ nats. Reversal therefore persists with much better do-soft prediction; it does not require both learners to remain poorly fitted. The enlarged gap is not itself evidence of quality gain.

Final-budget observational-soft checkpoints sharply worsen deployment fit: original and constrained KL masses are 14.8441 and 11.8653, versus their selected values 3.0522 and 2.5485. Common validation selection is consequently material to the reported observational-soft quality. The intervention establishes an empirical loss-recipe effect in this five-seed experiment, without identifying gradient conflict or representation competition. These newly trained models have not been tested for control return or certified by the structural equality rules.

Collection-policy comparison on shared deployment histories. We also compare the existing constrained, selected do-soft models trained at $\lambda = 0$ and $\lambda = 1$ (Table 10). This exploratory reanalysis uses the completed training runs and their shared deployment archives. At $\lambda = 0$, collection uses the fixed history-greedy rule, fed by exact do beliefs, with the same $\epsilon = 0.05$ uniform support mixture. Both collection arms use exact do-soft training targets, so the comparison concerns the collection policy and its training occupancy; it does not evaluate online student adaptation or DAgger (Ross et al., 2011).

On the three-group measure, $\lambda = 0$ yields KL mass 0.0872 versus 0.1466 at $\lambda = 1$; on the eight seen layouts, the corresponding values are 0.0308 and 0.1071. All five paired differences favor history-only collection. The fixed-log supervision comparison and the collection-policy comparison thus answer different questions: correcting the targets improves prediction within the privileged logs, while these data provide no prediction advantage for privileged over history-only collection after applying the same constrained do-soft recipe.

Under the constrained recipe, $\lambda = 0$ hard-label training is worse than $\lambda = 1$ do-soft (Appendix C.4), the reverse of the historical ordering, where the $\lambda = 0$ collector is the history-greedy rule on exact do beliefs (Appendix E.3). At $\lambda = 1$ the constrained-recipe hard-label arm repeats the historical arm's signs, better than do-soft on behavior and worse on deployment histories, at its early selected checkpoints.

Table 10: Collection-policy comparison of constrained, selected do-soft models on shared deployment histories. Lower unresolved KL mass is better (nats). Both collection arms use exact soft targets; the $\lambda = 0$ collector is the fixed history-greedy rule, which is also one of the four deployment policies. Means average five paired seeds; parentheses are descriptive 95% percentile bounds from all 5^5 ordered seed resamples. All five paired differences are negative. This reanalysis uses existing trained models.

Deployment layouts	$\lambda = 0$	$\lambda = 1$	$\lambda = 0$ minus 1
All deployment layouts	0.087	0.147	-0.059 $(-0.067, -0.053)$
Seen training layouts	0.031	0.107	-0.076 $(-0.085, -0.070)$

Table 11: All five seed values for the two primary selected intervention effects, the fixed-final do-soft effect, and the exploratory collection-policy comparison. Intervention differences are constrained minus all-target at $\lambda = 1$. Collection columns use constrained selected do-soft models, with Δ_{0-1} denoting $\lambda = 0$ minus $\lambda = 1$. The three-group measure retains the seen/unseen-A/unseen-B weights $1/2, 1/4, 1/4$ and equal weight for the four policies. Entries are fixed-60 unresolved KL loss masses (nats); the seen column uses the eight training layouts.

	Constrained minus all-target			Three-group measure			Seen
Seed	Obs sel.	Do sel.	Do final	$\lambda = 0$	$\lambda = 1$	Δ_{0-1}	Δ_{0-1}
201	-0.3826	-1.8163	-0.9186	0.0801	0.1535	-0.0734	-0.0932
202	-0.4715	-1.8502	-1.1863	0.0891	0.1519	-0.0628	-0.0756
203	-0.4167	-1.8633	-0.9727	0.0828	0.1359	-0.0531	-0.0667
204	-0.4978	-1.8343	-1.2576	0.0953	0.1484	-0.0531	-0.0725
205	-0.7499	-1.8387	-1.0931	0.0886	0.1431	-0.0545	-0.0737
Mean	-0.5037	-1.8406	-1.0857	0.0872	0.1466	-0.0594	-0.0763

Execution and retained evidence. Collection and exact references take 28.79 seconds, training with validation 797.21 seconds, and evaluation 77.74 seconds on one RTX PRO 6000 Blackwell GPU. An independent implementation checks all 80 selected/final weight records, the 3,240 validation candidates, the compact statistics, one raw episode per evaluation context and the retained logits, without rerunning training (paths in Appendix C.10; paired retraining is supplied as a command and

was not rerun). `collection_analysis/collection_tables.py` regenerates the collection comparison and per-seed table from the retained seed scores, with percentile bounds from all $5^5 = 3{,}125$ ordered seed resamples; this removes resampling Monte Carlo error without establishing small-sample coverage.

C.2 Exact distortion and finite predictions on a common measure

The following exploratory diagnostics rescore existing selected and final models without changing training, checkpoint selection or the prespecified primary comparisons. They apply equation 25 to the controlled channels and public-value training study separately. All terms within a comparison share histories, targets, masks, time weights and aggregation. With $p_j = b_j^{\mathrm{do}}$ and $b_j = b_j^{\mu}$, D_{marg} measures exact marginal distortion, E_{obs} and E_{do} measure own-target KL, and B_{obs} is the signed expectation of $\log b_j - \log q_{\mathrm{obs},j}$ under p_j. We calculate this expectation directly from the references and logits, then check $G = D_{\mathrm{marg}} - E_{\mathrm{do}} + B_{\mathrm{obs}}$. Neither E_{obs} nor a rearranged identity substitutes for that calculation. These are signed score terms, not independent causal effects. Per-policy and per-layout strata of Table 4 with E_{obs}, E_{do} and against-do scores are in Table 13; seed values and paired 95% intervals are in the core export's estimate and seed-score files (Appendix C.10).

Measures and normalization. The controlled channels retain all eleven targets, all 60 frames and equal episode weights. Public histories are matched across coupling conditions within each collection policy, not between behavior and deployment policies. Core summaries use cue retention 0.9 and equal weights for the two deployment policies in each five-seed group. At $\rho = 0$, the shared predictor gives $D_{\mathrm{marg}} = G = 0$; its own-target errors and B_{obs} can be nonzero and cancel in the identity.

RockSample retains the fixed-60 loss mass and public unresolved mask of Table 2. The three-group measure gives seen, unseen-A and unseen-B weights $1/2, 1/4, 1/4$, with equal layouts within each split and four equally weighted policies. Seen-only comparisons remain separate. Conditional per-unresolved-coordinate scores divide pooled loss mass by the same pooled coordinate mass, with numerator and denominator recomputed within every paired-seed bootstrap. No $1/L_i$ statistic is reinterpreted as fixed-time Bayes risk. On behavior histories, b^{μ} is the true posterior; scoring against do is a common-reference comparison. Expected reference scores and realized-label NLL are retained separately.

Table 12: Common-measure core summaries (nats). $A_{\mathrm{obs}} = \int \sum_j u_j D_{\mathrm{KL}}(p_j \| q_{\mathrm{obs},j})\, d(dh)$ is observational against-do KL. Expected and realized gaps are distinct. Study weights are those in Table 4; Channel rows use selected checkpoints; RockSample rows distinguish selected/final.

Study / predictors	D_{marg}	E_{obs}	E_{do}	B_{obs}	G	A_{obs}	NLL gap
Channels, initial	0.764	0.003	0.0004	-0.006	0.758	0.758	0.768
Channels, confirm.	0.766	0.003	0.0004	-0.005	0.760	0.761	0.763
RS, all / selected	18.083	2.569	1.987	-15.031	1.065	3.052	1.064
RS, constrained / selected	18.083	2.504	0.147	-15.535	2.402	2.549	2.396
RS, all / final	18.083	5.598	1.232	-3.239	13.612	14.844	13.642
RS, constrained / final	18.083	4.407	0.146	-6.218	11.719	11.865	11.727

What the decomposition explains. For the two controlled-channel groups, the selected expected deployment gaps are 0.758 and 0.760, close to exact distortions 0.764 and 0.766. Small do fitting errors and signed corrections account for this proximity. In the training study, selected constrained models instead have $D_{\mathrm{marg}} = 18.083$, $B_{\mathrm{obs}} = -15.535$ and $G = 2.402$. The observational predictor's against-do KL is 2.549, far below the exact behavior posterior's 18.083 on these same weighted histories. Its own-target error, 2.504, is a different expectation and cannot replace the negative signed correction. The smaller learned gap thus does not measure the full exact posterior distortion.

Removing publicly determined training terms leaves both reference posteriors and D_{marg} unchanged. Under the shared deployment measure, $\Delta G = \Delta R_{\pi}(q_{\mathrm{obs}}) - \Delta R_{\pi}(q_{\mathrm{do}})$. At selected checkpoints both predictors improve, with a larger do-soft gain; their gap grows by 1.337 nats. At final checkpoints both also improve, but observational-soft improves more, so the gap shrinks. A growing or shrinking gap alone is therefore not evidence of better absolute prediction. Final observational fitting and generalization are examined in Appendix C.3.

Policy and layout strata. Table 13 retains every full-mixing deployment policy and layout group for the constrained loss at selected checkpoints; the corresponding all-target-loss and final-checkpoint strata are in the core export's estimate files. Panel A of Table 4 uses the layout and four-policy weights of Table 2A, and panel B keeps the layout weights within each policy. Random actions have the smallest exact distortion but the largest selected constrained model gap. For the constrained selected models under Random, $D_{\mathrm{marg}} = 1.583$, $E_{\mathrm{obs}} = 4.077$, $E_{\mathrm{do}} = 0.313$, $B_{\mathrm{obs}} = +2.498$ (paired-seed 95% interval $(2.36, 2.63)$) and $G = 3.768$ (Table 13); at final checkpoints $B_{\mathrm{obs}} = +5.340$ and $G = 6.611$. Under the three history-driven policies (history-greedy, history-cautious and different-seed) B_{obs} is negative in every seed. In all twelve policy–layout cells, that realized gap is positive (means 1.742–3.976 nats). This pattern belongs to the training-study models and does not replace the historical Fresh random-policy result (Appendix E.3). Comparing separate confidence intervals does not test policy-effect differences. Policy-dependent occupancy and sensing are descriptive context, not identified causes of the score ordering.

Numerics, uncertainty and retained records. Bernoulli log probabilities use stable log-sigmoid expressions, and the training-study behavior references retain both log tails: 13,080 unresolved cells whose marginal rounds to one keep a nonzero complement in log space. The do reference, reconstructed from public records in log space, agrees with the cached marginals within 1.8×10^{-14}; no true posterior is clipped to force a finite score. The maximum closure error of the direct decomposition is 1.5×10^{-14} for RockSample and 4.5×10^{-16} for the channels, and the channel references reconstructed from their four-state recursion agree within 3.4×10^{-15}. Intervals are unadjusted 95% paired-seed percentile intervals over five training seeds with 10,000 draws and fixed layouts. Every seed, policy, mask, checkpoint identity, mixture or coupling condition and numerical check is retained in the training-study files `seed_scores.csv.gz` and `estimates.csv.gz` and the channel files `per_seed.csv` and `aggregate.csv`. The two prespecified primary comparisons keep their 97.5% sensitivity intervals, which these diagnostics do not replace.

Table 13: Full-mixing RockSample: constrained loss, selected checkpoints. Fixed-60 unresolved loss masses, averaged over five seeds. Layout and policy labels retain the original collections; Pooled rows use the three-group measure, with $1/2, 1/4, 1/4$ seen/unseen-A/unseen-B weights. Own-target errors use each learner's reference, while G and observational against-do KL (A_{obs}) use the do reference. The last column is the realized NLL difference. Complete unadjusted 95% seed intervals, lambda-zero controls, behavior scores and conditional-per-coordinate ratios are in the numerical files.

Layout	Policy	D_{marg}	E_{obs}	E_{do}	B_{obs}	G	A_{obs}	NLL gap
Seen	Random	1.587	4.219	0.284	2.687	3.991	4.274	3.976
Seen	greedy	24.295	1.747	0.047	-22.162	2.086	2.133	2.046
Seen	cautious	24.525	1.775	0.048	-22.346	2.130	2.178	2.147
Seen	Different seed	24.445	1.783	0.050	-22.413	1.982	2.032	1.962
Unseen A	Random	1.493	3.927	0.338	2.396	3.551	3.889	3.570
Unseen A	greedy	22.370	2.159	0.124	-20.400	1.845	1.970	1.840
Unseen A	cautious	22.338	2.195	0.126	-20.340	1.873	1.999	1.884
Unseen A	Different seed	22.573	2.175	0.122	-20.655	1.796	1.918	1.808
Unseen B	Random	1.665	3.945	0.347	2.221	3.539	3.886	3.545
Unseen B	greedy	21.034	2.178	0.145	-19.096	1.792	1.937	1.779
Unseen B	cautious	21.379	2.196	0.145	-19.382	1.853	1.997	1.904
Unseen B	Different seed	26.776	2.240	0.141	-24.833	1.803	1.944	1.742
Pooled	Random	1.583	4.077	0.313	2.498	3.768	4.081	3.767
Pooled	greedy	22.999	1.958	0.091	-20.955	1.952	2.043	1.928
Pooled	cautious	23.192	1.985	0.092	-21.103	1.997	2.088	2.020
Pooled	Different seed	24.560	1.995	0.091	-22.578	1.891	1.982	1.869

Exact distortion across collection mixtures. The training study collects behavior logs only at $\lambda \in \{0, 1\}$ with $\epsilon = 0.05$. Its deployment histories do not depend on the behavior mixture, so the exact behavior posterior of any other mixture in equation 1 can be computed on the same histories without new trajectories or training. Table 14 reports this post hoc, exploratory sweep with the masks, loss mass and weights of Table 4A; its pooled column is the three-group measure. Its $\lambda = 1$, $\epsilon = 0.05$ row reproduces the recorded seed values. On the three-group measure, D_{marg} rises monotonically with the teacher's privileged mixing, from zero at $\lambda = 0$ through 1.290 at $\lambda = 0.5$ to 18.083 at $\lambda = 1$. At $\lambda = 1$ it falls as the support floor rises, from 20.340 at $\epsilon = 0.01$ to 13.532 at $\epsilon = 0.2$.

Table 14: Exact marginal distortion D_{marg} (nats) on the deployment histories of Table 4A for other collection mixtures in equation 1 (post hoc, exploratory). Each row recomputes the exact behavior posterior b^{μ} for the stated λ and support floor ϵ on the same recorded histories; no trajectory is collected and no model is trained. Masks and fixed-60 loss mass follow Table 4. The pooled column weights training-layout, unseen-A and unseen-B histories by $1/2, 1/4, 1/4$ and the four deployment policies equally; each policy column applies the same split and layout weights within that policy. The row $\lambda = 1$, $\epsilon = 0.05$ is the recorded setting of Table 4 and appears in both panels. Second lines give unadjusted 95% paired-seed percentile intervals over five seeds with 10,000 draws.

λ	ϵ	Pooled	Random	History-greedy	History-cautious	Different-seed
A. Teacher mixing at $\epsilon = 0.05$						
0	0.05	0.000	0.000	0.000	0.000	0.000
		(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)
0.25	0.05	0.519	0.711	0.253	0.855	0.256
		(0.489, 0.553)	(0.679, 0.743)	(0.245, 0.259)	(0.754, 0.973)	(0.252, 0.261)
0.5	0.05	1.290	1.089	1.125	1.750	1.195
		(1.255, 1.321)	(1.042, 1.136)	(1.107, 1.142)	(1.636, 1.865)	(1.167, 1.229)
0.75	0.05	3.388	1.330	3.838	4.251	4.132
		(3.337, 3.438)	(1.277, 1.383)	(3.755, 3.920)	(4.122, 4.379)	(4.029, 4.300)
1	0.05	18.083	1.583	22.999	23.192	24.560
		(17.750, 18.405)	(1.525, 1.647)	(22.483, 23.505)	(22.704, 23.685)	(23.884, 25.460)
B. Support floor at $\lambda = 1$						
1	0.01	20.340	1.930	25.827	26.039	27.565
		(19.968, 20.698)	(1.854, 2.008)	(25.251, 26.392)	(25.499, 26.588)	(26.810, 28.564)
1	0.05	18.083	1.583	22.999	23.192	24.560
		(17.750, 18.405)	(1.525, 1.647)	(22.483, 23.505)	(22.704, 23.685)	(23.884, 25.460)
1	0.1	16.167	1.313	20.589	20.766	22.000
		(15.867, 16.457)	(1.268, 1.367)	(20.127, 21.045)	(20.324, 21.212)	(21.393, 22.816)
1	0.2	13.532	0.980	17.265	17.418	18.466
		(13.278, 13.776)	(0.948, 1.021)	(16.876, 17.651)	(17.038, 17.798)	(17.953, 19.165)

Both orderings hold in every seed, layout split and deployment policy. The 18.1-nat magnitude is therefore specific to fully privileged collection: at $\lambda = 0.5$ the three-group exact distortion is below Public's 1.967 nats against do. The sweep computes exact targets only. With learned predictors of one recipe (Appendix C.6), the deployment gap rises with λ while its share of D_{marg} falls. The historical $\lambda = 0.5$ runs used a different recipe, layouts and scoring, and their target contrasts remain group-specific (Appendix E.3).

Within-history heterogeneity. A post hoc, exploratory analysis compares rocks within one history using the recorded exact $\lambda = 1$ references on the 320 deployment contexts of Table 4A (training, unseen-A and unseen-B layouts, four history-only policies, five seeds); it trains no model. At a valid frame, i.e. a public history h_t, let $k_j = D_{\mathrm{KL}}(p_j \| b_j)$ with $p_j = b_j^{\mathrm{do}}$ and $b_j = b_j^{\mu}$ for each unresolved rock j; with the weights of Table 4 these terms reproduce its $D_{\mathrm{marg}} = 18.083$ nats. Every valid frame in these contexts has at least two unresolved rocks. A history is split if one unresolved rock has $k_j < 10^{-6}$ nats and do entropy at least 0.05 nats while another has $k_j > 1$ nat. Averaging the fraction of split histories within each context with the layout, policy and split weights of Table 4, 40.4% of deployment histories are split (seed range 39.7–40.9%); the share is 33.7% with 10^{-10} in place of 10^{-6} and 46.2% with 0.1 nats in place of 1 nat. The remaining statistics pool the 836,767 training and unseen-B histories without weights. Within-history deviations of k_j from the history mean carry 57.5% of the sum of squared deviations from the pooled mean over (history, unresolved rock) pairs. Among unresolved rocks with $k_j < 10^{-6}$ or $k_j > 0.1$ nats, a rule that labels all rocks of a history alike, as distorted or undistorted, errs on at least the smaller group at each history, in total at least 12.0% of these rocks. In histories whose summed KL exceeds 10^{-6} nats, the rock with the largest k_j carries a median share of 0.39 of the summed KL under the three history-driven policies and 0.93 under the random policy. Figure 5b shows a history with this property from the historical Fresh archives.

C.3 Training fit and checkpoint selection

We inspect all 81 recorded whole-validation scores for each model and independently rescore the 80 retained selected/final validation-logit archives. The curves subtract the entropy computed on

the same validation frames with the same weighting. The raw loss is partitioned into publicly determined and unresolved contributions. Only the unresolved contribution enters the shared, public-clamped selection score. The recorded training minibatch losses remain cross-entropies, because their corresponding minibatch entropies were not logged. Complete-training-set own-target KL is instead recomputed at the selected and final endpoints from the frozen weights. These diagnostics are post hoc analyses of the completed training study.

At $\lambda = 1$, both observational-soft branches select update 250 in all five seeds. All-target do-soft selects update zero in every seed, whereas constrained do-soft selects updates 19,250–20,000. Every noninitial all-target do candidate has worse unresolved validation fit than its initialization. This choice is not caused by determined coordinates entering the validation objective: the common clamp removes their contribution. At the final budget, all-target do-soft has unresolved own-target KL 0.00010 on the complete training set but 0.300 on the selection-validation layouts, in nats per unresolved coordinate. The constrained final model reduces these values to 0.00007 and 0.00447, respectively. Thus low training error alone does not account for the intervention's validation improvement.

Observational-soft exhibits a different endpoint pattern. For the constrained branch, continuing from its selected checkpoint to the final budget reduces own-target KL on training histories from 0.164 to 0.012, and on independent seen-layout behavior histories from 0.168 to 0.059. At the same time, own-target KL rises from 0.215 to 0.705 on independent held-out behavior layouts. Selected models retain training approximation error, while continued training substantially worsens generalization to the held-out layouts. On the same fresh seen-layout behavior histories, against-do KL increases from 0.184 to 1.241 despite the own-target fit improvement. This last comparison is a common-reference score on behavior histories, whose Bayes posterior remains b^{μ}. The results distinguish fitting the behavior supervision target from approximating the do target without identifying a unique internal optimization mechanism.

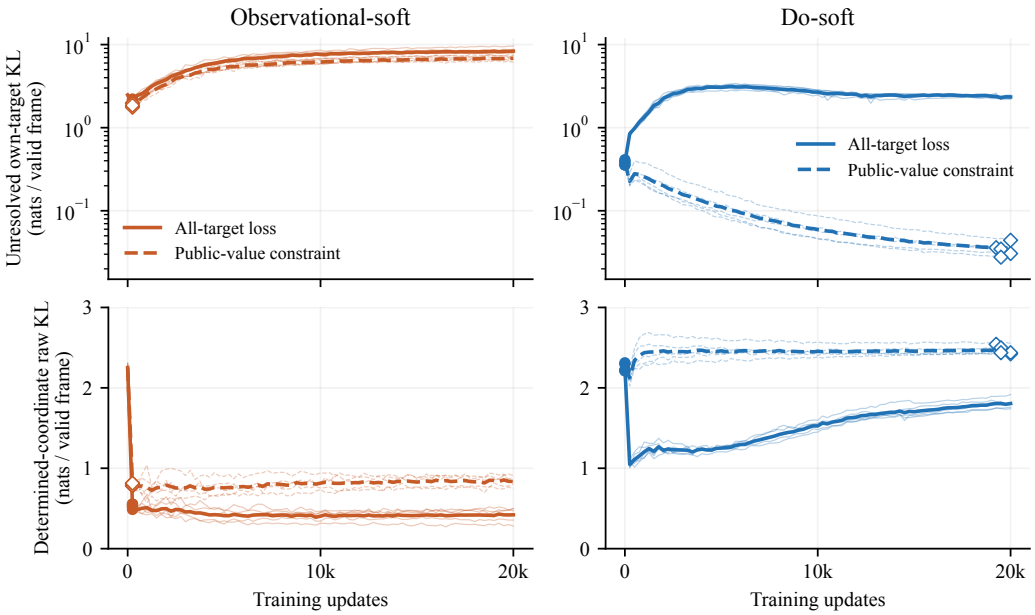


Figure 3: Validation fit and actual checkpoint selection at full mixing. Thin curves show all five seeds; thick curves show their means. Top: unresolved own-target KL (logarithmic axis). Bottom: determined-coordinate raw KL, excluded by common clamping. Both use nats per valid frame. Markers identify actual selections. The 81-point curves replay recorded whole-validation statistics; selected/final endpoints are additionally rescored from retained logits and weights. Validation participates in selection.

Table 15: Frozen-model own-target fit at selected and final checkpoints ($\lambda = 1$). Entries are nats per unresolved coordinate: pooled fixed-60 KL mass divided by the same pooled coordinate mass, over five seeds. All models use common public clamping. Training covers the complete training set; selection validation participates in checkpoint selection. Independent behavior columns use new trajectories on four held-out layouts or the eight training layouts. Deployment averages the four original policies equally. Unseen deployment pools its two four-layout sets. Full split results and exploratory paired-seed intervals accompany the numeric files.

Loss	Target	Checkpoint	Train μ	Select μ	Held-out μ	Seen μ	Seen π	Unseen π
All-target	Obs	Selected	0.174	0.267	0.230	0.177	0.294	0.317
All-target	Obs	Final	0.014	1.071	0.888	0.067	0.465	0.868
All-target	Do	Selected	0.045	0.049	0.044	0.045	0.237	0.236
All-target	Do	Final	0.00010	0.300	0.226	0.005	0.083	0.210
Constrained	Obs	Selected	0.164	0.241	0.215	0.168	0.283	0.313
Constrained	Obs	Final	0.012	0.878	0.705	0.059	0.349	0.700
Constrained	Do	Selected	0.00007	0.004	0.003	0.00068	0.013	0.022
Constrained	Do	Final	0.00007	0.004	0.003	0.00068	0.013	0.022

C.4 Observational fitting sensitivity and filter-free checkpoint selection

Two questions remain after Appendix C.1. First, observational-soft does not approach its own target at own-target-selected checkpoints in the forty-model study (own-target KL 0.168 versus 0.0007 nats per unresolved rock for do-soft on independent seen-layout behavior histories), so the size of the deployment gap in Table 2B mixes posterior distortion with fitting error. Second, own-target selection returned the initialization for all-target do-soft, and in the channels behavior-label validation selects the deployment-worse observational arm; a practitioner needs a selection stream that does not require the exact do filter. This study addresses both with new training runs on the retained histories, under a protocol frozen before training.

Design. At $\lambda = 1$ and the same five paired seeds, the same training and behavior-validation histories, the same initialization seed and the same minibatch seed as Appendix C.1, we train three optimization settings with the constrained loss and both supervision targets: the original recipe (128 units, learning rate 10^{-3}; also with the all-target loss), a wider network (256 units, 10^{-3}) and a smaller learning rate (128 units, 10^{-4}). Every validation candidate $0, 250, \ldots, 20{,}000$ is retained as weights. Two realized-label selection streams are recorded for every candidate, neither of which uses the do filter: *behavior-label* validation, the common-clamped realized-label NLL on the teacher-collected validation layouts, and *deployment-label* validation, the same score on the unseen-A history-only rollouts under the four deployment policies. Own-target checkpoint selection of Appendix C.1, which scores each arm against its own soft target, is also recomputed. Because unseen-A is consumed by selection here, this study's deployment measure pools the seen-layout and unseen-B deployment histories with equal weight; seen-layout and held-out behavior histories are scored as before. Intervals are descriptive paired five-seed 95% percentile bootstraps (10,000 draws) and are not adjusted across the tables. The initialization of the 128-unit runs reproduces the original parameter hash exactly; the installed PyTorch/CUDA build draws a different minibatch permutation from the same seed, so the 128-unit, 10^{-3} arm is an independent rerun of the original recipe rather than a bit-exact replication.

Own-target selection stops the observational arm at its first candidates in every setting. Table 16 reports the constrained arms. Own-target checkpoint selection stops observational-soft at update 250 in 13 of the 15 constrained setting–seed cases and at update 500 in the remaining 2 (10^{-4} learning-rate seeds). Behavior-label validation on the same histories stops it at update 250 in 12 pairs and at update 500 in 3, again all at the 10^{-4} learning rate. Choosing across arms, behavior-label validation returns observational-soft at update 250 in 11 pairs and at update 500 in 3, and do-soft at update 1,250 in the remaining pair. Among observational-soft candidates, deployment-label validation selects the initialization in all 15 (recorded selections in the core supplement). Observational against-do KL is 2.548, 2.871 and 2.556 nats against Public's 1.969 in the three settings, and do-soft's, E_{do}, is 0.151, 0.129 and 1.469. By equation 11, the expected deployment gap equals observational against-do KL minus E_{do}, so it follows both arms' against-do error rather than observational own-target fit. The smallest gap arises at the 10^{-4} learning rate, where do-soft is undertrained. At own-target-selected checkpoints, observational-soft own-target KL on independent seen-layout behavior histories is 0.168, 0.153 and 0.210 nats per unresolved rock for the 128-unit, 256-unit and 10^{-4} settings,

Table 16: Observational fitting under wider networks and smaller learning rates (constrained loss, $\lambda = 1$, five new paired seeds). Seen layouts are the training layouts; own-target-selected checkpoints minimize each arm's own-target validation loss. Own-target KL is per unresolved rock on independent seen-layout behavior histories. Realized NLL gaps are observational-soft minus do-soft; deployment pools seen-layout and unseen-B histories with equal weight under four history-only policies. Obs KL and Do KL are unresolved against-do loss masses, computed as in Table 2A on this study's deployment pool; the interval on the deployment gap is a paired five-seed 95% percentile bootstrap, and every seed value has the sign of its mean.

Setting	Checkpoint	Obs own	Do own	Gap μ	Gap π [95%]	Obs KL	Do KL
128 units, lr 10^{-3}	Own-target-selected	0.168	0.0007	-0.48	$+2.39\,(+2.31, +2.48)$	2.55	0.151
128 units, lr 10^{-3}	Final	0.059	0.0007	-1.02	$+11.68\,(+11.21, +12.09)$	11.88	0.150
256 units, lr 10^{-3}	Own-target-selected	0.153	0.0006	-0.56	$+2.72\,(+2.50, +2.94)$	2.87	0.129
256 units, lr 10^{-3}	Final	0.045	0.0006	-1.08	$+10.59\,(+10.42, +10.81)$	10.72	0.127
128 units, lr 10^{-4}	Own-target-selected	0.210	0.0272	-0.34	$+1.08\,(+0.81, +1.37)$	2.56	1.469
128 units, lr 10^{-4}	Final	0.092	0.0009	-0.85	$+7.22\,(+6.27, +8.02)$	7.69	0.451
Public reference	—	—	—	—	—	1.969	1.969

against 0.0007, 0.0006 and 0.027 for do-soft. The wider network helps little; the smaller learning rate hurts both arms within the fixed budget. Final checkpoints fit the training target better (0.059, 0.045 and 0.092) but generalize worse: the held-out behavior gap turns positive ($+2.09$, $+1.73$ and $+0.89$) and the deployment gap grows to $+11.7$, $+10.6$ and $+7.2$ nats, reproducing the overfitting pattern of Appendix C.3. The reversal itself holds in every setting and seed: at own-target-selected checkpoints the seen-layout behavior gaps are -0.48, -0.56 and -0.34, the held-out behavior gaps -0.23, -0.20 and -0.32, and the deployment gaps $+2.39$, $+2.72$ and $+1.08$ nats, with all fifteen seed values of each sign. The 128-unit rerun reproduces the original study's own-target-selected values to three decimals (own-target KL 0.168, behavior gap -0.480, held-out gap -0.228), so the values of Appendix C.1 are not artefacts of one minibatch order.

By equation 11, an observational predictor equal to b^{μ} on these histories would have expected deployment gap $D_{\mathrm{marg}} - E_{\mathrm{do}}$, where $D_{\mathrm{marg}} = 18.21$ nats on this pooled measure. Every expected gap G lies below this value because B_{obs} is negative on this measure in every seed (-15.67, -15.34 and -15.66 nats at own-target-selected checkpoints), whereas under the random policy alone fitting error inflates the gap (Table 4B). An exploratory reading of the recorded checkpoints gives a direction, not convergence: from the initialization to the own-target-selected checkpoint, observational E_{obs} on this pool falls while its against-do KL rises in all 15 setting–seed cases (3.852 to 2.510 and 1.990 to 2.548 nats at 128 units), and later training worsens this deployment-history fit in all 15 (final E_{obs} 4.528, 3.614 and 3.588 nats). We did not find a setting in which observational-soft approaches its target, so RockSample does not separate learnability from distortion, and we do not claim the gap magnitudes are semantic.

Table 17: Checkpoint and arm selection ($\lambda = 1$, five new paired seeds). Behavior-label validation scores realized labels on teacher-collected histories of the validation layouts; deployment-label validation scores them on history-only rollouts on the unseen-A layouts. The arm columns choose between arms by these two streams, and neither uses the do filter. Within do-soft, the update and KL columns compare own-target checkpoint selection (own), which scores the do-soft target and therefore needs the do filter, with deployment-label validation (dep.). Do chosen counts the seeds, out of five, in which each stream chooses do-soft; updates list the five selected do-soft candidates; deployment KL is the unresolved against-do loss mass of the chosen do-soft models on seen-layout and unseen-B deployment histories (equal weights), with paired five-seed 95% intervals.

Units, lr	Loss	Do chosen (beh. / dep.)	Do updates (own / dep.)	Do KL, own	Do KL, dep.
128, 10^{-3}	All-target	5/5 / 5/5	0 / 750–13,000	$1.99\,(1.97, 2.00)$	$1.28\,(1.02, 1.53)$
128, 10^{-3}	Constrained	0/5 / 5/5	19,250–20,000 / 19,250–20,000	$0.15\,(0.14, 0.16)$	$0.15\,(0.14, 0.16)$
256, 10^{-3}	Constrained	1/5 / 5/5	18,750–20,000 / 19,750–20,000	$0.13\,(0.12, 0.13)$	$0.13\,(0.12, 0.13)$
128, 10^{-4}	Constrained	0/5 / 5/5	500–2,250 / 12,750–19,750	$1.47\,(1.21, 1.72)$	$0.45\,(0.41, 0.50)$

Deployment-label validation picks do-soft and does no better than the last checkpoint. Table 17 compares the two realized-label streams for the arm choice and, within do-soft, deployment-label

validation with own-target selection. With the constrained loss, choosing between arms by behavior-label validation selects observational-soft in 14 of 15 setting–seed cases (the exception is one 256-unit seed), whereas deployment-label validation selects do-soft in all 15; with the all-target loss both streams select do-soft in every seed, because the all-target observational arm is poor on both history laws. Within the do-soft arm, own-target and deployment-label selection give almost the same deployment KL in the two constrained 10^{-3} settings (0.151 and 0.150 at 128 units, 0.129 and 0.128 at 256 units), although their checkpoints coincide in only 3 of five seeds at 128 units and 0 of five at 256 units. The two rules differ where own-target selection fails: for the all-target loss it again returns the initialization in every seed, whereas deployment-label validation selects updates 750–$13{,}000$ and lowers deployment KL from 1.99 to 1.28 $(1.02, 1.53)$; for the 10^{-4} learning rate it selects updates 500–$2{,}250$ with KL 1.47, whereas deployment-label validation selects $12{,}750$–$19{,}750$ with KL 0.45. For observational-soft, deployment-label validation selects the initialization in all 15 cases: no trained observational checkpoint improves on an untrained network under deployment histories, which restates the reversal as a selection outcome. The distortion alone predicts this outcome, so it does not show that the behavior posterior is hard to learn: on the pooled deployment measure the initialization's against-do KL is 1.990 nats at 128 units, whereas the exact behavior posterior's is $D_{\mathrm{marg}} = 18.21$, so an observational predictor equal to b^{μ} would also lose to the initialization. As in the channels, always choosing do-soft gives the same arm here: the observational arm's best deployment-label candidate is the initialization it shares with do-soft, so do-soft's best candidate never scores worse, and the driver (`r30_fit_selection.py`) breaks ties in favor of do-soft. Relative to own-target selection, the stream therefore changes do-soft deployment KL only in the all-target and 10^{-4} settings. The last checkpoint does as well without any labels. Its deployment KL is 1.21 under the all-target loss and 0.45 at the 10^{-4} learning rate, and its paired difference from the deployment-label-selected checkpoint is -0.069 $(-0.272, +0.128)$ and -0.001 $(-0.003, +0.002)$ nats. In the two constrained 10^{-3} settings the two means differ by at most 0.0003 nats. Within do-soft, the stream thus improves on own-target selection, not on stopping at the end of the fixed 20,000-update budget. The stream requires history-only rollouts with privileged labels, which a simulator provides at collection time. The rollouts use the four fixed deployment policies on the unseen-A layouts, with 64 episodes per layout and policy, 1,024 labeled history-only episodes per seed against 2,048 training and 256 behavior-validation episodes, and the deployment evaluation uses the same four policies on the seen and unseen-B layouts; three of the four policies act on exact do beliefs, so "no filter" refers to the labels and the scoring, not to running the rollout policy. The rollout policies are not the student itself; an adaptive version and the cost of on-policy collection are untested.

Execution and retained evidence. The forty new models (five seeds, three settings, two arms, two losses for the 128-unit 10^{-3} setting) trained in 1,180 seconds on one RTX PRO 6000 Blackwell GPU with PyTorch 2.10.0; evaluation of the 120 selected checkpoints on all seen, held-out, seen-deployment and unseen-B contexts followed. The core supplement includes the driver (`r30_fit_selection.py`), the table script, the frozen protocol, all candidate curves with both selection streams, the selections and the context-level scores; the 3,240 candidate weight archives (1.4 GiB) are retained locally and are regenerated by the documented command.

Hard-label arm. An extension adds the hard-label DSS arm to this study under a second protocol, fixed before any of its training. At $\lambda = 1$ the arm fits the realized rock qualities of the same training logs in all three settings, with the same five seeds, initialization, minibatch order, 20,000-update budget, candidate grid and selection rules. The constrained loss drops the same publicly determined coordinates, where the realized label equals the do target. The extension retrains observational-soft and do-soft alongside it: 2,430 of 2,430 candidate weight archives and 3,420 of 3,420 context-score rows are identical to those behind Table 16, so the three arms are paired within each seed. The hard-label arm's own target is the realized label, so its own-target selection is behavior-label validation and uses no posterior filter. A $\lambda = 0$ control trains do-soft and hard-label arms in the default setting on the history-only logs, where observational-soft coincides with do-soft and D_{marg} is zero, so at own-target-selected checkpoints it measures the cost of the hard labels' noise together with the difference in training length between the two arms' selected checkpoints. The extension's driver (`r34_hard_label.py`) derives from `r30_fit_selection.py`, which it leaves unchanged; its protocol, run log, candidate curves, selections, context-level scores and estimand summaries are recorded in the same form as this study's.

Table 18: Hard-label DSS under the constrained loss (five paired seeds per row). Entries are realized-label NLL differences (nats) on fresh behavior histories of the training layouts (Training), on behavior histories of the held-out behavior layouts (Held-out) and on the deployment pool, which weights deployment histories on the training and unseen-B layouts equally under four history-only policies. Block headers give the mixture, hidden units and learning rate. At $\lambda = 1$ the observational-soft and do-soft models are bit-identical replicates of those in Table 16 (every candidate weight, validation score and context score), and the hard-label arm shares their initialization and minibatch order. For the hard-label arm, own-target checkpoint selection is behavior-label validation and uses no filter; it selects updates 250, 250 and 250–750 in the three $\lambda = 1$ settings and 500–750 at $\lambda = 0$. Final is update 20,000. At $\lambda = 0$ the behavior policy is history-only and observational-soft coincides with do-soft. Each cell gives the seed mean with the number of seeds whose difference has its sign, and below it the paired five-seed 95% percentile bootstrap interval. Values of magnitude 10 or more are rounded to one decimal; three decimals are used where two would round a value to zero or print both interval endpoints alike.

Checkpoint	Training	Held-out	Deployment
A. Hard-label − do-soft			
$\lambda = 1$, 128 units, learning rate 10^{-3}			
Own-target	-0.41 (5/5)	-0.24 (5/5)	$+2.65$ (5/5)
	$(-0.43, -0.38)$	$(-0.27, -0.21)$	$(+2.44, +2.85)$
Final	$+10.5$ (5/5)	$+12.7$ (5/5)	$+32.5$ (5/5)
	$(+10.2, +10.7)$	$(+12.1, +13.5)$	$(+31.8, +33.5)$
$\lambda = 1$, 256 units, learning rate 10^{-3}			
Own-target	-0.47 (5/5)	-0.23 (5/5)	$+2.73$ (5/5)
	$(-0.50, -0.45)$	$(-0.28, -0.19)$	$(+2.55, +2.96)$
Final	$+16.7$ (5/5)	$+18.9$ (5/5)	$+43.3$ (5/5)
	$(+16.4, +17.0)$	$(+18.1, +19.7)$	$(+42.6, +44.3)$
$\lambda = 1$, 128 units, learning rate 10^{-4}			
Own-target	-0.34 (5/5)	-0.33 (5/5)	$+1.13$ (5/5)
	$(-0.37, -0.32)$	$(-0.36, -0.29)$	$(+0.80, +1.40)$
Final	-0.49 (5/5)	$+0.79$ (5/5)	$+5.95$ (5/5)
	$(-0.54, -0.44)$	$(+0.55, +1.05)$	$(+5.32, +6.55)$
$\lambda = 0$ (history-only behavior policy), 128 units, learning rate 10^{-3}			
Own-target	$+0.35$ (5/5)	$+0.30$ (5/5)	$+1.00$ (5/5)
	$(+0.33, +0.36)$	$(+0.28, +0.31)$	$(+0.92, +1.10)$
Final	$+22.6$ (5/5)	$+22.3$ (5/5)	$+23.0$ (5/5)
	$(+22.2, +23.0)$	$(+21.8, +22.7)$	$(+22.7, +23.4)$
B. Hard-label − observational-soft			
$\lambda = 1$, 128 units, learning rate 10^{-3}			
Own-target	$+0.07$ (5/5)	-0.01 (4/5)	$+0.26$ (5/5)
	$(+0.069, +0.074)$	$(-0.027, -0.002)$	$(+0.09, +0.45)$
Final	$+11.5$ (5/5)	$+10.6$ (5/5)	$+20.8$ (5/5)
	$(+11.3, +11.7)$	$(+10.0, +11.3)$	$(+20.1, +21.6)$
$\lambda = 1$, 256 units, learning rate 10^{-3}			
Own-target	$+0.08$ (5/5)	-0.03 (5/5)	$+0.005$ (4/5)
	$(+0.07, +0.10)$	$(-0.05, -0.02)$	$(-0.22, +0.15)$
Final	$+17.8$ (5/5)	$+17.2$ (5/5)	$+32.7$ (5/5)
	$(+17.5, +18.1)$	$(+16.5, +17.9)$	$(+32.1, +33.8)$
$\lambda = 1$, 128 units, learning rate 10^{-4}			
Own-target	-0.01 (3/5)	-0.002 (3/5)	$+0.05$ (2/5)
	$(-0.03, +0.01)$	$(-0.02, +0.01)$	$(-0.05, +0.15)$
Final	$+0.36$ (5/5)	-0.10 (4/5)	-1.27 (5/5)
	$(+0.33, +0.39)$	$(-0.41, +0.28)$	$(-1.49, -0.98)$

Hard labels inherit the reversal at early checkpoints. Table 18 reports the contrasts. At $\lambda = 1$ and own-target-selected checkpoints, hard-label minus do-soft realized NLL lies between -0.474 and -0.343 nats on training-layout behavior histories, between -0.326 and -0.232 on held-out behavior histories and between $+1.126$ and $+2.727$ on the deployment pool across the three settings. Each sign holds in all 15 setting–seed cases, and no interval includes zero. The hard-label arm is selected early: it stops at update 250 in both 10^{-3} settings and at 250–750 at 10^{-4}, never before observational-soft (250–500 at 10^{-4}), whereas do-soft is selected at 18,750–20,000 and, at 10^{-4}, 500–2,250. Final hard-label checkpoints overfit: they trail do-soft on held-out behavior histories in all 15 pairs and at deployment by $+5.951$ to $+43.342$ nats, and keep a training-layout advantage only at the 10^{-4} learning rate.

At own-target-selected checkpoints, hard labels bring no gain over observational-soft on the deployment pool: hard-label minus observational-soft there is $+0.260$ $(+0.094, +0.451)$ nats in the default setting, positive in every seed, and its intervals include zero in the other two settings. In the 256-unit setting the pool averages policies with opposite signs: with the two layout sets weighted equally, hard labels do better than observational-soft under each history-driven policy, where the difference is -0.233 to -0.193 nats and negative in every seed, and worse under the random policy, where it is $+0.661$ $(+0.299, +0.944)$ and positive in 4 of five seeds. The per-layout values are the differences of the two $\lambda = 1$ blocks of Table 19. At the final checkpoint of the 10^{-4} setting, by contrast, the pooled difference is -1.272 $(-1.488, -0.980)$ nats, negative in every seed, whereas under the random policy alone it is $+0.937$ $(+0.713, +1.121)$. On behavior histories the two arms differ by -0.032 to $+0.084$ nats. In the default setting the hard-label arm's against-do KL on the deployment pool is 2.802 nats, against 2.548 for observational-soft and 0.151 for do-soft.

Without distortion the hard labels cost accuracy. At $\lambda = 0$, hard-label minus do-soft is $+0.347$, $+0.296$ and $+0.998$ nats on the three history sets, positive in every seed, with the hard-label arm selected at updates 500–750 and do-soft at 19,000–20,000. On the shared deployment contexts of the default setting, $\lambda = 0$ hard-label minus $\lambda = 1$ do-soft is $+0.940$ nats and $\lambda = 0$ hard-label minus $\lambda = 1$ observational-soft is -1.448, and $\lambda = 0$ do-soft minus $\lambda = 1$ do-soft is -0.058, each with its sign in all five seeds. Unlike in the historical study (Appendix E.3), $\lambda = 0$ hard-label training is worse here than $\lambda = 1$ do-soft. These cross-mixture contrasts change the collection mixture, the first two also the label form, and none estimates distortion.

With the hard-label arm as a third candidate, behavior-label validation chooses it in 11 of the 15 $\lambda = 1$ cases and observational-soft in the other 4, never do-soft, whereas deployment-label validation chooses do-soft in all 15. As for observational-soft, the hard-label arm's best deployment-label candidate is its initialization in every $\lambda = 1$ case: even its best trained checkpoint scores 0.263–1.072 nats worse than the initialization on the unseen-A rollouts, whereas at $\lambda = 0$ its deployment-label-selected checkpoint improves on the initialization by +0.930 nats, in every seed. At $\lambda = 0$ both streams choose do-soft in every seed. These three-arm counts are selection outcomes on the same models and are reported separately from the two-arm counts of Table 17, which the retrained arms reproduce (14 and 15 of 15).

Behavior-label validation restricted to later checkpoints. Every non-do choice of behavior-label validation at $\lambda = 1$ lies at update 250 or 500, for two and for three arms. After all runs were complete we therefore applied the same rule post hoc, with ties going to do-soft, then observational-soft, then hard-label and then the earliest update, to every arm's recorded validation curve restricted to updates of at least T; `r38_late_selection.py` computes it without training or inference, and $T = 0$ reproduces the recorded choices above. With $T = 2{,}000$ the rule picks do-soft in 10 of the 15 $\lambda = 1$ cases with two arms and in 10 with three. The other five are the five seeds of the 10^{-4} learning rate, where it picks observational-soft or DSS at updates 2,000–2,750, just after the cutoff. On the unseen-A rollouts, realized NLL of these picks minus that of the do-soft checkpoint the rule picks within do-soft is $+1.286$ and $+1.207$ nats for two and three arms, positive in all five seeds. With $T = 10{,}000$ it picks do-soft in 15 of 15 cases with two arms and 14 with three, the exception a DSS checkpoint at update 10,000. At $\lambda \in \{0.25, 0.5, 0.75\}$ and in the $\lambda = 0$ control it picks do-soft at every T. The deployment pool is not recorded per candidate, so only the unseen-A rollouts of deployment-label validation score the picked checkpoints. Within these runs, the selection failure at $\lambda = 1$ is concentrated at early checkpoints.

By deployment policy and layout set. Table 19 splits the deployment entries of Tables 16 and 18 at own-target-selected checkpoints by history-only policy and layout set. This split (`r35_policy_breakdown.py`) is post hoc and descriptive, since the frozen protocol pools the four policies.

Table 19: Deployment contrasts of Tables 16 and 18 by history-only policy and layout set at own-target-selected checkpoints (post hoc, descriptive; realized-label NLL differences in nats). Setting gives hidden units and learning rate. A cell averages, within each seed, the layouts of one policy (eight training, four unseen-B); the equal-weight average of the eight cells of a setting is its deployment entry in Table 18 or, for observational-soft − do-soft, its Gap π entry in Table 16. Each cell gives the five-seed mean with the number of seeds sharing its sign and, below it, the paired five-seed 95% percentile bootstrap interval. The top rows give the exact distortion D_{marg} of each policy and layout set at $\lambda = 1$ (seed mean), which depends on the histories only; at $\lambda = 0$ it is zero and observational-soft coincides with do-soft.

Setting	Layouts	Random	History-greedy	History-cautious	Different-seed
Exact distortion D_{marg}, $\lambda = 1$, all settings					
	Training	1.59	24.3	24.5	24.4
	Unseen-B	1.66	21.0	21.4	26.8
Hard-label − do-soft, $\lambda = 1$					
128, 10^{-3}	Training	+4.81 (5/5) (+4.57, +5.09)	+2.01 (5/5) (+1.84, +2.18)	+2.12 (5/5) (+1.98, +2.24)	+2.00 (5/5) (+1.86, +2.15)
	Unseen-B	+4.37 (5/5) (+3.98, +4.76)	+1.91 (5/5) (+1.65, +2.17)	+2.05 (5/5) (+1.86, +2.30)	+1.90 (5/5) (+1.70, +2.10)
256, 10^{-3}	Training	+4.76 (5/5) (+4.39, +5.17)	+2.17 (5/5) (+2.00, +2.33)	+2.29 (5/5) (+2.18, +2.39)	+2.14 (5/5) (+2.00, +2.30)
	Unseen-B	+4.31 (5/5) (+3.99, +4.75)	+2.00 (5/5) (+1.80, +2.25)	+2.15 (5/5) (+2.01, +2.30)	+2.00 (5/5) (+1.82, +2.22)
128, 10^{-4}	Training	+2.23 (5/5) (+1.73, +2.74)	+0.87 (5/5) (+0.52, +1.12)	+0.91 (5/5) (+0.59, +1.17)	+0.84 (5/5) (+0.54, +1.10)
	Unseen-B	+2.01 (5/5) (+1.65, +2.36)	+0.68 (5/5) (+0.40, +0.87)	+0.75 (5/5) (+0.45, +1.04)	+0.71 (5/5) (+0.45, +0.93)
Observational-soft − do-soft, $\lambda = 1$					
128, 10^{-3}	Training	+3.98 (5/5) (+3.82, +4.13)	+2.05 (5/5) (+2.01, +2.10)	+2.15 (5/5) (+2.12, +2.18)	+1.96 (5/5) (+1.85, +2.08)
	Unseen-B	+3.54 (5/5) (+3.37, +3.76)	+1.78 (5/5) (+1.59, +1.97)	+1.90 (5/5) (+1.75, +2.02)	+1.74 (5/5) (+1.62, +1.86)
256, 10^{-3}	Training	+4.13 (5/5) (+3.92, +4.36)	+2.43 (5/5) (+2.22, +2.65)	+2.54 (5/5) (+2.40, +2.68)	+2.35 (5/5) (+2.20, +2.53)
	Unseen-B	+3.63 (5/5) (+3.37, +3.89)	+2.20 (5/5) (+1.89, +2.54)	+2.33 (5/5) (+2.07, +2.58)	+2.17 (5/5) (+1.88, +2.53)
128, 10^{-4}	Training	+2.26 (5/5) (+1.83, +2.75)	+0.81 (5/5) (+0.52, +1.10)	+0.85 (5/5) (+0.55, +1.16)	+0.77 (5/5) (+0.49, +1.06)
	Unseen-B	+2.02 (5/5) (+1.70, +2.34)	+0.61 (5/5) (+0.38, +0.78)	+0.67 (5/5) (+0.41, +0.95)	+0.63 (5/5) (+0.43, +0.84)
Hard-label − do-soft, $\lambda = 0$ (history-only behavior policy)					
128, 10^{-3}	Training	+3.01 (5/5) (+2.68, +3.44)	+0.26 (5/5) (+0.25, +0.27)	+0.29 (5/5) (+0.28, +0.30)	+0.64 (5/5) (+0.60, +0.68)
	Unseen-B	+2.74 (5/5) (+2.45, +3.11)	+0.24 (5/5) (+0.18, +0.29)	+0.25 (5/5) (+0.21, +0.29)	+0.55 (5/5) (+0.47, +0.62)

Hard-label minus do-soft is positive in every seed in all 24 setting–layout–policy cells, from $+0.679$ to $+4.813$ nats, with every interval above zero, and each unseen-B value lies below the training-layout value of its setting and policy. The gap is largest under the random policy in every setting and layout set: at learning rate 10^{-3} it is $+4.313$ to $+4.813$ nats, against $+1.895$ to $+2.287$ under the history-driven policies. Yet the random policy has the smallest exact distortion, $D_{\mathrm{marg}} = 1.59$ and 1.66 nats on the two layout sets against 21.0 to 26.8, and the $\lambda = 0$ control, which has no distortion, is also largest under that policy ($+2.741$ and $+3.014$ nats against $+0.243$ to $+0.638$). The random-policy excess therefore reflects fitting error, as in the random-policy exception of Section 5.5. Observational-soft minus do-soft shows the same pattern, from $+0.605$ to $+4.130$ nats and largest under the random policy in every row.

Decomposition of the hard-label gap. Hard labels have the population target b^{μ} of observational-soft, so equation 11 holds with the hard-label arm in the observational slot; B_{hard} denotes its signed

fitting term and G its expected gap to do-soft. We evaluate the terms on the context scores behind Table 19, with its pooling (`r36_dss_decomposition.py`; post hoc and descriptive), at $\lambda = 1$ and own-target-selected checkpoints. In the 18 history-driven setting–layout–policy cells, D_{marg} ranges from 21.034 to 26.776 nats, but B_{hard} ranges from -25.038 to -18.898 and is negative in every seed in 18 of them, so G stays far below D_{marg}, at 0.700 to 2.261. In the six random-policy cells, D_{marg} is only 1.587 to 1.665, whereas B_{hard} is 2.960 to 3.754 and positive in every seed in 6 of them; G (2.005 to 4.820) exceeds D_{marg} in 6 of the six, as for observational-soft in Table 4B. On the deployment pool of the default setting, $D_{\mathrm{marg}} = 18.213$, $B_{\mathrm{hard}} = -15.411$, negative in 5 of five seeds, and $G = 2.652$ nats. The $\lambda = 0$ control, where $b^{\mu} = b^{\mathrm{do}}$, so that $D_{\mathrm{marg}} = 0$ and B_{hard} is the hard-label arm's KL error against b^{do}, shows the random-policy excess without distortion: B_{hard} is 3.057 under the random policy, with the two layout sets weighted equally, against 0.276 to 0.669 in the six history-driven cells. The hard-label gap thus keeps its sign in every cell, while its size follows the fitting term rather than D_{marg}.

C.5 More training layouts at full privilege

At $\lambda = 1$ observational-soft stops near initialization and fits its own target poorly on held-out layouts, so its gap mixes exact distortion with fitting error. This study asks whether more training layouts improve that fit and whether a better fit moves the learned gap G toward D_{marg}. It was designed after the eight-layout results were known and is therefore post hoc. Its protocol, conditions, estimands and decision rules were fixed before any new history was collected, and it is run by `r39_more_layouts.py` and tabulated by `r39_more_layouts_tables.py`.

Design. All conditions use the default constrained recipe of Appendix C.4 at $\lambda = 1$: 128 units, learning rate 10^{-3}, 20,000 updates, candidates every 250 updates, the three arms observational-soft, do-soft and DSS, and the initializations of the recorded seeds 201–205. The recorded study trains on 8 layouts with 256 teacher episodes each. The primary condition uses 64 layouts with 256 episodes each, and a data control uses the original 8 layouts with 2,048 episodes each, the same 16,384 episodes in total. Two secondary conditions use 64 layouts with 32 episodes each, matching the recorded total, and 256 layouts with 64 episodes each. The 248 new layouts are pairwise distinct, differ from the 56 layouts used elsewhere in this work and are nested across conditions. The original eight layouts and their recorded histories are kept. Validation, held-out behavior and deployment histories and their references are the recorded ones, so D_{marg} is the same in every condition. The primary contrasts, paired by seed against the recorded study, are observational-soft's own-target KL per unresolved rock on the held-out behavior layouts and G/D_{marg} on the deployment pool, and both validation streams' picks are the third estimand. The fit counts as improved if the first contrast's interval lies below zero and all five seeds are lower. The gap counts as approaching the distortion if the second contrast's interval lies above zero with all five seeds higher and the pooled E_{obs} lower in at least four seeds. The selection reversal is that of Appendix C.6, and a late reversal also requires every non-do behavior-label pick to fall at or after update 2,000. Before training, all recorded training archives were regenerated bit for bit, propensity and public-value errors were zero, recomputed references matched the stored ones to floating-point rounding, the per-context D_{marg} equalled the recorded one exactly, and retraining recorded seed 201 reproduced every row of its candidate curve. Driver changes after the freeze affected only the engineering smoke test or split training into one process per seed, and left the design unchanged.

Fit improves but the gap does not approach the distortion. Table 20 reports every condition. The reversal holds in every seed of every condition: observational-soft wins on teacher histories of the training and held-out layouts and loses on the deployment pool, where its realized-label gap is $+2.518$ nats ($+2.178$, $+2.902$) with 64 layouts. With 64 layouts, observational-soft's held-out own-target KL falls from 0.215 to 0.158 nats per unresolved rock, a paired change of -0.057 (-0.060, -0.054) in all five seeds, and its pooled E_{obs} falls from 2.510 to 1.473. The fit rule is met. Yet G stays at 13.8% of D_{marg} against 13.2%, a per-seed change of $+0.7$ points (-0.9, $+2.7$) that is positive in only 3 of 5 seeds, because B_{obs} stays at -15.605 nats. The gap rule is not met. Eight times as many episodes on the original layouts change neither quantity: the held-out KL changes by -0.0005 and G/D_{marg} by $+0.1$ points. With 32 episodes on each of 64 layouts the fit also improves, but G/D_{marg} moves away from the distortion, by -4.3 points (-5.0, -3.7) in all five seeds. With 256 layouts the held-out KL falls to 0.131 and the per-seed ratio rises by $+2.0$ points (-0.5, $+4.6$) in 4 of 5 seeds, which also misses the rule. The negative B_{obs} is not confined to deployment histories: on teacher histories of the

Table 20: More training layouts at $\lambda = 1$ (constrained loss, 128 units, learning rate 10^{-3}, five seeds, 20,000 updates). The study was designed and its protocol fixed after the eight-layout results were known. Recorded: the eight-layout study. Observational-soft (obs) against do-soft (do); own-target-selected checkpoints unless marked final. Seed means; lower panel: paired change, seeds in its direction and paired five-seed 95% percentile bootstrap interval. Picks: counts over the three arms and the range of chosen updates. $D_{\mathrm{marg}} = 18.213$ nats on the deployment pool and 12.400 on the held-out behavior histories for every column; G/D_{marg} is a ratio of seed means, and its paired change uses per-seed ratios.

Layouts × episodes	Recorded 8 × 256	More data 8 × 2,048	More layouts 64 × 256	Diversity 64 × 32	Most layouts 256 × 64
Own-target KL of obs per unresolved rock					
held-out behavior layouts	0.215	0.215	0.158	0.188	0.131
held-out, final checkpoint	0.704	0.569	0.240	0.377	0.152
unseen-B deployment histories	0.313	0.311	0.188	0.246	0.165
do-soft, held-out layouts	0.003	0.002	0.001	0.001	0.001
Deployment pool					
E_{obs}	2.510	2.493	1.473	2.003	1.383
E_{do}	0.151	0.084	0.092	0.170	0.098
B_{obs}	-15.666	-15.718	-15.605	-16.436	-15.356
G	$+2.397$	$+2.412$	$+2.516$	$+1.607$	$+2.759$
G/D_{marg} (%)	13.2	13.2	13.8	8.8	15.1
realized-label NLL gap	$+2.388$	$+2.407$	$+2.518$	$+1.608$	$+2.765$
G/D_{marg} (%), final	64.4	55.4	39.6	38.6	30.7
Behavior histories (obs against do)					
NLL gap, training layouts (fresh)	-0.480	-0.481	-0.724	-0.453	-0.707
NLL gap, held-out layouts	-0.228	-0.226	-0.489	-0.342	-0.618
B_{obs}, held-out layouts	-11.464	-11.426	-11.090	-11.651	-11.058
G/D_{marg} (%), held-out layouts	7.4	7.8	10.5	6.0	10.8
Selection					
obs own-target update	250	250	2,000–3,500	500–750	3,500–6,000
behavior-label pick obs/do/DSS	1/0/4	1/0/4	5/0/0	5/0/0	5/0/0
at update	250	250	2,000–2,750	500–750	3,500–6,000
deployment-label pick obs/do/DSS	0/5/0	0/5/0	0/5/0	0/5/0	0/5/0
at update	19,250–20,000	17,000–20,000	9,500–13,250	3,750–6,000	7,750–12,000
Paired change against the recorded study (seeds in the direction of the mean)					
held-out own-target KL per rock	–	-0.0005 (2/5) $[-0.003, +0.002]$	-0.057 (5/5) $[-0.060, -0.054]$	-0.027 (5/5) $[-0.031, -0.024]$	-0.084 (5/5) $[-0.088, -0.081]$
pool E_{obs}	–	-0.018 (3/5) $[-0.041, +0.006]$	-1.038 (5/5) $[-1.063, -1.004]$	-0.508 (5/5) $[-0.589, -0.404]$	-1.127 (5/5) $[-1.201, -1.056]$
pool G/D_{marg} (points)	–	$+0.1$ (3/5) $[-0.4, +0.5]$	$+0.7$ (3/5) $[-0.9, +2.7]$	-4.3 (5/5) $[-5.0, -3.7]$	$+2.0$ (4/5) $[-0.5, +4.6]$

held-out layouts, which involve no policy swap, it moves only from -11.464 to -11.090 nats with 64 layouts. A better fit in the b^{μ}-weighted sense of E_{obs} therefore does not make observational-soft reproduce the log-probabilities of b^{μ} where the do posterior places its mass. More layouts also curb overfitting: at the final checkpoint the held-out KL falls from 0.704 to 0.240, and the final-checkpoint G/D_{marg} from 64.4% to 39.6%.

Selection. With either quantity of data on the original layouts, own-target selection stops observational-soft at update 250. With 64 layouts it stops at updates 2,000–3,500, and behavior-label validation across the three arms picks observational-soft in 5 of 5 seeds at updates 2,000–2,750, while deployment-label validation picks do-soft in 5 of 5 at updates 9,500–13,250. The selection reversal therefore persists and meets the late-reversal rule, although the earliest pick lies exactly at its threshold. With 256 layouts the picks are at updates 3,500–6,000 and 7,750–12,000. The planned extension to 40,000 updates was not triggered, because no condition selected observational-soft at update 20,000.

C.6 Learned predictors at intermediate teacher mixtures

This study retrains the three constrained-loss arms of the hard-label study (observational-soft, do-soft and hard-label DSS; 128 units, learning rate 10^{-3}, 20,000 updates) at teacher mixture $\lambda = 0.5$ (primary) and $\lambda \in \{0.25, 0.75\}$ (secondary), with new behavior logs from the same collector and the seeds, initialization, minibatch order, deployment pool, selection rules and paired bootstrap of Appendix C.4, so that every seed pairs with the recorded $\lambda \in \{0, 1\}$ runs. The protocol was fixed after the $\lambda \in \{0, 1\}$ results and the exact sweep were known and before any intermediate-mixture log was collected. It prespecified two outcomes: a deployment gap, detected when observational-soft minus do-soft realized-label NLL on the deployment pool is positive in all five seeds with a 95% interval

excluding zero, and a selection reversal, in which behavior-label validation picks a non-do arm in at least four seeds while deployment-label validation picks do-soft in at least four. A one-line change to the driver after the freeze and before any log was collected restricts a post-training consistency check to the study's seeds and changes no reported value. The $\lambda = 0.75$ run was interrupted once and rerun from scratch with unchanged inputs, driver and protocol. The files the two runs share are byte-identical, and the tables report the completed rerun.

The deployment gap is detected at all three mixtures (Table 21A). At $\lambda = 0.25$, 0.5 and 0.75 it is $+0.498$, $+0.731$ and $+1.652$ nats of realized NLL, against $+2.388$ nats at $\lambda = 1$. As in Table 4C, the marginal gap G is 95.3, 56.0, 48.0 and 13.2% of exact D_{marg} at the four mixtures. There is no selection reversal at any intermediate mixture: behavior-label and deployment-label validation both pick do-soft in every seed (Table 21C). The behavior side of the reversal appears only at high mixing. On training-layout behavior histories observational-soft is worse than do-soft at $\lambda = 0.25$ ($+0.100$ nats, all five seeds), indistinguishable from it at 0.5 (-0.016, interval -0.038 to $+0.004$) and better at 0.75 (-0.299, all five seeds), yet behavior-label validation, which scores the separate validation layouts, still picks do-soft there. The hard-label increment over $\lambda = 0$ is null at 0.25 ($+0.002$, interval -0.086 to $+0.089$) and positive at 0.5 ($+0.240$, four of five seeds) and 0.75 ($+0.814$, all five). The increment is descriptive, not a distortion estimate, and stays well below the exact distortion: on the same pool it is 0.3, 18.2 and 23.6% of D_{marg} at these mixtures, against 9.1% at $\lambda = 1$, and at $\lambda = 0.25$ its interval spans zero although D_{marg} is 0.527 nats. The pool weights the training and unseen-B layouts equally and excludes unseen-A, so its D_{marg} differs from the three-group values of Table 14. The evidence is one setting with five seeds per mixture and unadjusted intervals.

C.7 Per-target action information and the certificate in RockSample

Corollary 2 and the finite examples of Table 5 show that per-target action information and distortion can separate. This post hoc, exploratory analysis measures how closely they track each other in RockSample. Thresholds, weights, cell sets and outcome rules were fixed before the full computation, after an exploratory probe on the cells of eight of the deployment contexts used below. On the 320 deployment contexts of Table 14 at $\lambda \in \{0.5, 1\}$, it evaluates every unresolved target of a valid frame, the cells of D_{marg}, with the exact per-target KL and the cumulative per-target action information $\mathrm{CAI}_k(h_t) = \sum_{s<t} I_\mu(A_s; Z_k \mid H_s = h_s)$ under the exact behavior posterior, a history-wise counterpart of ι^{act} and ι^{fwd} in equation 53, which coincide here because an unsampled rock keeps its quality. It collects no trajectory and trains no model.

At both mixtures a zero CAI is a near-exact sufficient check for no distortion (Table 22): 99.992% ($\lambda = 1$) and 99.999% ($\lambda = 0.5$) of its target mass is undistorted, and none of its cells has KL above 0.1 nats. It recovers fewer undistorted targets than the certificate of Proposition 5 on the same cells and tolerance, 41.1% against 73.9% at $\lambda = 1$ (difference -32.8 points, 95% interval -33.1 to -32.4) and 47.2% against 83.7% at $\lambda = 0.5$ (-36.6 points, -36.9 to -36.1). Only 24.7% ($\lambda = 1$) and 21.1% ($\lambda = 0.5$) of the unresolved target mass is undistorted, so the certificate clears 18.2% (17.9 to 18.5) and 17.6% (17.3 to 18.0) of that mass. The certificate covers targets that the teacher's actions do inform: 45.3% and 44.1% of certified mass has positive CAI, up to 21.0 nats at $\lambda = 1$. CAI and per-target distortion are positively but imperfectly monotone (Spearman 0.68 and 0.80), and of targets with CAI above 0.1 nats, 4.3% and 2.3% are undistorted. Soundness depends on the zero threshold of 10^{-20} nats. At 10^{-10} nats it is 97.51% and 99.28%, and at 10^{-6} nats 92.34% and 90.28%, below the 99% level of a near-exact check fixed in advance except at $\lambda = 0.5$ with 10^{-10}. At these thresholds no passed cell exceeds 0.1 nats, and at this tolerance recall stays below the certificate's. CAI needs the exact behavior posterior over all 2^{11} quality configurations on every history prefix, and that posterior already gives b^μ_k and hence the exact per-target KL; the certificate needs only integer counts on the public support and the teacher's preference map. In our single-thread runs on these contexts the CAI recursion took 13.4 times as long as the exact behavior posterior at $\lambda = 1$, and evaluating the certificate for both mixtures in one pass took 3.7 times as long as the exact behavior posteriors at the two mixtures (descriptive and machine dependent). These ratios time diagnostics on deployment contexts; the ratio of Appendix F.4 instead times building complete selective training targets against do relabeling of the training logs.

Tolerance. The comparisons above count a target as undistorted only within 10^{-10} in probability. Table 23 loosens this tolerance with the point-estimate grid that the same analysis recorded, rendered post hoc by `r39_recorded_tolerance_grid.py` after the grid values had been inspected. The certificate's pass set does not depend on the tolerance, so its recall falls as more targets count as

Table 21: Constrained-loss arms at the intermediate teacher mixture $\lambda = 0.25$, $\lambda = 0.5$, $\lambda = 0.75$ beside the recorded $\lambda = 0$ and $\lambda = 1$ runs of the hard-label study (128 units, learning rate 10^{-3}, five paired seeds with the same initialization and minibatch order). The intermediate-mixture protocol was fixed before its logs were collected, after the $\lambda \in \{0, 1\}$ results were known. Rows A and B give realized-label NLL differences (nats) at own-target-selected checkpoints unless marked final. The deployment pool weights deployment histories on the training and unseen-B layouts equally under four history-only policies; behavior histories are those of the trained mixture on the training layouts (Training) and on the held-out behavior layouts (Held-out). Exact D_{marg} is the marginal distortion of the behavior posterior on the pool, and $G = D_{\mathrm{marg}} - E_{\mathrm{do}} + B_{\mathrm{obs}}$ is the observational-soft minus do-soft against-do KL gap. The increment is (hard-label − do-soft) at the given λ minus the same difference at $\lambda = 0$, paired by seed. Each cell gives the seed mean with the number of seeds whose value has its sign, and below it the paired five-seed 95% percentile bootstrap interval. Row C counts the arm (observational-soft/do-soft/hard-label) chosen out of five seeds among all candidates of all arms trained at that λ; at $\lambda = 0$ observational-soft coincides with do-soft and is not trained.

$\lambda = 0$	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 0.75$	$\lambda = 1$
A. Deployment pool				
Exact D_{marg}				
0.000	0.527	1.318	3.448	18.213
Observational-soft − do-soft				
—	+0.50 (5/5)	+0.73 (5/5)	+1.65 (5/5)	+2.39 (5/5)
	(+0.45, +0.54)	(+0.69, +0.77)	(+1.58, +1.72)	(+2.31, +2.48)
G (against-do KL)				
—	+0.50 (5/5)	+0.74 (5/5)	+1.66 (5/5)	+2.40 (5/5)
	(+0.46, +0.54)	(+0.70, +0.77)	(+1.58, +1.74)	(+2.31, +2.51)
B_{obs}				
—	+0.06 (4/5)	−0.49 (5/5)	−1.69 (5/5)	−15.7 (5/5)
	(+0.02, +0.11)	(−0.52, −0.45)	(−1.76, −1.65)	(−16.0, −15.3)
Hard-label − do-soft				
+1.00 (5/5)	+1.00 (5/5)	+1.24 (5/5)	+1.81 (5/5)	+2.65 (5/5)
(+0.92, +1.10)	(+0.93, +1.07)	(+1.14, +1.35)	(+1.75, +1.86)	(+2.44, +2.85)
increment over $\lambda = 0$				
—	+0.002 (2/5)	+0.24 (4/5)	+0.81 (5/5)	+1.65 (5/5)
	(−0.09, +0.09)	(+0.09, +0.37)	(+0.68, +0.94)	(+1.48, +1.84)
Observational-soft − do-soft, final				
—	+0.95 (5/5)	+2.23 (5/5)	+5.33 (5/5)	+11.7 (5/5)
	(+0.92, +0.98)	(+2.10, +2.36)	(+5.27, +5.40)	(+11.2, +12.1)
Hard-label − do-soft, final				
+23.0 (5/5)	+24.6 (5/5)	+27.2 (5/5)	+31.5 (5/5)	+32.5 (5/5)
(+22.7, +23.4)	(+24.2, +25.0)	(+26.6, +27.6)	(+31.0, +31.9)	(+31.8, +33.5)
B. Behavior histories of the trained λ				
Observational-soft − do-soft, Training				
—	+0.10 (5/5)	−0.02 (3/5)	−0.30 (5/5)	−0.48 (5/5)
	(+0.09, +0.11)	(−0.038, +0.004)	(−0.34, −0.26)	(−0.51, −0.45)
Observational-soft − do-soft, Held-out				
—	+0.18 (5/5)	+0.11 (5/5)	−0.01 (3/5)	−0.23 (5/5)
	(+0.15, +0.20)	(+0.09, +0.13)	(−0.08, +0.08)	(−0.26, −0.20)
Hard-label − do-soft, Training				
+0.35 (5/5)	+0.33 (5/5)	+0.18 (5/5)	−0.21 (5/5)	−0.41 (5/5)
(+0.33, +0.36)	(+0.30, +0.36)	(+0.14, +0.21)	(−0.29, −0.13)	(−0.43, −0.38)
Hard-label − do-soft, Held-out				
+0.30 (5/5)	+0.34 (5/5)	+0.24 (5/5)	+0.02 (3/5)	−0.24 (5/5)
(+0.28, +0.31)	(+0.31, +0.37)	(+0.21, +0.29)	(−0.05, +0.10)	(−0.27, −0.21)
C. Arm chosen (observational-soft/do-soft/hard-label)				
Behavior-label validation				
–/5/0	0/5/0	0/5/0	0/5/0	1/0/4
Deployment-label validation				
–/5/0	0/5/0	0/5/0	0/5/0	0/5/0

Table 22: Cumulative per-target action information (CAI) against exact per-target distortion on the 320 RockSample deployment contexts of Table 14 (post hoc, exploratory; thresholds, weights and outcome rules fixed before computation). Cells are unresolved targets of valid frames, the cells of D_{marg}. A target is undistorted if $|b_k^{\mu} - b_k^{\mathrm{do}}| \le 10^{-10}$, the near-equality tolerance used for certificate coverage; CAI≈0 means $\sum_{s<t} I_{\mu}(A_s; Z_k \mid h_s) \le 10^{-20}$ nats. Undist.: undistorted share (%). Cells in millions. ρ: Spearman correlation of KL and CAI over pooled cells, values at numerical zero tied; per-seed values differ from it by at most 0.005. Sound: P(undistorted|CAI≈0). Recall: P(CAI≈0|undistorted). Cert.: P(certified|undistorted) for the residual-selection certificate on the same cells. Δ: Recall minus Cert. (percentage points). Pos.: P(undistorted|CAI> 0.1). FS: cells with CAI≈0 that are not undistorted; their largest probability difference is 8.5×10^{-10} and largest KL 3.5×10^{-16} nats, and no cell with CAI≈0 has KL above 0.1 nats. Shares use the D_{marg} weights; second lines give unadjusted 95% paired-seed percentile intervals over five seeds with 10,000 draws.

	Cells	Undist.	ρ	Sound	Recall	Cert.	Δ	Pos.	FS
A. All unresolved targets									
$\lambda = 1$	10.33	24.7	0.68	99.992	41.1	73.9	−32.8	4.3	79
		(24.5, 24.9)		(99.991, 99.994)	(40.3, 41.8)	(73.2, 74.4)	(−33.1, −32.4)	(4.3, 4.4)	
$\lambda = 0.5$	10.33	21.1	0.80	99.999	47.2	83.7	−36.6	2.3	11
		(20.8, 21.3)		(99.998, 100.000)	(46.3, 48.0)	(83.1, 84.3)	(−36.9, −36.1)	(2.2, 2.3)	
B. Unresolved targets with do entropy at least 0.05 nats									
$\lambda = 1$	9.21	23.3	0.69	99.992	43.8	78.8	−35.0	3.7	79
		(23.1, 23.5)		(99.990, 99.993)	(43.0, 44.6)	(78.3, 79.2)	(−35.4, −34.6)	(3.7, 3.8)	
$\lambda = 0.5$	9.21	19.6	0.82	99.999	50.8	90.9	−40.1	1.9	11
		(19.4, 19.8)		(99.998, 99.999)	(50.0, 51.8)	(90.5, 91.4)	(−40.5, −39.6)	(1.9, 2.0)	

Table 23: Recall of undistorted unresolved targets, in percent of their D_{marg}-weighted mass, by the certificate (Cert.) and by near-zero cumulative per-target action information at four thresholds in nats, as the total-variation tolerance that defines an undistorted target loosens. Parentheses: soundness of the action-information check, the percent of its passed mass that is undistorted at that tolerance. The certificate passes no target with total-variation error above 1.8×10^{-14}, so its soundness is 100% throughout. Last row of each panel: passed cells with KL above 0.1 nats. Point estimates from the grid recorded by the analysis of Table 22; the first row of each panel reproduces the recall columns of that table.

		Action-information threshold			
Tolerance	Cert.	10^{-20}	10^{-6}	10^{-4}	10^{-2}
A. $\lambda = 1$					
10^{-10}	73.9	41.1 (99.99)	47.9 (92.34)	65.8 (78.15)	80.9 (65.15)
10^{-6}	58.2	32.3 (100.00)	39.5 (96.53)	57.6 (86.89)	75.2 (76.84)
10^{-4}	49.6	27.6 (100.00)	34.5 (98.93)	52.4 (92.69)	70.7 (84.79)
10^{-2}	39.4	21.9 (100.00)	27.7 (100.00)	44.8 (99.75)	64.0 (96.55)
KL > 0.1 passed	0	0	0	2	10,372
B. $\lambda = 0.5$					
10^{-10}	83.7	47.2 (>99.99)	55.3 (90.28)	74.1 (69.20)	87.9 (50.31)
10^{-6}	68.5	38.6 (100.00)	47.1 (93.98)	67.8 (77.40)	84.5 (59.16)
10^{-4}	53.0	29.8 (100.00)	38.2 (98.70)	61.2 (90.36)	81.0 (73.33)
10^{-2}	35.6	20.0 (100.00)	26.0 (>99.99)	45.5 (99.88)	69.6 (93.78)
KL > 0.1 passed	0	0	0	7	1,791

undistorted, to 39.4% and 35.6% at 10^{-2}. With a threshold of at most 10^{-6} nats, action information recovers fewer targets than the certificate at every tolerance in the table, and at tolerances up to 10^{-6} it does so for thresholds up to 10^{-4} nats. With a threshold of 10^{-4} nats it recovers more once the tolerance reaches 10^{-4}, 52.4% against 49.6% at $\lambda = 1$. At 10^{-2} it recovers 44.8% and 45.5% with 99.75% and 99.88% of its passed mass within the tolerance, but it passes 2 and 7 targets whose KL exceeds 0.1 nats, and a threshold of 10^{-2} nats passes 10,372 and 1,791 such targets. The certificate is therefore the stronger check at near-exact tolerances and needs no divergence threshold or numerical action probabilities, whereas a practitioner who accepts distortions of order 10^{-4} or more in probability clears more targets with a looser action-information threshold, at the cost of the exact behavior posterior and a few large misses. Appendix C.8 repeats this comparison at explicit KL tolerances with seed intervals and recomputes both checks under misspecified parameters.

C.8 The certificate at explicit KL tolerances and under misspecified parameters

The comparison above counts a target as undistorted within a tolerance in probability and reports the certificate's recall without a direct counterpart for the false passes of a looser action-information threshold. This post hoc, exploratory analysis, run by `r39_tolerance.py`, scores both at explicit per-target KL tolerances and recomputes them under misspecified parameters. Its tolerances, rules, analyst models and outcome classes were fixed in a preregistration before any of its results were computed. The only earlier computation was a single-context timing probe, which the preregistration discloses. It collects no trajectory and trains no model. Four deviations are logged with its outputs: a renamed summary file, one consistency check that was not preregistered, table columns chosen after the results, and a second rendering of the tables to remove an overfull line, which changed no number.

Design. Part A recomputes the per-cell arrays of Table 22 with the code of that analysis, and they reproduce every recorded estimate exactly. A target is ϵ-safe if its KL is at most ϵ, for ϵ from 10^{-6} to 0.1 nats. The rules are the certificate, CAI at most τ_c on a grid of thresholds, CAI at the threshold that matches the certificate's coverage, and the oracle KL. The first decision compares the certificate's recall of the ϵ-safe mass with that of the best grid threshold that clears no target above ϵ; the second counts the targets above ϵ that the coverage-matched threshold clears. Part B recomputes the certificate, CAI and a plug-in KL under 11 analyst models that change the sensor, the teacher's and the student's action probabilities, the support floor and, at $\lambda = 0.5$ only, the mixture weight, singly and jointly, and scores them against the true per-target KL. At $\lambda = 0.5$ the certificate also uses the history-only student's preferred action, which the analyst either takes from the log or recomputes under its own sensor model.

Tolerance. The certificate clears no target above any tolerance, and the largest KL among its targets is 3.5×10^{-16} nats (Table 24). At every tolerance and both mixtures it recovers more of the ϵ-safe mass than every grid threshold of CAI that clears no target above ϵ, in all five seeds: at 10^{-6} nats 47.9% against 29.1% at $\lambda = 1$ and 48.1% against 28.1% at $\lambda = 0.5$, and at 0.1 nats 35.0% against 24.6% and 22.6% against 16.5%. A looser threshold reverses the order. CAI at most 10^{-4} nats recovers more than the certificate at every tolerance and both mixtures, in all five seeds, by 4.8 and 4.9 points at $\lambda = 1$ for tolerances of 10^{-6} and 0.1 nats and by 11.1 and 6.3 at $\lambda = 0.5$. It does so by clearing targets above the tolerance, 76,102 and 87,953 at 10^{-6} nats and 2 and 7 at 0.1 nats, the largest at 0.28 and 0.20 nats, although all targets it clears together carry at most 4.5×10^{-4}% of D_{marg}. One of them has CAI 3.5×10^{-6} nats and KL 0.28 nats. At the certificate's coverage, $\tau_c = 2.7 \times 10^{-5}$ at $\lambda = 1$ and 1.0×10^{-5} at $\lambda = 0.5$, CAI recovers as much as the certificate from 10^{-4} nats on, with intervals including zero, but less at 10^{-6} nats, -1.1 points (-1.5, -0.7) and -0.6 (-1.0, -0.1), and it clears 43,367 and 21,255 targets above 10^{-6} nats and 1 and 4 above 0.1 nats. In recorded single-thread time on these contexts the certificate for both mixtures took 3.7 times as long as the two exact behavior posteriors, which already give the per-target KL, so in this implementation the oracle KL is the cheapest of the rules.

Misspecified parameters. At $\lambda = 1$ no certified target is distorted under any analyst model or student arm (Table 25). No change of a numeric parameter alters a certified target when the student's action is taken from the log. This follows from the construction, since the certificate reads none of these parameters, and the computation confirms it. A sensor model without an exact reading at the rock, a structural change, removes 82,022 certified targets at $\lambda = 1$ and 83,969 at $\lambda = 0.5$ and, with the student's action taken from the log, adds none. At $\lambda = 0.5$ the certificate is exact while the

Table 24: The certificate against thresholded cumulative action information (CAI) at explicit per-target KL tolerances ϵ, on the 320 deployment contexts and cells of Table 22 (post hoc, exploratory; tolerances, rules and outcome classes fixed before computation; the recomputed arrays reproduce every recorded estimate of that table exactly). A target is ϵ-safe if its KL is at most ϵ. Safe: ϵ-safe share of the unresolved mass (%). Cert.: share of the ϵ-safe mass the certificate clears (recall, %); it clears no target above any ϵ (largest cleared KL 3.5×10^{-16} nats). Clean CAI: the grid threshold τ_c with the highest recall among those that clear no target above ϵ, its recall, and Δ = its recall minus Cert. (points). CAI$\leq 10^{-4}$: recall, Δ, and the number of cleared targets above ϵ ($> \epsilon$). Match: targets above ϵ cleared by CAI at the certificate's coverage ($\tau_c = 2.7 \times 10^{-5}$ at $\lambda = 1$ and 1.0×10^{-5} at $\lambda = 0.5$). In the recorded TV grid of Table 22, a looser CAI threshold matches or beats the certificate's recall: at $|b_k^{\mu} - b_k^{\mathrm{do}}| \leq 10^{-2}$, CAI$\leq 10^{-4}$ recovers 44.8% against 39.4% ($\lambda = 1$) and 45.5% against 35.6% ($\lambda = 0.5$). CAI$\leq 10^{-4}$ also recovers more ϵ-safe mass than the certificate at every KL tolerance tested (10^{-6} to 10^{-1} nats), with Δ intervals above 0, but it clears targets above ϵ, the largest at 0.28 ($\lambda = 1$) and 0.20 nats ($\lambda = 0.5$); all targets it clears together carry at most 4.5×10^{-4}% of D_{marg}. Every grid threshold of CAI that clears no target above ϵ recovers less than the certificate. Second lines: unadjusted 95% paired-seed percentile intervals (five seeds, 10,000 draws); the clean and matched thresholds are selected on the same cells.

			Clean CAI			CAI$\leq 10^{-4}$			Match
ϵ (nats)	Safe	Cert.	τ_c	Recall	Δ	Recall	Δ	$> \epsilon$	$> \epsilon$
A. $\lambda = 1$									
10^{-6}	38.1	47.9	10^{-10}	29.1	-18.8	52.7	4.8	76,102	43,367
		(47.2, 48.5)		(28.5, 29.8)	(-19.1, -18.5)	(52.3, 53.0)	(4.4, 5.2)		
10^{-4}	42.3	43.1	10^{-8}	27.2	-15.9	48.9	5.8	10,895	3,002
		(42.4, 43.7)		(26.5, 27.8)	(-16.2, -15.6)	(48.4, 49.2)	(5.5, 6.2)		
10^{-3}	44.9	40.6	10^{-8}	25.6	-15.0	46.3	5.7	1,578	463
		(39.9, 41.2)		(25.0, 26.2)	(-15.3, -14.7)	(45.8, 46.6)	(5.3, 6.0)		
10^{-2}	47.7	38.3	10^{-6}	26.9	-11.4	43.7	5.4	73	7
		(37.6, 38.8)		(26.3, 27.5)	(-11.6, -11.1)	(43.2, 43.9)	(5.1, 5.7)		
10^{-1}	52.0	35.0	10^{-6}	24.6	-10.4	40.0	4.9	2	1
		(34.4, 35.6)		(24.1, 25.2)	(-10.7, -10.1)	(39.5, 40.3)	(4.6, 5.2)		
B. $\lambda = 0.5$									
10^{-6}	36.7	48.1	10^{-10}	28.1	-19.9	59.2	11.1	87,953	21,255
		(47.4, 48.8)		(27.5, 28.8)	(-20.2, -19.7)	(58.7, 59.6)	(10.7, 11.5)		
10^{-4}	44.3	39.8	10^{-8}	24.0	-15.9	50.8	11.0	7,288	779
		(39.2, 40.5)		(23.4, 24.5)	(-16.1, -15.6)	(50.3, 51.2)	(10.6, 11.3)		
10^{-3}	49.5	35.7	10^{-8}	21.5	-14.2	45.6	9.9	955	36
		(35.0, 36.3)		(20.9, 22.0)	(-14.5, -14.0)	(45.1, 45.9)	(9.6, 10.3)		
10^{-2}	58.7	30.0	10^{-8}	18.1	-12.0	38.4	8.4	52	8
		(29.5, 30.6)		(17.6, 18.5)	(-12.2, -11.8)	(38.0, 38.7)	(8.1, 8.7)		
10^{-1}	78.1	22.6	10^{-6}	16.5	-6.1	28.9	6.3	7	4
		(22.2, 23.0)		(16.2, 16.8)	(-6.2, -5.9)	(28.6, 29.1)	(6.1, 6.5)		

student's preferred action comes from the log, but when the analyst recomputes it under a wrong sensor model, 287 to 613 certified targets per model are distorted, up to 0.69 nats. Below $\lambda = 1$ the certificate therefore needs the student's preference map as well as the teacher's. Under the same models CAI at most 10^{-20} nats changes status on up to 19,565 targets but never clears one above 10^{-6} nats, CAI at most 10^{-4} nats clears up to 360 and 370 targets above 0.1 nats, with true KL up to 16.3 nats, against 2 and 7 under the true parameters, and a plug-in KL at most 10^{-3} nats computed from the analyst's posteriors clears up to 243,790 and 877,805 targets above 10^{-3} nats, with true KL up to 24.4 nats.

Table 25: The certificate, CAI and a plug-in KL computed by an analyst with misspecified parameters, on the cells of Table 24 (post hoc, exploratory; models, arms and decision rules fixed before computation). Truth: sensor correctness $\frac{1}{2}(1 + 2^{-d/20})$, exact at the rock; teacher 0.95 on its preferred action; history-only student 0.9; support floor 0.05. Each row changes the listed parameter(s) (half distance: the 20 in the sensor term); "no exact reading" multiplies the sensor term by 0.9. Student: logged uses the preferred action the student took under the recorded do marginals; rederived recomputes it under the analyst's sensor. Joint combines half distance 40, teacher 0.80, student 0.75 and floor 0.20, with $\lambda' = 0.75$ in panel B. Certificate: certified targets lost and gained; $> 10^{-10}$: certified targets whose true $|b_k^{\mu} - b_k^{\mathrm{do}}|$ exceeds 10^{-10}. CAI$\leq 10^{-20}$: targets whose status changes (none of the targets it clears has true KL above 10^{-6} in any row). CAI$\leq 10^{-4}$: cleared targets with true KL above 0.1 nats. KL$' \leq 10^{-3}$: targets the analyst's $\mathrm{KL}(b_k^{\mathrm{do}\prime} \,\|\, b_k^{\mu\prime}) \leq 10^{-3}$ clears whose true KL exceeds 10^{-3}, and the largest true KL among the targets it clears. No change of a numeric parameter (sensor half distance, teacher, student, support floor, λ') alters a certified target in the logged arm. An analyst sensor with no exact reading removes 82,022 ($\lambda = 1$) and 83,969 ($\lambda = 0.5$) certified targets and adds none. At $\lambda = 1$ no certified target is distorted under any model. At $\lambda = 0.5$ the certificate also uses the student's preferred action; when the analyst recomputes it under a wrong sensor, 287 to 613 certified targets are distorted, up to 0.69 nats. Counts over 10.33 million targets per mixture.

		Certificate			CAI$\leq 10^{-20}$	CAI$\leq 10^{-4}$	KL$' \leq 10^{-3}$	
Analyst model	Student	lost	gained	$> 10^{-10}$	changed	> 0.1	$> 10^{-3}$	max KL
A. $\lambda = 1$								
true parameters	–	0	0	0	0	2	0	1.0×10^{-3}
half distance 10	logged	0	0	0	8,893	2	7,975	1.70
half distance 10	rederived	0	0	0	8,893	2	7,975	1.70
half distance 40	logged	0	0	0	10,320	62	140,870	1.12
half distance 40	rederived	0	0	0	10,320	62	140,870	1.12
no exact reading	logged	82,022	0	0	10,312	18	10,603	24.38
no exact reading	rederived	82,022	0	0	10,312	18	10,603	24.38
teacher 0.80	logged	0	0	0	9,848	38	64,095	1.10
teacher 0.99	logged	0	0	0	5,608	2	5,525	0.29
student 0.75	logged	0	0	0	0	2	0	1.0×10^{-3}
support floor 0.01	logged	0	0	0	5,211	2	5,201	0.41
support floor 0.20	logged	0	0	0	9,265	38	60,650	0.84
joint	logged	0	0	0	19,565	360	243,790	7.76
joint	rederived	0	0	0	19,565	360	243,790	7.76
B. $\lambda = 0.5$								
true parameters	–	0	0	0	0	7	0	1.0×10^{-3}
half distance 10	logged	0	0	0	6,687	11	34,308	1.05
half distance 10	rederived	1,305	573	287	6,586	16	38,814	20.59
half distance 40	logged	0	0	0	8,061	14	263,726	1.29
half distance 40	rederived	1,028	1,170	613	8,026	15	266,864	3.74
no exact reading	logged	83,969	0	0	9,224	9	27,362	5.42
no exact reading	rederived	85,059	856	466	9,188	14	34,122	18.80
teacher 0.80	logged	0	0	0	1,012	7	155,856	1.08
teacher 0.99	logged	0	0	0	328	7	7,181	0.03
student 0.75	logged	0	0	0	801	7	63,209	0.85
support floor 0.01	logged	0	0	0	601	7	18,478	0.15
support floor 0.20	logged	0	0	0	1,593	11	101,369	1.44
$\lambda' = 0.25$	logged	0	0	0	1,129	19	877,805	2.07
$\lambda' = 0.75$	logged	0	0	0	748	19	50,833	6.91
joint	logged	0	0	0	10,067	369	230,584	21.52
joint	rederived	1,028	1,170	613	10,129	370	231,281	21.52

C.9 CLOSED-LOOP CONTROL WITH THE CONSTRAINED PREDICTORS

The offline scores above measure predictions at fixed histories. This study measures native return when a controller acts on the predictions. It was designed after the offline results and an earlier closed-loop study with all-target GRUs were known and is therefore post hoc. Its protocol, run by `r39_closed_loop.py`, was fixed before any return of these predictors on the evaluation layouts was computed. Before the freeze the driver was tested end to end on one smoke layout and four further layouts that are disjoint from every evaluation layout and random stream, and those returns are not reported and changed no part of the design.

Design. The two history controllers that define the greedy and cautious deployment policies, with their recorded thresholds, act on a predictor's eleven marginals instead of the exact do marginals, in the official simulator. The predictors are the constrained-loss arms of Appendices C.4 and C.6, 128 units and learning rate 10^{-3}, for seeds 201–205 at every λ, at the own-target-selected and the final checkpoints. Coordinates that the history makes public are set to their public value before the controller reads a learned predictor, as in every offline score; unclamped marginals are a sensitivity analysis. The references are the exact do marginals, the uninformed prior and the exact behavior posterior b^{μ} of each mixture, propagated online on the realized history. Each seed's conditions share reset, transition and action streams on the training, unseen-B and unseen-A layouts, 64 episodes per layout, for 430,080 rollouts in total. Returns average both controllers and pool the training and unseen-B layouts equally, as in the deployment pool. The two primary contrasts are do-soft minus observational-soft at $\lambda = 1$ at the selected and at the final checkpoints; each counts as positive if its paired five-seed interval lies above zero and all five seeds are positive. If the selected contrast were positive and the final one not, the protocol required reporting that the selected observational-soft checkpoint lies near initialization. The forked rollout code reproduced sixteen recorded batches of the earlier study bit for bit, online and full-sequence logits agreed, the online b^{μ} matched the offline recursion exactly, and 1,260 episodes were replayed step by step against the official scalar environment.

Results. Table 26 reports every condition. Both primary contrasts are positive. At $\lambda = 1$ do-soft returned 36.92, 87% of the range from the uninformed prior, 9.96, to the exact do marginals, 40.79. The selected observational-soft checkpoint, at update 250, returned 11.23: it sampled 0.17 good rocks per episode against 3.75 for do-soft. The contrast was $+25.69$ ($+24.45$, $+26.69$) in all five seeds, and it is positive on each layout group, under each controller and without clamping. Because that checkpoint lies near initialization, the final checkpoints separate the targets from under-training: there observational-soft returned 25.10 and the contrast was $+11.83$ ($+10.63$, $+12.78$), again in all five seeds. At $\lambda = 0.25$, 0.5 and 0.75 the selected contrasts were $+7.29$, $+11.78$ and $+19.39$, growing with λ, although at $\lambda = 0.25$ and 0.5 the exact targets rank the other way, as shown next.

Target and learning. Without learning, the exact b^{μ} returned 31.33 at $\lambda = 1$, so exact do minus exact b^{μ} was $+9.47$ ($+9.07$, $+9.81$). Final observational-soft returned -6.22 relative to its own target, and selected do-soft -3.87 relative to its target. At $\lambda = 0.25$ and 0.5, however, the exact b^{μ} returned more than the exact do marginals, with exact do minus exact b^{μ} of -2.23 (-3.01, -1.74) and -0.52 (-0.98, -0.21). The controllers are fixed heuristics rather than planners, and with the exact do marginals 60% of episodes reach the 60-step cap, so return here measures what these controllers do with a belief, not how accurate the belief is. DSS minus do-soft return was negative at every λ, including $\lambda = 0$, where it was -10.09 at the selected and -20.03 at the final checkpoints. There the two targets coincide in population, so that deficit comes from fitting sampled labels, not from distortion.

Relation to the earlier study. The earlier closed-loop study evaluated the sixty historical all-target GRUs under the same two controllers on four layouts. Their do-soft stayed at the uninformed prior's level, and do-soft minus observational-soft return was -0.926 (-1.385, -0.566) at full mixing, mainly through later episode exits under the greedy controller, and $+2.684$ (0.615, 4.732) at half mixing. With the constrained loss do-soft tracks the exact do marginals offline and in closed loop, and the sign at full mixing reverses.

Table 26: Closed-loop native return of the two fixed history controllers acting on each predictor's marginals (constrained loss, 128 units, learning rate 10^{-3}, five paired seeds). The protocol was fixed after the offline results and the historical closed-loop study were known and before any return of these predictors on the evaluation layouts was computed. Returns are averaged over both controllers and pooled as 0.5 training layouts (61101–08) + 0.5 unseen-B layouts (65101–04), 64 episodes per layout with random streams shared within seed; 430,080 rollouts in total. Publicly determined coordinates are set to their public value before the controller reads a learned predictor. Sel: own-target-selected checkpoint; Fin: update 20,000. Exact b^μ propagates the exact observational marginals of the λ mixture on the realized history; at $\lambda = 0$ it equals the exact do marginals, and observational-soft coincides with do-soft. Contrasts give the seed mean, the number of seeds with its sign, and the paired five-seed 95% percentile bootstrap interval.

	$\lambda = 0$	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 0.75$	$\lambda = 1$
A. Return (pool)					
Observational-soft, Sel	—	31.66	26.81	18.81	11.23
Observational-soft, Fin	—	34.66	32.85	30.08	25.10
Do-soft, Sel	38.23	38.95	38.59	38.20	36.92
Do-soft, Fin	38.37	38.98	38.65	38.14	36.94
DSS, Sel	28.15	25.80	22.50	18.38	10.86
DSS, Fin	18.34	18.28	17.80	17.35	18.47
Exact b^μ	—	43.02	41.32	37.62	31.33
Exact do − exact b^μ	—	−2.23 (5/5)	−0.52 (5/5)	+3.17 (5/5)	+9.47 (5/5)
		(−3.01, −1.74)	(−0.98, −0.21)	(+2.76, +3.51)	(+9.07, +9.81)
Exact do marginals 40.79; uninformed prior 9.96					
B. Share of episodes reaching the 60-step cap (%, pool)					
Observational-soft, Sel	—	55	44	9	0
Observational-soft, Fin	—	49	32	12	3
Do-soft, Sel	64	63	62	64	65
Do-soft, Fin	63	61	63	64	65
DSS, Sel	45	46	26	7	0
DSS, Fin	40	40	39	28	14
Exact b^μ	—	41	20	7	1
Exact do marginals 60; uninformed prior 0					

C. Paired contrasts, $\lambda = 1$	Sel		Fin	
Do-soft − observational-soft	+25.69 (5/5)	(+24.45, +26.69)	+11.83 (5/5)	(+10.63, +12.78)
training layouts	+28.71 (5/5)	(+28.04, +29.39)	+10.62 (5/5)	(+9.91, +11.28)
unseen-B	+22.66 (5/5)	(+20.94, +24.33)	+13.05 (5/5)	(+11.46, +14.36)
unseen-A	+28.57 (5/5)	(+27.76, +29.51)	+17.43 (5/5)	(+16.73, +17.91)
greedy controller	+29.55 (5/5)	(+28.77, +30.12)	+16.07 (5/5)	(+15.29, +16.56)
cautious controller	+21.82 (5/5)	(+20.12, +23.26)	+7.60 (5/5)	(+5.89, +9.06)
unclamped marginals	+26.12 (5/5)	(+24.95, +27.06)	+11.93 (5/5)	(+10.72, +12.88)
DSS − do-soft	−26.06 (5/5)	(−27.01, −24.91)	−18.47 (5/5)	(−20.00, −17.02)
Do-soft − exact do	−3.87 (5/5)	(−4.97, −2.91)	−3.85 (5/5)	(−4.96, −2.76)
Do-soft − exact b^μ	+5.60 (5/5)	(+4.74, +6.42)	+5.61 (5/5)	(+4.75, +6.46)
Observational-soft − exact b^μ	−20.09 (5/5)	(−20.41, −19.71)	−6.22 (5/5)	(−6.66, −5.78)

D. Other mixtures	Sel		Fin	
Do-soft − obs.-soft, $\lambda = 0.25$	+7.29 (5/5)	(+6.41, +8.67)	+4.33 (5/5)	(+4.02, +4.68)
Do-soft − obs.-soft, $\lambda = 0.5$	+11.78 (5/5)	(+10.42, +13.01)	+5.80 (5/5)	(+5.04, +6.66)
Do-soft − obs.-soft, $\lambda = 0.75$	+19.39 (5/5)	(+18.60, +20.59)	+8.06 (5/5)	(+7.50, +8.87)
DSS − do-soft, $\lambda = 0$	−10.09 (5/5)	(−11.75, −8.95)	−20.03 (5/5)	(−20.78, −19.26)
DSS − do-soft, $\lambda = 0.25$	−13.14 (5/5)	(−13.74, −12.54)	−20.70 (5/5)	(−21.12, −20.28)
DSS − do-soft, $\lambda = 0.5$	−16.09 (5/5)	(−17.29, −14.49)	−20.85 (5/5)	(−21.90, −19.80)
DSS − do-soft, $\lambda = 0.75$	−19.82 (5/5)	(−20.92, −18.59)	−20.79 (5/5)	(−21.70, −19.87)

C.10 CORE EVIDENCE, REPRODUCIBILITY PATHS AND ARCHIVE COVERAGE

The core reproduction directories distinguish three operations. *Recorded-statistic replay* aggregates saved context or episode scores, recomputes seed intervals and reproduces Table 2. *Frozen-model rescoring* runs the stored GRU weights on retained public inputs and reconstructs the scores from reference posteriors. *Paired retraining* fits new weights from the same inputs; the original forty models were not retrained, but the sensitivity and selection study of Appendix C.4 trained forty new models with the supplied driver and the same data, initialization seed and minibatch seed. Model manifests bind seeds, layouts, policy, posterior target, loss, checkpoint update, selection objective, input and weight hashes; the machine recipe key `original` denotes the *all-target loss*, not the historical Original group.

Executed core paths. The training-study export contains all 80 selected/final weights, 560 raw-history files and 880 reference records, along with validation and training logs. A separate path-configurable scoring implementation evaluates all 880 model/data contexts and reproduces the 220 retained Table 2 summary fields within 5.4×10^{-15}; complete frozen-model scoring in a clean directory takes 38.0 seconds on the recorded CUDA setup, and statistical replay reproduces the summary files byte for byte. For the channels, clean-directory CPU rescoring of 60 soft-branch checkpoints and 60 test-history files takes 20.85 seconds, and all 3,240 original aggregate comparisons satisfy the fixed absolute-plus-relative Float32 tolerance. These are runs of separate scoring code within the same research workspace, not third-party replications or new training. The scorers' inputs (71 public channels for the training study, 38 for the channels) contain no current action, hidden state, posterior or future observation, which perturbation checks confirm; a missing required file is an error, and scoring never substitutes paper constants. The README files give relative-path commands, tested software versions, input hashes and tolerances for `c16_table1_replay.py`, `c16_rescore.py`, `training_diagnostics.py`, `training_endpoint_summary.py` and `channels_rescore.py`. Intermediate weights and full logits were not retained at every validation candidate, so the 81-point curves are scalar-record replay and independent rescoring covers only the selected/final endpoints.

Archive coverage. The single submitted supplement is this core export: the public-value training study (all 560 history archives, 80 selected/final weights, validation records and curves, and the retraining and rescoring code), the channel study (both five-seed groups, all 60 soft-predictor checkpoints and 60 test-history archives), the binary residual-selection enumeration, the collection-policy reanalysis and the figures. Under `rocksample/recorded/` it also holds the records of the sensitivity study (`r30`), of its hard-label extension and $\lambda = 0$ control and the exact mixture sweep (`r34`, `r34_exact_mixture_sweep`), of the intermediate-mixture study (`r38_mid_mixture`), of the per-target action-information analysis (`r38_target_action_info`), of the late-checkpoint selection analysis (`r38_late_selection`), of the more-layout study (`r39_more_layouts`), of the tolerance grid and the tolerance and misspecification analysis (`r39_recorded_grid`, `r39_tolerance`) and of the closed-loop study (`r39_closed_loop`), with the CartPole seed summaries behind Section 5.6 (`r38_inputs`): protocols, run logs, candidate validation curves, selections, context-level scores and summaries as applicable, but no candidate weights. Under `channels/recorded/r39_per_target_learning` it holds the protocol, run log, exact enumeration checks and seed-level metrics of the paired per-target learning test, without weights. `verify` checks every packaged file, `replay` recomputes the training-study and channel tables without inference, and `frozen` reruns every included checkpoint. The tables and macros of the sensitivity, hard-label, intermediate-mixture, late-selection, action-information, more-layout, tolerance, closed-loop and per-target learning analyses are regenerated by their own scripts, listed in the reproducibility statement, not by `replay`. The export does not include the historical Original/Fresh models and histories, the CartPole archives and checkpoints, or the certificate reproductions. The estimates reported for those studies in Appendices E and F are reconstructible from seed vectors recorded in the authors' local study archives, which are not part of the export, and their checkpoints cannot be rescored from the submission. Hardware and backend differences can change retrained weights.

Table 27: Protocol inventory and glossary of the learning studies. Each supervision comparison changes only the label (Appendix D.5); Appendix E.3 summarizes the historical Original/Fresh study.

	Controlled channels	RockSample
Models and seeds	two five-seed groups of the three arms; innovation coupling $\rho \in \{0, 0.5, 1\}$ (Section 5.2)	new study: forty models (five seeds, $\lambda \in \{0, 1\}$, both soft arms, all-target and constrained losses); sensitivity study: constrained recipe at $\lambda = 1$, three optimization settings, one with 256 units, extended by a hard-label arm in each setting and a $\lambda = 0$ do-soft and hard-label control (Appendix C.4); intermediate-mixture study: the three constrained arms of the default setting at $\lambda \in \{0.25, 0.5, 0.75\}$, same seeds (Appendix C.6)
Layouts	none	24, namely eight training, four validation, four held-out behavior (distinct from unseen-A/B), and deployment groups unseen-A (deployment-label validation only) and unseen-B, four each
Behavior histories	teacher action parity equals the queried X bit with probability 0.9	teacher mixture equation 1 at the stated λ, $\epsilon = 0.05$
Deployment histories	uniform policy; a policy that follows the public cue bit	random (uniform over the 16 actions) and three history-driven policies on exact do marginals, all history-only: history-greedy, history-cautious (belief thresholds $0.55/0.65$ and $0.70/0.85$) and different-seed (greedy with another tie rule)
Scoring	eleven target bits; every episode has 60 frames	fixed-60 (episode loss sum divided by 60, zero after termination; Appendix D.4)
Deployment weighting	uniform and cue-following policies equally	four policies equally. Table 2A and Table 4: training, unseen-A and unseen-B layouts weighted $1/2$, $1/4$, $1/4$; Table 2B: training layouts; sensitivity study: training and unseen-B layouts equally
Status	protocol fixed before collection; hard-label, $\rho = 0.5$ and cue-retention 0.6 analyses declared secondary	fixed before collection: new study; fixed before training: sensitivity study and its hard-label extension, with descriptive intervals; post hoc: the collection-policy comparison and the decomposition of Table 4

Term	Meaning in RockSample
Unresolved	a rock whose current quality the public history does not determine
Public	the baseline that outputs determined qualities and $1/2$ otherwise
Public clamping	evaluation-time replacement of every predictor's determined outputs by the Public values
Valid frame	a pre-termination frame, $t < L_i$
Loss mass	a loss summed over unresolved coordinates and valid frames, not divided by the coordinate count; per-rock values divide pooled loss by pooled coordinate mass
Seen	the eight training layouts, scored on new histories
Selected, final	the checkpoint chosen by the stated rule; the checkpoint at update 20,000
Own-target selection	minimum of an arm's own training loss on validation-layout behavior histories, after public clamping; for a hard-label arm it coincides with behavior-label validation
Behavior-label validation	publicly clamped realized-label NLL on teacher-collected validation-layout histories, the same score for every arm
Deployment-label validation	the same score on simulator-labeled history-only rollouts from the unseen-A layouts under the four deployment policies, without a posterior filter

D Experimental Protocol and Reproducibility

D.1 Behavior and target policies

The Bandit/Chain grids and original RockSample replay use equation 1 with $\epsilon = 0.05$ and $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$. The complete action distribution is stored at each transition. Target trajectories are collected independently. Within each fixed study protocol, model and hyperparameter selection does not use that study's target-test outcomes; earlier test results informed the post hoc historical follow-ups (Appendix E.3). Every online belief path receives the current public observation and, from the second step onward, the preceding realized action. Privileged state is used only for training targets, behavior propensities and evaluation.

Population support, finite coverage, and model recovery. With $\epsilon = 0.05$, the uniform component in equation 1 assigns every action probability at least $0.05/16 = 0.003125$ in RockSample and $0.05/2 = 0.025$ in CartPole. For a fixed public context, shared initial distribution and invariant environment kernels, any finite target-policy prefix with positive probability also has positive behavior probability. This is a population support statement. It neither ensures that a finite log contains all target state–action regions nor bounds estimation error.

The main RockSample study uses known kernels to construct exact references. Training and held-out layouts are disjoint public contexts; the shared environment structure permits filtering on each new layout, without asserting empirical training coverage of those layouts. The factor-estimation study listed in the guide to the appendix estimated initial rock probabilities, a sensor parameter and teacher confidence within specified families, retaining known transitions and the preference map, and so tested parameter recovery only in that structure. CartPole uses an approximate, regularized dynamics ensemble and has no certified finite-sample recovery bound. Its same-parent action branches cover actions at logged parent states, without exhausting the continuous state space.

D.2 Exact synthetic grid

ActionLeak-Bandit crosses four observation-error values, five teacher accuracies, five mixture levels, four target policies, and three analytic seed replicas. The resulting 1,200 condition–policy keys are evaluated by direct atom enumeration. ActionLeak-Chain additionally crosses three horizons and three target action biases, giving 2,700 keys evaluated by a separate dynamic program. For each key we persist exact PSPD, both target NLLs, their identity error, TAL, action-channel trajectory KL, and its TAL decomposition error. The consolidated file contains 31,200 rows. At horizon 20, observation-error probability 0.4, teacher accuracy 0.99 and $\lambda = 1$, the three Chain target policies share cumulative episode TAL 0.668549 nats but have time-averaged PSPD 1.758053 (action-one bias), 3.285992 (random) and 5.872944 (action-zero bias). Collapsing the three identical analytic seed copies leaves 1,300 settings (400 Bandit and 900 Chain), or 260 five-mixture curves with four adjacent comparisons each (1,040 total). The separate binary residual-selection control is specified in Appendix E.1.

The independent RockSample replay uses three seeds and five mixture levels, with public layout identity as the split unit (four training, two validation and two test layouts) and an episode horizon of 60. It covers 15 archives, 300 episodes and 1,275 metric rows, and encodes the 2,048 latent quality configurations, movement, sample termination, sensing likelihoods, teacher probabilities and public mask updates separately from the main filter implementation. Its exact semantic metric file holds every matched observational, do and no-action score and supplies the RockSample row of Table 28.

D.3 RockSample GRU template and deployment policies

Every RockSample learner uses one template: a one-layer GRU with 128 units (256 in one sensitivity setting) and eleven Bernoulli output logits, with 78,603 parameters at 128 units. Its 71 input channels contain 33 raw current-observation values, a 16-way previous-action encoding (all zero initially), and the 22 public ordered-layout coordinates divided by 10. Exact beliefs and hidden qualities are training targets or evaluation references, never model inputs. Adam uses learning rate 10^{-3} (10^{-4} in one sensitivity setting), batches of 64 complete padded sequences, and gradient clipping at norm 1. Each arm receives the same initial parameters, ordered minibatches, and 20,000 updates. Loss sums over rocks and averages valid time steps within a batch; padding contributes no loss. GRU optimization and inference use float32 on CUDA without TF32 or mixed precision; exact joint references and log scores use float64 on CPU, and training targets are converted to float32.

Collection uses the pinned official POBAX transitions (Tao et al., 2025) with a 60-action horizon. The four history-only deployment policies are random (uniform over the 16 actions), history-greedy and history-cautious (candidate and sampling thresholds 0.55 and 0.65, and 0.70 and 0.85, on exact do marginals with distance penalty 0.015) and different-seed (the greedy rule with a seed-derived tie offset).

D.4 Metrics, inference and fixed-time risk

For categorical targets, NLL is the negative log probability of the realized state. PSPD is exact in finite domains, and synthetic TAL is the exact conditional mutual information in equation 36. CartPole NLL is the two-dimensional diagonal-Gaussian log score in raw velocity units; CartPole method effects are paired on seed, mixture and target policy, and condition-cell contrasts are averaged within seed before the five seed means receive equal weight. Unless stated otherwise, percentile intervals resample the training seed as one cluster for 10,000 replicates. Exact synthetic equalities are reported with their maximum absolute numerical discrepancy rather than a confidence interval.

Fixed-time risks and episode termination. For a fixed seed, policy and layout group with N episodes, let L_i be the valid pre-action length and $\ell_{it}(q)$ the sum of eleven binary log losses, so that the fixed-time risk

$$R_t^0(q) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{t < L_i\}\ell_{it}(q), \qquad t = 0, \ldots, 59, \tag{70}$$

counts terminated episodes as zero loss while keeping them in the denominator. The survival proportion $a_t = N^{-1}\sum_i \mathbf{1}\{t < L_i\}$ is determined at the scored history, unlike the episode weight $1/L_i$ of the historical score, which divides each episode's loss sum by its realized length and so depends on later outcomes. R_t^0/a_t is the risk conditional on survival when $a_t > 0$, and the fixed-horizon summary $60^{-1}\sum_{t=0}^{59} R_t^0$ is the fixed-60 score of the new and sensitivity studies.

D.5 Information requirements and related learning objectives

Related approaches differ from the matched supervision comparison in the information they use and in what they evaluate: Swamy et al. (2022) contrast off-policy and on-policy context inference in online imitation, Vuorio et al. (2024) deconfound imitation with unlabeled static context, Cai et al. (2024) distill an expert with active privileged exploration, and Warrington et al. (2021) evaluate student control return with a jointly adapted privileged expert. Here fixed state-labelled logs and paired recurrent learners isolate one intervention, which posterior supplies the training labels, and the object of evaluation is physical target-marginal prediction risk under policy replacement. None of these contrasts is a performance comparison.

Prior identities and collection regimes. Corollary 1 is the conditional importance-weight identity of Rowland et al. (2020, Prop. 5.1) with a state target Z_t in place of the return and the do law as the sampling measure. For finite actions, accepting each do path with probability $W_{t-1}^{\mu} \le 1$ yields the behavior posterior, so the criterion reads $S \perp\!\!\!\perp Z_t \mid H_t$ for this selection event S, the distribution-level form of the d-separation condition that characterizes s-recoverability (Bareinboim et al., 2014, Thm. 1); the cancellation example in equation 34 meets it although X enters the selection weight. Fixed logs of a policy acting on hidden state ($\lambda = 1$ in equation 1) train an offline driving world model (Hu et al., 2022), a driving model trained on a privileged rule-based expert (Shao et al., 2023), benchmark baselines trained on logs of an RL expert that uses ground-truth states (Jia et al., 2024), and learned filters whose logs include a true-pose shortest-path policy (Jonschkowski et al., 2018); in that policy-transfer test (their Sec. V-A.4), an LSTM filter fails once training and test policies differ. We do not claim that these systems show target-specific distortion. RMA (Kumar et al., 2021), locomotion students (Lee et al., 2020; Miki et al., 2022) and the final DAgger stage of Learning by Cheating (Chen et al., 2020) train on student rollouts ($\lambda = 0$), whereas Scaffolder (Hu et al., 2024), Guided Policy Optimization (Li et al., 2026) and a high-speed flight student that follows its own plans only near a privileged reference (Loquercio et al., 2021) keep privileged action dependence ($\lambda > 0$). Recovering the do recursion from full Markov labels requires invariant kernels, state–action coverage and model recovery rather than a specified teacher mixture, and the integer-count certificate additionally needs a product do posterior, the preference map and an action family with shared probability ratios. High-dimensional observations, learned black-box teachers

and unknown dynamics require estimation and coverage analysis that the controlled tasks do not evaluate.

E Additional Results

The exact synthetic grid evaluates 1,200 ActionLeak-Bandit and 2,700 ActionLeak-Chain condition–policy keys (Appendix D). Mean exact PSPD over the grid is 0.081 in Bandit (0.242 at $\lambda = 1$) and 0.221 in Chain (0.682 at $\lambda = 1$), and the identity between PSPD and the excess target NLL of the behavior posterior holds with maximum error 8.88×10^{-16}. Both TAL and PSPD vanish at $\lambda = 0$ and rise with mixing; at $\lambda = 1$ mean TAL is 0.189 in Bandit against mean PSPD 0.242, and 0.318 in Chain against 0.682. The two quantities are rank-correlated across unique analytic conditions (Spearman 0.983 in Bandit and 0.972 in Chain; Figure 4), yet policies with the same episode TAL can have different PSPD, as the Chain example of the exact synthetic grid protocol shows. Table 28 records the matched-filter sign reversal in the Bandit, Chain and RockSample replays. At $\lambda = 1$ the do filter has zero calibration error in Bandit and Chain and 0.063 in the RockSample replay, against 0.180, 0.178 and 0.199 for the observational filter; the corresponding target RMSEs are 0.371, 0.313 and 0.367 against 0.440, 0.420 and 0.477.

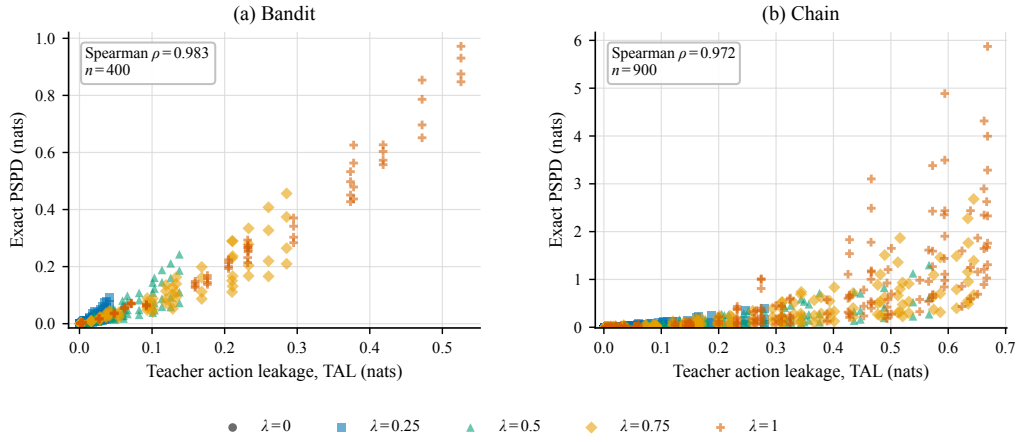


Figure 4: Teacher-action leakage versus exact PSPD for every unique analytic condition. Colors and markers identify the privileged mixture coefficient.

Figure 5 shows the target-level picture in RockSample. At one fixed history, two certified uncertain rocks keep their marginals while another is strongly distorted, the within-history heterogeneity quantified in Appendix C.2; panel (a) and the example are described in Appendix F.3.

Table 28: Matched filter semantics at $\lambda = 1$. Bandit and Chain are a deterministic analytic diagnostic; RockSample is a registered sampled replay scored by exact filtering. Entries are NLL differences (left minus do-filter), so negative Δ_B means observational filtering fits behavior histories better and positive target differences favor the do-filter. Bold contrasts show the behavior-to-target sign reversal.

Environment	Δ_B obs–do	Δ_T obs–do	Δ_T noact–do
ActionLeak-Bandit	$\mathbf{-0.156}$	$\mathbf{+0.243}$	0
ActionLeak-Chain	$\mathbf{-0.176}$	$\mathbf{+0.498}$	$+0.187$
RockSample	$\mathbf{-2.094}$	$\mathbf{+19.462}$	$+0.003$

E.1 A controlled residual-selection example with fixed information

The following exact example isolates residual selection while holding both teacher-action information and target uncertainty fixed. It complements the same-TAL target-policy comparisons in the main finite-state experiments. Let $S_1 = (X, Y) \in \{0, 1\}^2$, with

$$P(X = x, Y = y) = \begin{cases} (1 + \rho)/4, & x = y, \\ (1 - \rho)/4, & x \neq y, \end{cases} \qquad 0 \leq \rho \leq 1. \tag{71}$$

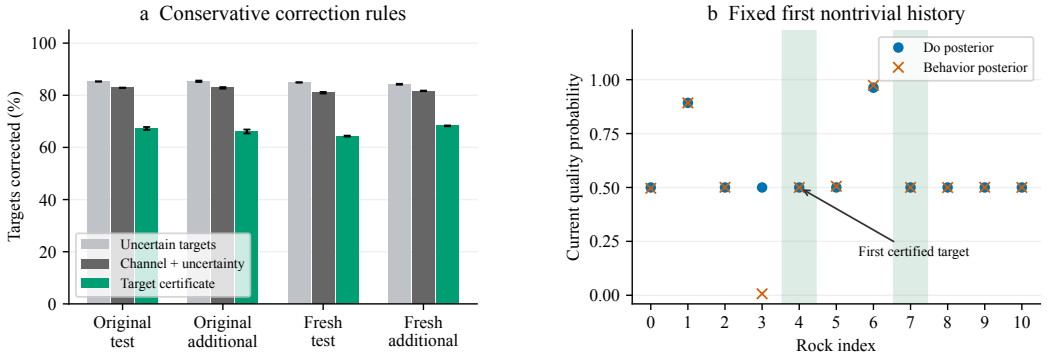


Figure 5: Target-wise supervision selection at full mixing. (a) Conservative correction fractions on all target histories, with 95% paired collection-seed intervals. Each rule retains do-equivalent target marginals numerically. (b) All eleven marginals at the fixed first nontrivial history. Green bands mark certified targets; both remain uncertain while another target is strongly distorted. Uncertified targets are corrected conservatively, without asserting that their correction is necessary.

Both marginals are uniform and $P(X = Y) = (1 + \rho)/2$. The state remains unchanged, $S_2 = S_1$, and $O_1 = O_2 = 0$. Thus H_1 is constant and $H_2 = (O_1, O_2, A_1)$ carries only the realized action. The teacher chooses $A_1 = X$ with probability $r \in (1/2, 1)$, whereas the target policy chooses each action with probability $1/2$. We score $Z_2 = Y$. All actions have positive probability; structural zeros in the state prior at $\rho = 1$ are shared by the behavior and target laws.

Since A_1 depends on S_1 only through X, its marginal is uniform and

$$\mathrm{TAL}_1 = I_\mu(S_1; A_1 \mid H_1) = I_\mu(X; A_1) = \log 2 - h(r), \tag{72}$$

where h is binary entropy in nats. This quantity is independent of ρ. The target action is independent of both state components, so the do-posterior of Y is uniform for either history and $H_\pi(Y \mid H_2) = \log 2$ throughout the sweep.

Write $c = \rho(2r - 1)$. Direct marginalization gives

$$b_2^{\mu,Y}(Y = A_1 \mid H_2) = \frac{1 + c}{2}, \qquad b_2^{\mathrm{do},Y}(Y = A_1 \mid H_2) = \frac{1}{2}. \tag{73}$$

For a fixed action a, the path weight is $W_1^\mu = \mu(a \mid X, Y)$, its do expectation is $1/2$, and its conditional expectation given $Y = a$ is $(1 + c)/2$ (given $Y \neq a$, it is $(1 - c)/2$). The residual-selection ratio in Corollary 1 is therefore $1 + c$ or $1 - c$, respectively. Its variation, rather than the unchanged full-state teacher information, determines

$$\mathrm{PSPD}_2(\mu \to \pi) = \frac{1}{2} \log \frac{1}{1 + c} + \frac{1}{2} \log \frac{1}{1 - c} = -\frac{1}{2} \log(1 - \rho^2(2r - 1)^2). \tag{74}$$

At $\rho = 0$, Y remains maximally uncertain and $\mathrm{TAL}_1 > 0$, yet $\mathrm{PSPD}_2 = 0$. Thus zero residual selection does not require a history-sufficient target. For fixed r, increasing ρ increases PSPD while preserving both TAL and target conditional entropy.

We enumerate all eight (X, Y, A_1) atoms using rational probability masses, then compute posterior KL, state log risks, trajectory KL, and the conditional weight ratio by separate aggregations. The sweep uses $r = 0.9$ and 21 values of ρ from 0 to 1. TAL remains 0.368064 nats and target conditional entropy remains 0.693147 nats, while PSPD grows from 0 to 0.510826 nats. These are time-2 PSPD and one-action cumulative TAL, not time-averaged quantities or independent sampling replicates.

E.2 Position-Only CartPole

The Easy confirmation uses Position-Only CartPole from POPGym (Morad et al., 2023) with five training seeds, five mixture levels and four history-only target policies; each seed–mixture condition has 10,000 training, 2,000 validation and 4,000 test episodes, and the state-aware teacher component is a five-parameter CEM policy. DSS is a one-layer 128-unit GRU that outputs a diagonal Gaussian over the two hidden velocities. Do-filter distillation shares its architecture, behavior histories and

Table 29: Exact residual-selection control at teacher accuracy $r = 0.9$. Varying the correlation ρ between teacher-relevant X and scored Y leaves both TAL_1 and the target conditional entropy fixed. At $\rho = 0$, Y remains uncertain but $\mathrm{PSPD}_2 = 0$. Entries are exact-enumeration quantities in nats, not sampled estimates. Bold identifies positive TAL with zero target PSPD.

ρ	$P^{\mu}(Y = A_1 \mid A_1)$	TAL_1	$H_{\pi}(Y \mid H_2)$	PSPD_2
0	0.500	**0.368**	0.693	**0.000**
0.25	0.600	0.368	0.693	0.020
0.5	0.700	0.368	0.693	0.087
0.75	0.800	0.368	0.693	0.223
1	0.900	0.368	0.693	0.511

Table 30: Five-seed Position-Only CartPole Easy confirmation. Primary effects aggregate the complete Easy grid; the final two rows are supplementary $\lambda = 1$ minus $\lambda = 0$ DSS endpoints. Effects are paired left-minus-right contrasts. Directional gates require the registered sign in at least four of five seeds; no-action and factor rows are two-sided, and the branch row uses symmetric equivalence classification. All intervals use 10,000 percentile bootstrap resamples of training seeds as clusters. Bold estimates have paired intervals excluding zero (unadjusted). The branch row instead emphasizes its registered equivalence decision.

Estimand	Effect	95% interval	Decision	
Behavior observational − do NLL, overall Easy grid	$\mathbf{-0.029}$	$(-0.039, -0.019)$	negative; `pass`	5/5;
Target observational − do NLL, overall Easy grid	$\mathbf{+0.154}$	$(+0.100, +0.218)$	positive; `pass`	5/5;
Target no-action − do NLL, overall Easy grid	$\mathbf{+5.058}$	$(+4.816, +5.300)$	two-sided	
Target do-filter − DSS NLL, overall Easy grid	$\mathbf{-8.092}$	$(-11.308, -5.816)$	two-sided	
20k branch − passive target NLL, overall Easy grid	$+4.49 \times 10^{-4}$	$(+2.36 \times 10^{-4}, +6.57 \times 10^{-4})$	two-sided; $\epsilon = .10$: **equivalent**	
DSS behavior NLL, $\lambda = 1$ minus 0	$\mathbf{-1.006}$	$(-1.171, -0.858)$	two-sided	
DSS target NLL, $\lambda = 1$ minus 0	$\mathbf{+15.496}$	$(+12.161, +19.002)$	two-sided	

optimizer and minimizes $D_{\mathrm{KL}}(q_{\mathrm{filter}} \| q_{\mathrm{GRU}})$ to the passive do-filter's velocity moments; action dropout and IPS use the same architecture. The fitted filters use 256 particles and a ten-member regularized dynamics ensemble, and every baseline selects the earliest minimum of its own objective on behavior-validation histories.

In the ideal real-valued environment two position observations and the preceding action determine the velocities, so exact PSPD is zero (Appendix A.6); on 600 held-out episodes the reconstruction error is at most 1.1791×10^{-5}, a finite-precision error rather than a bound on posterior KL. The fitted filter nonetheless shows an observational-minus-do target contrast of 0.4973 nats at 1,024 particles and 0.4843 at 4,096, which falls to 0.0896 when the observation scale is reduced from 0.005 to 0.001; none of these identifies the true environment's posterior KL.

Over the grid (Table 30), mean target NLL is 3.24 nats for DSS and -1.08 for do-filter distillation, against -4.70 and -4.85 for the observational and passive do filters; action dropout scores 7.18–19.63 and IPS 6.22–12.96. From $\lambda = 0$ to $\lambda = 1$ DSS target NLL changes by $+15.496$ nats ($+12.227$, $+19.248$). Distillation, whose targets do not depend on the behavior policy, also rises, from -2.423 to $+3.598$ nats ($+6.021$, positive in 5 of 5 seeds), so part of the DSS rise reflects the changed training histories; the DSS-minus-distillation increment is $+9.474$ nats ($+5.954$, $+13.304$), positive in 5 of 5 seeds. The intervals of the DSS and DSS-minus-distillation increments enumerate all 3,125 ordered resamples of the five seeds, whereas the table and the $\lambda = 0$ contrast below use 10,000 random resamples, so the DSS interval differs slightly from the table's. At $\lambda = 0$, DSS already trails the passive do filter by 2.501 nats (1.474, 3.937) despite a zero observational–do contrast. The three prespecified fixed-head return contrasts are zero in every paired seed, and the matched-head branch-minus-passive and passive-minus-DSS return effects, $+0.018$ and $+0.017$, have intervals that span zero. Raw IPS is unstable across seeds (seed 1 scores 36.085 against 3.944 for DSS) but beats DSS at full mixing (6.430 against 13.118).

E.3 Historical Original/Fresh RockSample study (summary)

Before the studies of Appendix C, two five-seed groups trained the template of Appendix D.3 at $\lambda \in \{0, 0.5, 1\}$ with hard-label DSS, do-soft and observational-soft arms, each selecting its own behavior-validation minimum over 81 candidates. The Original group (seeds 0–4) chose its soft targets and selection rule post hoc, after fixed-final checkpoints had overtrained. The Fresh group (seeds 5–9, new layouts in every split) fixed the three-arm design before collection, but that design was informed by the Original results. Each group has eight training and four validation layouts and two held-out groups of four (test and additional layouts), and scores divide each episode's eleven-rock loss sum by its realized length L_i rather than by 60.

At full mixing, observational-soft minus do-soft target NLL is 0.726 (0.402, 1.052) on Original and 0.510 (0.300, 0.719) on Fresh test layouts, positive in every seed, while the behavior contrasts are -0.499 and -0.493. On Original test histories, exact marginal distortion is 18.413 nats against a learned expected-risk gap of 0.846, so the learned gap does not estimate exact distortion. The hard-label and do-soft arms select update 250 in every Original $\lambda = 1$ run. The hard-label arm is better than do-soft on behavior and worse on target histories at $\lambda = 1$ in every Original seed and both held-out groups, and its target NLL stays above the uninformed $11 \log 2$ reference. Its target gap to do-soft is already 0.636 (0.551, 0.745) at $\lambda = 0$, and the $\lambda = 1$ minus $\lambda = 0$ increment is 0.187 $(-0.025, 0.373)$ on test layouts, so the evidence for an increment is mixed; in the Fresh group the full-mixing hard-label target gap does not exceed its zero-mixing gap on the test layouts. At $\lambda = 0$ the hard-label arm scores 6.541 against 5.905 for both soft arms on Original test layouts, yet $\lambda = 1$ do-soft is worse than it on the same target histories in all five seeds, a contrast that changes collection mixture and label form together. At full mixing the random-policy gaps on test layouts do not exclude zero in either group (-0.079 Original, -0.100 Fresh).

Half mixing. At $\lambda = 0.5$, observational-soft minus do-soft target-test NLL is 0.031 $(-0.127, 0.241)$ in the Original group and 0.156 $(0.020, 0.357)$ in the Fresh group, positive in four of five Fresh seeds (unadjusted). One interval crossing zero and the other excluding it does not establish a difference between groups, and the Fresh design was informed by the Original results. The rise of the equal-policy pooled contrast from Original to Fresh mainly follows the random-policy estimate, which changes from -0.413 to $+0.372$ with both intervals crossing zero, whereas all three nonrandom means decrease (from 0.152/0.154/0.231 to 0.093/0.092/0.069). Seeds, layouts, trajectories and fitted networks all change between groups, and this comparison isolates none of those causes.

Layout-by-policy re-evaluation. The twenty selected $\lambda = 1$ soft GRUs were rescored, with unchanged weights, on 51,200 new episodes that cross each group's eight training layouts and eight new layouts with the behavior policy and the four deployment policies; fixed-60 estimands were fixed before collection, and the design followed inspection of the earlier results. On training layouts both groups reverse, including under Bonferroni sensitivity intervals: the Original realized-NLL gaps are -0.156 under behavior and $+1.116$ under pooled deployment, and the Fresh gaps have the same signs. On unseen behavior histories, clamped observational-soft has no observed advantage (the Original interval includes zero and the Fresh interval is positive), which limits a claim of uniform reversal. Do-soft's unresolved own-target error on training layouts rises from 0.074 nats/rock under behavior to 0.228 (Original) and 0.219 (Fresh) under deployment; deployment also raises sensing from 6.8–6.9 to 33.6–34.0 checks per 100 valid frames, a change in information histories that is not identified as the cause of the error. Absolute quality stays near Public: clamped do-soft improves on Public by 0.070 and 0.093 nats, against 1.971 and 1.938 for exact do (3.6% and 4.8% of that gain).

F Target-wise Supervision Certification in RockSample

F.1 A sufficient certificate from residual selection

Let $X \in \{0,1\}^{11}$ contain the initial rock qualities. At a fixed pre-action public history h_t, the current quality $Z_{t,j}$ equals X_j if rock j has not been sampled, and zero otherwise. Under the do law, independent initial qualities, observations that sense individual rocks, and publicly determined sampling transitions preserve the product form

$$P^{\mathrm{do}}(X = x \mid h_t) = \prod_{j=1}^{11} p_{t,j}(x_j). \tag{75}$$

The support $\mathcal{S}_t = \prod_j \mathrm{supp}(p_{t,j})$ is determined by public observations: a check at an unsampled rock's location resolves its initial bit, and sampling resolves its current quality, while its initial bit remains a latent nuisance unless earlier observations resolved it.

Proposition 5 (Safe reuse of a target marginal)**.** *Assume equation 75 and positive behavior action likelihoods. A current target is safe if it is publicly determined. An unsampled, unresolved target j is also safe if*

$$W^{\mu}_{t-1}(x; h_t) = W^{\mu}_{t-1}(x \oplus e_j; h_t) \quad \textit{for all } x \in \mathcal{S}_t, \tag{76}$$

where $x \oplus e_j$ flips bit j. Every safe target satisfies $b^{\mu,Z_j}_t = b^{\mathrm{do},Z_j}_t$.

Proof. For a determined target, both posteriors concentrate on the same value. Otherwise the product law makes X_{-j} conditionally independent of X_j. Equation 76 makes W a function of X_{-j} on the supported configurations. Therefore $\mathbb{E}_{\mathrm{do}}[W \mid X_j, h_t] = \mathbb{E}_{\mathrm{do}}[W \mid h_t]$, and Corollary 1 gives the result. □

The product-law assumption is essential. With correlated initial qualities, even a weight depending only on another bit can distort the target, as in Appendix E.1. The certificate is sufficient: the conditional mean of W can be target-independent even when the pointwise bit-flip condition fails.

F.2 Exact integer implementation and supervision rule

The supported RockSample behavior mixture has two action-probability levels, according to whether the privileged teacher prefers the realized action. At full mixing $W = C(h_t)r^{c(x)}$, where $c(x)$ counts prior realized actions preferred under initial configuration x and $r > 1$ is the ratio of preferred to nonpreferred action probabilities; at half mixing the counts are grouped by whether the public-history student prefers the realized action. Equality of every integer count under the target-bit flip implies equation 76 without a divergence threshold and for every positive value of the shared ratios; if ratios may vary over time, each preference indicator is tested separately on current support. At zero mixing every target is safe. The implementation enumerates the 2,048 initial configurations compatible with the public support, and counts at time t include actions strictly before t. With $m_{t,j}(h_t)$ indicating certification, the selective supervision target

$$\widetilde{p}_{t,j} = m_{t,j} p^{\mu}_{t,j} + (1 - m_{t,j}) p^{\mathrm{do}}_{t,j} \tag{77}$$

equals $p^{\mathrm{do}}_{t,j}$ for every history and target by Proposition 5, so it preserves the full do-soft objective. The mask uses public observations, past actions, the layout and the specified behavior policy; posterior divergences, hidden-state labels and network predictions do not select it, and exact posteriors only verify its output. Uncertified coordinates can still satisfy posterior invariance, so an unverified result does not establish distortion.

Applying the certificate. The criterion is fixed from the public history, the environment dependence structure and the specified preference map and action family before any posterior discrepancy is audited. Dependent targets or approximate weights need justified dependence, bit-flip and log-weight error budgets, which fitted probabilities alone do not supply. No selectively supervised predictor was trained; Appendix F.4 measures coverage and cost on the training histories.

F.3 Complete diagnostic protocol and results

The diagnostic covers all 80 target archives of the historical study (seeds 0–9, four policies, both held-out layout groups): 20,480 complete episodes at all three mixtures, with the design specified after the earlier prediction results but before any certificate outcome was computed. Correction fractions divide by valid frames times eleven targets and average episodes, layouts, policies and seeds equally; paired percentile intervals use 10,000 draws. Besides no, full and uncertain-target correction, a channel-and-uncertainty gate corrects uncertain targets once an earlier teacher action distribution varies over the public support; unlike the certificate, it does not localize dependence of the realized path weight, and neither computes TAL.

At full mixing, the certificate corrects 67.21% and 66.09% of targets in the original group, and 64.34% and 68.28% in the fresh group (Table 31), between 13.39 and 16.75 percentage points fewer than the channel-and-uncertainty gate. The selective targets match complete do supervision numerically, whereas leaving all behavior targets unchanged gives exact sum-marginal distortions of

18.413, 17.628, 17.907 and 17.591 nats per frame. A nontrivial subset excludes determined targets and constant weights: certified targets with conditional entropy at least 0.05 nats and nonconstant W on public support. Its episode-weighted fractions of all valid target cells are 7.79%, 8.24%, 8.17% and 5.54%. Pooling target cells over episodes and policies (`seed_pooled_metrics.csv`), 9.07, 9.61, 9.53 and 6.53% of unresolved cells are nontrivially certified, and 19.3, 20.8, 22.7 and 17.3% are certified in total.

Because the certificate is only sufficient, these rates do not measure how much of the posterior-equality set it recovers. At full mixing, among unresolved targets whose two posteriors agree within 10^{-10} in probability, the cumulative certificate covers 77.51% and 79.21% in the original test and additional splits and 83.09% and 74.74% in the fresh splits (post hoc, on the same archives); at tolerance 10^{-4} this coverage falls to 50.39–61.15%. Retained marginals match do within 1.57×10^{-14} across all mixtures. The first nontrivial history in lexicographic order (fresh seed 5, different-seed policy, test episode 0, frame 3) leaves all eleven targets uncertain; rocks 4 and 7 are certified at probability 0.5 under both posteriors, while rock 3 has do probability 0.5 and behavior probability 0.006975 (Figure 5b).

Table 31: Selective supervision in RockSample at $\lambda = 1$. Percentage of target coordinates corrected and residual exact sum-marginal KL per frame. Entries pool the four fixed policies; intervals condition on the reported layouts. The certificate chooses targets from public histories and action factors; posterior KL is used only for verification. Bold marks the fewest corrections among zero-residual rules.

Group / split	Rule	Corrected (%)	Residual KL	Extra safe (%)
Original / test	None	0.0	18.413	–
	All coordinates	100.0	$< 10^{-10}$	–
	Uncertain targets	85.3	$< 10^{-10}$	–
	Channel + uncertainty	82.9	$< 10^{-10}$	–
	Target certificate	**67.2**	$< 10^{-10}$	15.7
Original / additional	None	0.0	17.628	–
	All coordinates	100.0	$< 10^{-10}$	–
	Uncertain targets	85.4	$< 10^{-10}$	–
	Channel + uncertainty	82.8	$< 10^{-10}$	–
	Target certificate	**66.1**	$< 10^{-10}$	16.8
Fresh / test	None	0.0	17.907	–
	All coordinates	100.0	$< 10^{-10}$	–
	Uncertain targets	84.9	$< 10^{-10}$	–
	Channel + uncertainty	81.0	$< 10^{-10}$	–
	Target certificate	**64.3**	$< 10^{-10}$	16.6
Fresh / additional	None	0.0	17.591	–
	All coordinates	100.0	$< 10^{-10}$	–
	Uncertain targets	84.2	$< 10^{-10}$	–
	Channel + uncertainty	81.7	$< 10^{-10}$	–
	Target certificate	**68.3**	$< 10^{-10}$	13.4

F.4 Coverage and cost on actual training histories

A further study, fixed after the deployment-history results and before its own coverage or timing results were computed, applies the gates to all 30 behavior-training archives (seeds 0–9, $\lambda \in \{0, 0.5, 1\}$, 61,440 episodes, 3 009 266 valid frames). Full do supervision runs the exact eleven-marginal recursion; the factor rule and the cumulative certificate copy an existing behavior-soft target at certified coordinates, assumed already available, and construct do targets elsewhere. At full mixing, cumulative reuse is 37.87–38.58% of training targets, only 0.040–0.042 percentage points above the factor rule (0.49–0.84 points on pooled deployment histories), and the maximum selective-target error is 1.06×10^{-14}.

Net preprocessing cost. Neither selective procedure saves time in this implementation (one pinned CPU per process; each seed's median of three repetitions). Across the positive mixtures, complete selective output takes 87.1–362.8 times the full-do time; at full mixing, full do takes 0.044–0.045

seconds per archive, the factor procedure 4.923–5.055 and the cumulative procedure 3.836–3.949 seconds. The gates enumerate quality configurations, whereas full do maintains eleven marginals, and their peak memory also exceeds full do. Both gates avoid about 30% of full-do sensing updates at full mixing, and the extra cumulative reuse avoids almost none, so reuse rates count retained labels, not filtering work eliminated. Full do is the cheaper target-generation choice on these logs, and this known-family comparison supplies no dependence or weight-error budget for unknown structure.