DISCREPANCY IS NOT ADVICE UTILITY: ON PRIVILEGED MULTI-AGENT DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

When should decentralized agents imitate a privileged teacher? Reproducibility-based weighting favors teacher actions that a local auxiliary policy can predict, but predictability alone need not make an imitation update beneficial. In a two-agent cooperative game with an optimal teacher and an exact auxiliary, we show that discrepancy-weighted Jensen–Shannon divergence (JSD) imitation yields strictly lower team return than constant weighting after a sufficiently small update of equal joint-logit norm. The weighting shifts update mass toward the agent whose change is harmful to the current team. Yet the ordering reverses at the limiting policies under continuous-time pure JSD imitation. The counterexample thus rules out a general local-improvement guarantee for this weighting rule, without implying that it harms learning in general. Separately, exact-reference calibration shows how auxiliary capacity and fitting affect discrepancy at fixed information. A five-task neural study compares combined quality/discrepancy weighting with quality-only weighting and reference systems. Two comparisons match both selected imitation coefficients and learning rates; the others compare separately selected procedures. For the fixed teachers and selected procedures, combined weighting improves over RL on two warehouse tasks but shows no statistically resolved gain over quality-only weighting on any of the five tasks. These results distinguish two questions: whether an agent can reproduce a teacher's actions, and whether following those actions improves team return.

1 INTRODUCTION

A privileged teacher can coordinate actions using information unavailable to individual agents at deployment. Distillation offers a way to transfer that knowledge while retaining decentralized execution. The difficulty is deciding which targets to imitate. An agent may be unable to infer the teacher's action from its own observation. Even when it can, that action may work well with the teacher's teammates and poorly with its current ones.

Adaptive imitation addresses the information constraint by fitting a local auxiliary to the teacher and using their agreement to regulate supervision (Weihs et al., 2021). This suggests favoring targets that a local predictor can reproduce. But does better reproducibility imply a better learning update? The distinction is easy to state and harder to isolate: poor outcomes can also come from a weak teacher, an underfit auxiliary, or a larger update. A useful test must remove those explanations.

Our central result does so in a two-agent coordination game. The teacher is optimal, the auxiliary is exact, and the gated and constant-weight updates have the same joint logit norm. Nevertheless, discrepancy weighting produces a strictly lower return after a sufficiently small simultaneous update. The mechanism is a change in direction: the gate allocates more of the step to the agent whose movement is less favorable to the current team. An intrinsic equal-KL construction gives the same qualitative result. Crucially, the disadvantage is local. In the same game, continuous-time pure JSD imitation reaches a better limiting policy under the gate. Neither the immediate effect nor its persistence follows from teacher reproducibility alone.

We study exponential discrepancy weights applied to an additive JSD imitation loss. The neural experiments combine that weight with a value-based quality weight and compare the product with its components and simpler controls. We test a design heuristic: using local reproducibility to

allocate an imitation budget. The contribution is a controlled comparison of directions that removes teacher quality, auxiliary fitting, and step size as explanations for their ordering.

Two studies place the counterexample in context. First, exact calibration shows what changes when the auxiliary must be learned: a measured discrepancy mixes unavailable information with fitting error. Second, a complete five-task, 400-run neural comparison asks whether the additional discrepancy factor improves a selected learning procedure. Combined weighting beats local RL on two warehouse tasks, but no adjusted comparison resolves an improvement over quality-only weighting. Selected coefficients or learning rates differ in three tasks; we distinguish those procedure comparisons from the two that match both settings. These results constrain attribution of the observed gains, rather than proving that the gate has no effect.

Our contributions are:

- An optimal-teacher counterexample with an exact auxiliary and equal update budget, together with an equal-KL extension and a contrasting long-run result (Section 3).
- An exact calibration study that separates auxiliary fitting effects from conditional teacher variation (Section 4).
- A complete neural comparison and selection controls that distinguish gains from teacher supervision from evidence for the additional discrepancy factor (Section 5).

2 Discrepancy and supervision utility

We consider cooperative learning with a frozen privileged teacher and independently acting students. The teacher observes X, while agent i has local information H_i. Let $T_i(\cdot \mid X)$ be its teacher target, $\pi_i(\cdot \mid H_i)$ its executable policy, and $q_i(\cdot \mid H_i)$ a separately trained auxiliary that predicts the teacher. The deployed joint policy factorizes as $\prod_i \pi_i(a_i \mid H_i)$. In the local distillation experiments, H_i contains the current local observation and agent identity; the students share actor parameters, and the auxiliary has its own parameters and optimizer.

2.1 What the auxiliary estimates

Under a fixed joint distribution of (X, H_i), minimizing population cross-entropy gives the conditional average of the teacher distribution (Weihs et al., 2021):

$$q_i^*(a \mid H_i) = \mathbb{E}[T_i(a \mid X) \mid H_i]. \tag{1}$$

This target yields a useful decomposition of forward KL:

$$\underbrace{\mathbb{E}\,\mathrm{KL}(T_i \| q_i)}_{\text{measured discrepancy}} = \underbrace{\mathbb{E}\,\mathrm{KL}(T_i \| q_i^*)}_{\text{conditional variation}} + \underbrace{\mathbb{E}\,\mathrm{KL}(q_i^* \| q_i)}_{\text{fitting error}}. \tag{2}$$

The identity follows by conditioning the difference of the first two terms on H_i and substituting Equation 1. For finite information and action spaces, positive fitted probabilities make the terms finite; more general spaces require integrability. Thus a large residual can reflect missing information, imperfect fitting, or both.

The gates studied here use Jensen–Shannon divergence (JSD), which is bounded by $\log 2$ with natural logarithms. Equation 2 is a forward-KL identity, not an additive decomposition of JSD. We therefore calibrate the fitted JSD directly against two exact population statistics:

$$R_i^* = \mathbb{E}_{H_i} \mathbb{E}_{X \mid H_i}\, \mathrm{JSD}(T_i(\cdot \mid X), q_i^*(\cdot \mid H_i)), \tag{3}$$

$$P_i = \mathbb{E}_{H_i} \mathbb{E}_{X, X' \overset{\text{iid}}{\sim} p(\cdot \mid H_i)}\, \mathrm{JSD}(T_i(\cdot \mid X), T_i(\cdot \mid X')). \tag{4}$$

The conditional-mean residual R_i^* compares the teacher with its local cross-entropy target. The pairwise statistic P_i compares two teacher targets consistent with the same local information. They measure different aspects of conditional variation; neither is an estimate of reward.

2.2 THE GATE AND THE DECISION IT CONTROLS

For a visited state, define the teacher–auxiliary discrepancy as

$$d_i = \mathrm{JSD}(T_i, q_i). \tag{5}$$

The discrepancy and quality weights are

$$w_i^{\mathrm{dec}} = \exp(-\alpha d_i / \log 2),$$
$$w^{\mathrm{quality}} = \sigma\left(\kappa \frac{\widehat{V}_T - \widehat{V}_S}{\max(\mathrm{mean}_{\mathrm{batch}}\,|\widehat{V}_T|, 0.1)}\right). \tag{6}$$

Here σ is the logistic sigmoid, $\widehat{V}_T$ and $\widehat{V}_S$ are teacher and student value estimates, and $\alpha, \kappa > 0$. We compare ungated imitation, each weight separately, and their product, called *combined* weighting. The student objective adds $\lambda\,\mathrm{mean}[\mathrm{stopgrad}(w) D(T, \pi)]$ to the PPO policy and entropy loss, with an ungated value-regression loss. Here $\mathrm{stopgrad}$ stops gradients through the weight. The main analysis and neural comparison use $D = \mathrm{JSD}$; historical controls also include forward and reverse KL. Because the mean divides by sample count, a gate changes both *where* supervision is applied and *how much* is applied. This distinction motivates the constant and shuffled controls below.

The quality weight introduces a second distinction. Even exact values $V_T(s)$ and $V_S(s)$ compare complete continuation policies. Taking one joint teacher action and then returning to the student instead has utility

$$U(s) = \mathbb{E}_{a \sim T(\cdot|s)}[Q_S(s, a)] - V_S(s). \tag{7}$$

Consequently, the value gap need not estimate either $U(s)$ or the return change from a distillation update. Appendix F.1 gives two games with the same teacher and student root-state values and zero discrepancy, but opposite signs of $U(s)$. Reproducibility, policy quality, and intervention utility require separate measurements.

3 EXACT DISCREPANCY CAN FAVOR A WORSE LOCAL UPDATE

We first remove the two most immediate explanations for a poor distillation update: an inaccurate teacher and an underfit auxiliary. A two-agent game shows that discrepancy weighting can still favor a worse update than constant weighting, even when both explanations are absent and the step norms match.

A hidden bit $Z \sim \mathrm{Bernoulli}(3/4)$ determines the correct action. The team receives reward one only when both agents choose Z. The teacher observes the bit, chooses (Z, Z), and achieves return one. Each student observes a constant and independently chooses action one with probability p_i. The exact auxiliary is therefore $q = (1/4, 3/4)$, while team return is

$$R(p_1, p_2) = \tfrac{3}{4} p_1 p_2 + \tfrac{1}{4}(1 - p_1)(1 - p_2) = p_1 p_2 - \tfrac{1}{4}(p_1 + p_2) + \tfrac{1}{4}. \tag{8}$$

The more frequent target, action one, has lower discrepancy from q. Every decreasing exponential gate consequently assigns weights $w_1 > w_0 > 0$. Its preference is correct about reproducibility, yet it need not be correct about the current team's return.

3.1 A COMPARISON AT EQUAL UPDATE NORM

Let $\theta_i = \mathrm{logit}(p_i)$, and let g and v be negative logit gradients of the population JSD imitation losses with target weights (w_0, w_1) and constant weight $m = (w_0 + 3w_1)/4$, respectively. Both update all agents simultaneously. Normalizing the directions removes the difference in total step size, so the comparison tests the allocation of the update.

Proposition 1 (Local disadvantage with an optimal teacher) *For every $0 < p_1 < p_2 < 1/4$ and finite $w_1 > w_0 \geq 0$, define $d_g = g/\|g\|_2$, $d_m = v/\|v\|_2$, and $\delta = \nabla_\theta R \cdot (d_g - d_m)$. Then $\delta < 0$, and*

$$R(\theta + \eta d_g) < R(\theta + \eta d_m) \qquad \textit{for every } 0 < \eta < -4\delta.$$

Both joint logit displacements have norm exactly η.

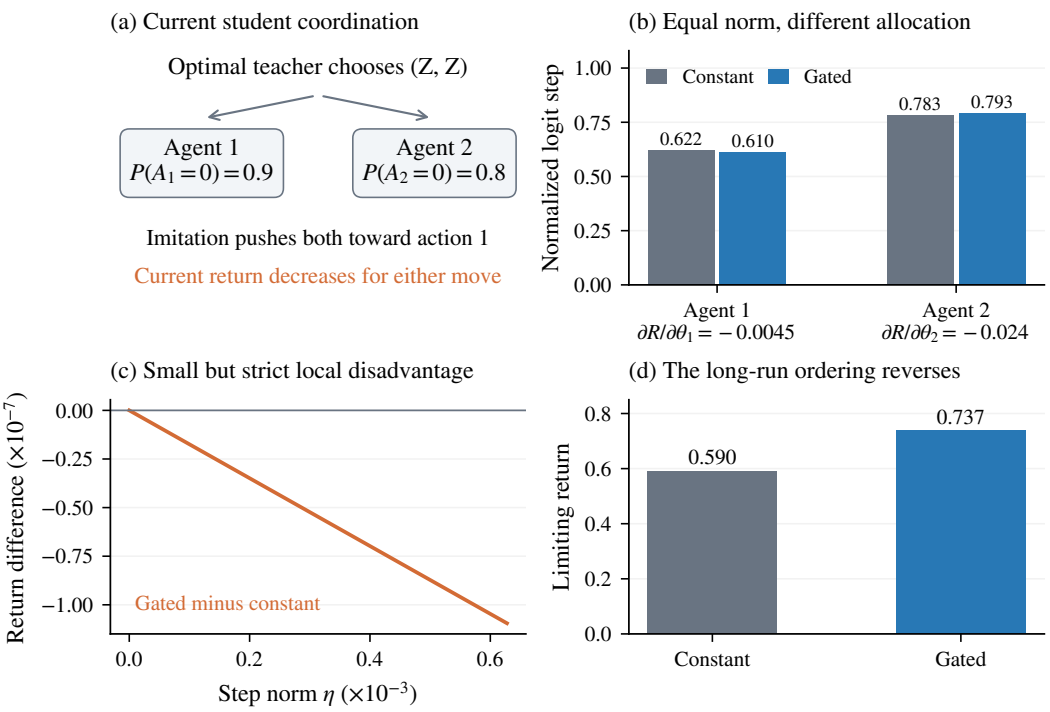


Figure 1: **A local counterexample, with a different long-run outcome.** (a) Both students initially prefer action zero; moving either toward one reduces current team return. (b) With equal joint logit norm, the gate shifts the direction toward agent 2, whose return derivative is more negative. (c) The resulting return difference is small but strictly negative throughout the displayed interval guaranteed by Proposition 1. (d) Continuous-time pure JSD imitation favors the gate at its limiting policy. All panels evaluate the game analytically, with $(p_1, p_2) = (0.1, 0.2)$ and $w_z = \exp[-2\,\mathrm{JSD}(\delta_z, q)/\log 2]$; they are not neural training results.

Proof sketch. Below probability $1/4$, both imitation directions increase both logits, whereas both components of the return gradient are negative. Relative to the constant direction, discrepancy weighting allocates more of the normalized step to the second agent, whose change has the less favorable return derivative. Ordering the two direction slopes gives $\delta < 0$. A global bound of $1/4$ on the operator norm of the return's logit Hessian then gives $R(\theta+\eta d_g) - R(\theta+\eta d_m) \le \eta\delta + \eta^2/4$. Appendix P provides the complete proof.

The gate changes how a move toward action one is allocated between agents. That move can reduce the current team's chance of agreeing, despite the optimal teacher and exact auxiliary.

The strict inequality requires unequal initial probabilities; the normalized directions coincide at $p_1 = p_2$. A separate natural-gradient construction in Appendix T preserves the local disadvantage at equal intrinsic Fisher distance and exactly equal joint forward KL.

3.2 Why the time horizon changes the conclusion

The local comparison does not predict the endpoint of repeated imitation. For fixed positive weights $w_1 > w_0$, continuous-time JSD imitation converges to a common probability p_w for both students. Constant weighting converges to p_c, with $p_w > p_c > 1/2$. Equation 8 is increasing along (p, p) for $p > 1/4$, so the gated endpoint has higher return (Figure 1d). Appendix P also establishes an early cumulative disadvantage followed by eventual benefit along normalized imitation flow.

The counterexample rules out a specific implication: optimal teaching and exact local prediction do not guarantee that discrepancy weighting improves an imitation step at a fixed budget. The small effect in panel (c) establishes that logical failure, not a practically large training loss. The proposition

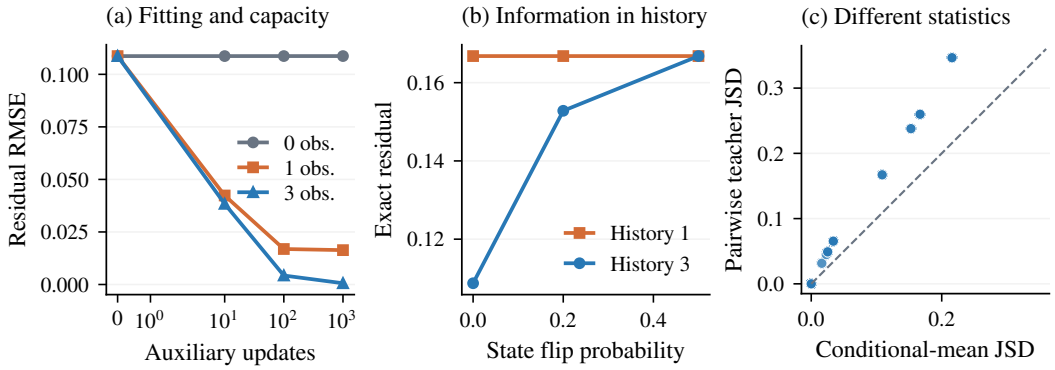


Figure 2: **Fitted discrepancy depends on the auxiliary, not just the information gap.** (a) Error relative to the exact conditional-mean residual, with history length three and 4,096 fitting samples. Curves vary the number of recent observations available to the auxiliary. (b) Exact residual with a correct teacher and observation noise 0.25. Longer histories help only when states are persistent. (c) Pairwise teacher JSD and conditional-mean residual across all 54 exact conditions; the dashed line is equality. Panel (a) reports RMSE over the specified condition/replicate grid; (b,c) are exact population quantities.

also concerns pure JSD imitation rather than PPO or ADVISOR. Predicting the effect on a complete learning procedure requires evidence beyond this local construction.

4 Calibrating discrepancy against an exact reference

The counterexample assumes an exact auxiliary. We now ask what the measured signal means when that auxiliary must be fitted. An enumerable model makes the distinction observable: we can hold the information available to the student fixed while varying capacity, data, and optimization.

We use a stationary binary hidden state that flips with probability f. Each observation independently misreports the state with probability p, and a local history contains one or three observations. The teacher assigns probability $1 - e$ to the current hidden-state label. Enumerating states and histories gives the exact conditional mean q_i^* and both population discrepancy statistics.

The study covers 54 information/teacher conditions and 21,600 fitting cases, with ten data replicates per configuration. It varies observation noise, state persistence, teacher accuracy, accessible history, sample size, and fitting duration. A direct empirical conditional mean controls for finite data, and independent held-out samples separate measurement error from auxiliary fitting error. Appendix E.1 specifies the grid and fitting procedure.

Fitting error persists without missing information. With noiseless observations, a correct teacher, and the full history class, the exact conditional-mean residual is zero. The fitted residual is nevertheless positive: its mean falls from 0.215762 at initialization to 0.077731, 0.009116, and 0.001017 after 10, 100, and 1,000 updates. Figure 2a separates this optimization effect from limited auxiliary capacity. An auxiliary that sees no observations cannot remove its error by training longer, while one that sees the complete history approaches the exact reference.

A symmetry check separates agreement from accuracy. Replacing a correct deterministic teacher ($e = 0$) with an inverted teacher ($e = 1$) preserves the exact residual in every matched information condition. Accuracy changes from one to zero, but action-label symmetry leaves the discrepancy unchanged. This is a sanity check on the statistic's meaning, rather than a new failure mechanism: agreement with a local predictor does not encode whether the teacher's labels are correct.

History and the choice of statistic also matter. At observation noise 0.25, three observations reduce the residual when the state persists. The benefit disappears at flip probability 0.5, when pre-

Table 1: **Distillation gains and gate gains are different comparisons.** Final return, mean (sample SD), over ten training seeds and 320 evaluation episodes per policy, using 19,988,480 student steps. † marks matched coefficients and learning rates for combined and quality-only; the other tasks have different selections. Bold marks the focal procedure. Figure 3 gives adjusted quality-only contrasts.

	LBF		RWARE		
Method	Full	Partial	Large†	Small	Tiny†
Local baseline and JSD procedures (feedforward)					
RL	1.000 (0.001)	0.987 (0.009)	3.963 (2.628)	5.308 (3.070)	13.438 (4.857)
Ungated JSD	1.000 (0.000)	0.979 (0.009)	10.834 (0.384)	12.433 (0.383)	13.818 (0.263)
Quality-only JSD	1.000 (0.000)	0.985 (0.007)	10.744 (0.270)	12.440 (0.597)	14.425 (0.404)
Combined JSD	0.999 (0.002)	0.990 (0.005)	10.928 (0.327)	12.386 (0.321)	14.417 (0.505)
Linear-decay JSD	1.000 (0.000)	0.988 (0.005)	10.552 (0.507)	12.700 (0.422)	14.848 (0.750)
Alternative objective (feedforward actor)					
ADVISOR adaptation	1.000 (0.000)	0.985 (0.004)	11.557 (0.598)	14.598 (0.186)	17.160 (0.823)
Reference systems (recurrent actors)					
MAPPO	1.000 (0.000)	1.000 (0.001)	0.009 (0.007)	0.207 (0.536)	3.509 (6.207)
MAGPO	1.000 (0.000)	0.999 (0.002)	12.614 (0.561)	14.042 (1.435)	17.909 (1.072)

vious observations carry no information about the current state (Figure 2b). Pairwise teacher JSD also differs from the conditional-mean residual in nontrivial conditions (Figure 2c). A calibration result for one statistic cannot be transferred to the other. These effects are larger than the measurement error in this study: held-out residual estimates have pooled RMSE 0.000465 relative to the fitted auxiliary's exact population residual. Numerical checks of Equation 2 agree to 2.22×10^{-16}.

5 Does the discrepancy factor improve neural learning?

We next compare combined and quality-only weighting in neural learning. RL and other systems provide context for the added discrepancy factor.

5.1 Evaluation protocol and comparison scope

The primary study includes fully and partially observed Level-Based Foraging (LBF) and three RobotWarehouse sizes: large, small, and tiny (Papoudakis et al., 2021). LBF uses a 100-step horizon. Eight procedures each receive ten fresh confirmation seeds, for 400 runs in total. Selected recipes are locked before confirmation; every run receives 19,988,480 joint environment steps and a separate 320-episode final-policy evaluation. The analysis specification was fixed after launch but before complete-grid aggregation (Appendix J).

The local distillation procedures use a frozen autoregressive MAT teacher (Wen et al., 2022), a feedforward decentralized actor, and a separate auxiliary. Targets condition on preceding sampled student actions. The discrepancy and quality weights use $\alpha = 2$ and $\kappa = 4$. We also include a code-style ADVISOR adaptation: it samples $A \sim T$, weights teacher-action cross-entropy by $q(A)^{\alpha_A}$ with $\alpha_A = 4$, and weights policy and entropy terms by the complement. Reproduced MAGPO and MAPPO use recurrent actors; MAGPO co-trains its guider (Li et al., 2026b; Yu et al., 2022). These reference systems match tasks and student steps, but not architectures or total compute.

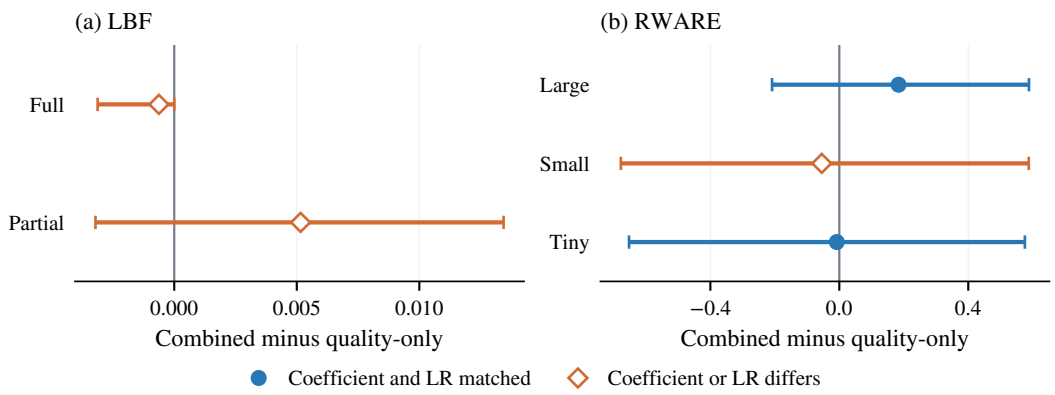


Figure 3: **The additional discrepancy factor has no resolved return benefit.** Combined minus quality-only final return in all five tasks. Filled circles mark matched selected coefficients and learning rates; open diamonds mark comparisons where one differs. Bars are the stored nominal 99.857% bootstrap intervals adjusted over all 35 primary contrasts, with ten training seeds per method. The two panels use different return scales. Positive favors combined; all intervals include zero. The complete seven-comparator results remain in Appendix J.

Which comparisons isolate the added factor most closely? Combined and quality-only select the same imitation coefficient and learning rate on RWARE-large and RWARE-tiny. On small, their coefficients differ (0.3 versus 0.1); on both LBF tasks, their learning rates differ. We mark this distinction in Table 1 and Figure 3. The latter three comparisons evaluate selected procedures rather than changing only the discrepancy factor. All five condition on one fixed teacher per task. Appendices J.1–J.2 give teacher references and the selection protocol.

Uncertainty. We average evaluation episodes within each trained policy and independently bootstrap training runs within each task and method. The primary family contains all 35 combined-minus-comparator final-return contrasts. Stored percentile intervals use 100,000 draws and a Bonferroni adjustment across the family. These small-sample intervals are approximate; tasks retain separate return scales, and episodes are not independent training replicates (Agarwal et al., 2021).

5.2 No resolved gain from the additional discrepancy factor

The two comparisons with matched coefficients and learning rates give small point estimates of opposite signs. Combined minus quality-only is $+0.184$ on RWARE-large, with adjusted interval $[-0.209, 0.588]$, and -0.008 on RWARE-tiny, with interval $[-0.652, 0.576]$. Thus neither resolves an improvement from the additional factor. The other three intervals also include zero (Figure 3), with recipe differences further limiting their interpretation. These intervals describe uncertainty about the effect; they are not evidence of equivalence. The pointwise 95% intervals also include zero on both tasks (Table 23), so the unresolved direction is not solely due to the 35-comparison adjustment.

Distillation gains are a different comparison. Combined exceeds local RL on RWARE-large and RWARE-small by 6.965 and 7.077, respectively, with adjusted intervals above zero. Table 1 shows why this does not identify a benefit of the discrepancy factor: quality-only attains similar returns. Reference systems also change the ranking. Combined falls below ADVISOR on small and tiny, and below MAGPO on LBF-partial, large, and tiny under the same adjusted intervals. Near-ceiling LBF-full returns offer little separation. Appendix K reports a separate wider-network comparison.

5.3 Do simpler allocation controls explain the gain?

A separate allocation-control study replaces the combined weight array by its mean (constant) or permutes its entries (shuffled). Both controls preserve the current learner's batch mean while changing allocation across samples. Linear decay provides a schedule independent of discrepancy. The

methods retain independently selected recipes and ten confirmation seeds each. This cohort uses 50-step LBF and 128 evaluation episodes, so it is not pooled with the primary study.

Table 2: **Simple allocation controls leave the extra gate benefit unresolved.** Combined minus comparator in the allocation-control study. Each cell gives the mean difference and its unadjusted per-contrast 95% interval, from ten training seeds per method. Positive favors combined. This separate cohort uses 50-step LBF and 128 evaluation episodes per policy; it is not pooled with Table 1.

Task	Constant	Shuffled	Linear decay
LBF-full	+0.0000 [0.0000, 0.0000]	+0.0008 [0.0000, 0.0023]	+0.0000 [0.0000, 0.0000]
LBF-partial	+0.0047 [-0.0094, 0.0191]	-0.0004 [-0.0125, 0.0109]	+0.0117 [-0.0008, 0.0246]
RWARE-small	+0.0234 [-0.4477, 0.4937]	+0.0227 [-0.4664, 0.5211]	-0.9180 [-1.3930, -0.4391]
RWARE-tiny	+0.0070 [-0.4578, 0.4555]	+0.2719 [-0.1781, 0.7023]	-2.0312 [-2.5320, -1.4609]

Every final-return and monitoring-AUC interval comparing combined with constant or shuffled weighting includes zero (Table 2 and Appendix B). Linear decay exceeds combined in final return by 0.918 on small and 2.031 on tiny. These results leave attribution to state-specific selection unresolved. Preserving each learner's own batch mean also does not match supervision mass across different training trajectories; Appendix F gives the refresh and normalization rules.

Secondary evidence. The supplementary policy diagnostics and prospective advice studies use warehouse teachers weaker than the corresponding RL students. Their findings therefore cannot separate gate quality from limited useful advice (Appendix A). This post-hoc placement retains all results; the primary comparison includes all five tasks.

6 Related work

Privileged imitation and local realizability. ADVISOR identifies the imitation gap caused by privileged information and fits an auxiliary policy to balance imitation with reinforcement learning (Weihs et al., 2021). Its conditional policy average provides the population target used here, and its Appendix A.4 discusses finite estimation and online fitting lag. Our calibration quantifies these effects against an exact reference; the target itself is not a new result. Song et al. (2025) analyze privileged distillation in perturbed Block MDPs and continuous control, including settings where estimated belief error does not predict suboptimality across tasks. We instead isolate the return direction of a local imitation update in cooperative learning and compare it at a fixed update budget.

Coordination under decentralized execution. AgentMixer studies the asymmetry between fully observable joint policies and local product policies (Li et al., 2025). MAGPO addresses decentralized realizability by co-training a privileged guider with its learner (Li et al., 2026b), whereas our distillation arms freeze the teacher. Relative overgeneralization also exposes the dependence of an agent's preferred action on its partners (Zhao et al., 2024). Our coordination game uses this established difficulty to compare gated and constant-weight updates with an optimal teacher, an exact auxiliary, and a fixed update budget. The resulting local disadvantage and long-run reversal delimit the guarantee that agreement can support; they do not constitute a new account of coordination failure. Table 5 separates these antecedents from our equal-budget and continuous-time comparisons.

Distillation objectives and schedules. Distillation objectives can change optimization dynamics and the resulting policy (Czarnecki et al., 2019). Generalized Knowledge Distillation compares divergences on student-generated language-model sequences (Agarwal et al., 2024), while experiments in partially observed driving find that linear decay can outperform adaptive guidance under severe occlusion (Haklidir, 2026). These results motivate objective and schedule controls alongside the RL baseline.

Selection and continuation utility. Language-model studies examine low divergence on failed trajectories (Ma, 2026), fixed-context teachability (Wang et al., 2026), and guidance evaluated through student continuations (Chen et al., 2026; Qu et al., 2026). CROP matches selection budgets (Li et al., 2026a); Cal-OPD uses retention-matched controls to distinguish attenuation from selection (He et al., 2026). We study these attribution questions in cooperative learning, using exact information references and an optimal-teacher counterexample to delimit what discrepancy can establish.

7 Discussion and conclusion

An optimal teacher and an exact local auxiliary do not guarantee that a discrepancy-weighted imitation step improves on constant weighting at the same budget. Our game isolates why: the weight changes the joint update direction, and its return effect depends on the students' current coordination. Under continuous-time pure JSD imitation, the ordering reverses at the limiting policies. A local return comparison therefore does not determine the endpoint.

This result concerns an additive JSD objective. ADVISOR uses a different coupling of imitation and reinforcement learning and can outperform the combined rule in our benchmarks. The practical lesson is to evaluate a selection rule against the alternative it replaces. Gains over RL establish the value of a complete distillation procedure; they do not identify the incremental value of its discrepancy factor. For the fixed teachers and matched coefficient/LR settings on large and tiny, the adjusted intervals still allow return gains up to about 0.59 and 0.58, respectively, as well as losses. These comparisons leave the increment unresolved; they do not establish that it is practically negligible. Calibration supports the same separation of questions by showing how fitting changes the signal at fixed information.

Limitations. The counterexample uses small pure-imitation steps in a two-agent game; it neither predicts a PPO trajectory nor establishes a practically large performance loss. Calibration uses exogenous observations and small action spaces. Neural comparisons condition on one teacher per task, and selected recipes, architectures, and teacher-training costs limit attribution. Intervals containing zero remain inconclusive. Multiple capable teachers, matched recipes, and prospectively fixed selection rules would provide a stronger empirical test; no such rerun is included here.

Reproducibility statement

The supplement specifies task conditions, training and evaluation protocols, and uncertainty estimates. Episode-level records are available for the 300 local-method confirmation runs; the 100 reference runs and five teacher checkpoints require the original files for full replay (Appendix J.2).

Ethics statement

The reported experiments use simulated cooperative tasks and synthetic hidden-state models. The experiments measure policy return and discrepancy; they do not establish safety for physical multi-agent deployment.

AI use statement

Language models assisted with research design, theoretical models and proofs, simulation and training code, experiment orchestration, result interpretation, literature checks, and manuscript drafting.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 29304–29320. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f514cec81cb148559cf475e7426eed5e-Paper.pdf.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.13649.

Dimitri Bertsekas. Multiagent reinforcement learning: Rollout and policy iteration. *IEEE/CAA Journal of Automatica Sinica*, 8(2):249–272, 2021. doi: 10.1109/JAS.2021.1003814. URL https://web.mit.edu/dimitrib/www/Multiagent_Sinica_2020.pdf.

Dimitri P. Bertsekas and David A. Castañon. Rollout algorithms for stochastic scheduling problems. *Journal of Heuristics*, 5:89–108, 1999. doi: 10.1023/A:1009634810396. URL https://web.mit.edu/dimitrib/www/quiz.pdf.

Clément Bonnet, Daniel Luo, Donal Byrne, Shikha Surana, Sasha Abramowitz, Paul Duckworth, Vincent Coyette, Laurence I. Midgley, Elshadai Tegegn, Tristan Kalloniatis, Omayma Mahjoub, Matthew Macfarlane, Andries P. Smit, Nathan Grinsztajn, Raphael Boige, Cemlyn N. Waters, Mohamed A. Mimouni, Ulrich A. Mbou Sob, Ruan de Kock, Siddarth Singh, Daniel Furelos-Blanco, Victor Le, Arnu Pretorius, and Alexandre Laterre. Jumanji: a diverse suite of scalable reinforcement learning environments in jax. arXiv preprint arXiv:2306.09884v2, 2024. URL https://arxiv.org/abs/2306.09884v2.

Chishui Chen, Yaoyou Fan, Te Sun, Yi Yang, Chenghao Sun, Delin Mao, Hongbo Qiao, Zuowei Zhang, Junxi Wang, Chenxing Sun, Yangen Hu, Lu Pan, Xuyang Liu, and Linfeng Zhang. Look ahead before you distill: Future trajectory validation of teacher guidance for agentic on-policy distillation. *arXiv preprint arXiv:2608.01953*, 2026. URL https://arxiv.org/abs/2608.01953v2.

Wojciech M. Czarnecki, Razvan Pascanu, Simon Osindero, Siddhant Jayakumar, Grzegorz Swirszcz, and Max Jaderberg. Distilling policy distillation. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1331–1340. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/czarnecki19a.html.

Ruan de Kock, Omayma Mahjoub, Sasha Abramowitz, Wiem Khlifi, Callum Rhys Tilbury, Claude Formanek, Andries Smit, and Arnu Pretorius. Mava: a research library for distributed multi-agent reinforcement learning in jax. arXiv preprint arXiv:2107.01460v2, 2023. URL https://arxiv.org/abs/2107.01460v2.

Mehmet Haklidir. When does adaptive guidance help? belief-aware privileged distillation for autonomous driving under partial observability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 907–915, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026W/WAD/html/Haklidir_When_Does_Adaptive_Guidance_Help_Belief-Aware_Privileged_Distillation_for_Autonomous_CVPRW_2026_paper.html.

Qiangqiang He, Jin Li, and MingCai Chen. Calibrating teacher–student discrepancy for on-policy distillation. *arXiv preprint arXiv:2609.21619*, 2026. URL https://arxiv.org/abs/2609.21619v1.

Enhan Li, Junhao He, and Hongyang Du. Crop: Task relevance via counterfactuals for selective on-policy distillation. *arXiv preprint arXiv:2608.13387*, 2026a. URL https://arxiv.org/abs/2608.13387v4.

Yueheng Li, Guangming Xie, and Zongqing Lu. Multi-agent guided policy optimization. In *International Conference on Learning Representations*, 2026b. URL https://arxiv.org/abs/2507.18059.

Zhiyuan Li, Wenshuai Zhao, Lijun Wu, and Joni Pajarinen. Agentmixer: Multi-agent correlated policy factorization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(17):18611–18619, 2025. doi: 10.1609/aaai.v39i17.34048. URL https://doi.org/10.1609/aaai.v39i17.34048.

Guoqing Ma. Outcome-confounded local supervision in on-policy distillation. *arXiv preprint arXiv:2607.23731*, 2026. URL https://arxiv.org/abs/2607.23731v1.

Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS)*, 2021. URL http://arxiv.org/abs/2006.07869.

Zikun Qu, Min Zhang, Mingze Kong, Zhiwei Shang, Zhengyu Chen, Yikun Ban, Shuang Qiu, and Zhongxiang Dai. Spot: Sparse probing and outcome calibration for on-policy distillation. *arXiv preprint arXiv:2608.04419*, 2026. URL https://arxiv.org/abs/2608.04419v2.

Yuda Song, Dhruv Rohatgi, Aarti Singh, and J. Andrew Bagnell. To distill or decide? understanding the algorithmic trade-off in partially observable RL. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 89559–89606. Curran Associates, Inc., 2025. doi: 10.52202/085713-2996. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/81343f8ebe529de8d0bb654af7523184-Paper-Conference.pdf.

Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. Not all disagreement is learnable: Token teachability in on-policy distillation. *arXiv preprint arXiv:2605.26844*, 2026. URL https://arxiv.org/abs/2605.26844v1.

Luca Weihs, Unnat Jain, Iou-Jen Liu, Jordi Salvador, Svetlana Lazebnik, Aniruddha Kembhavi, and Alex Schwing. Bridging the imitation gap by adaptive insubordination. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 19134–19146. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/9fc664916bce863561527f06a96f5ff3-Paper.pdf.

Muning Wen, Jakub Kuba, Runji Lin, Weinan Zhang, Ying Wen, Jun Wang, and Yaodong Yang. Multi-agent reinforcement learning is a sequence modeling problem. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 16509–16521. Curran Associates, Inc., 2022. doi: 10.52202/068431-1201. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/69413f87e5a34897cd010ca698097d0a-Paper-Conference.pdf.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Wenshuai Zhao, Yi Zhao, Zhiyuan Li, Juho Kannala, and Joni Pajarinen. Optimistic multi-agent policy gradient. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61186–61202. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhao24v.html.

A GUIDE TO THE EVIDENCE AND REPORTING SCOPE

The main text starts with an optimal-teacher counterexample, then asks what a fitted discrepancy measures and whether the additional factor improves a complete neural learning procedure. The supplement retains the full set of reported comparisons, including adverse and inconclusive results. Table 3 distinguishes these roles.

Table 3: **Evidence used in the main text and retained in the supplement.** Placement was revised after the outcomes were known. No run was deleted, no result was recomputed under a new exclusion rule, and no new training was run.

Study	Role	Interpretation boundary
Exact calibration	Main: measurement error	Exogenous binary histories
Optimal-teacher game	Main: local counterexample	Pure imitation; small steps
400-run primary study	Main: selected-system comparison	Five tasks retained; recipes differ
Allocation controls	Main: selection diagnostics	Separate 50-step LBF protocol
Policy diagnostics	Supplement: advice utility	Weak warehouse teachers
Prospective advice	Supplement: frozen decision rules	Same weak teachers; one action
Finite updates / refitting	Supplement: local interventions	No full-training inference; refitting audit failed
Fixed-coefficient extension	Supplement: as-run training	Mass-control criterion failed

A.1 TEACHER QUALITY IN THE SUPPLEMENTARY DIAGNOSTICS

The warehouse teachers used in the policy diagnostics have lower complete-policy return than the corresponding RL students (Table 4). The prospective studies use the same teachers. This limits how much useful advice they can test, although complete-policy return does not determine the utility of every teacher action. These studies remain supplementary; the primary comparison retains all five tasks.

Table 4: **Teacher quality limits the supplementary advice studies.** Teacher references use one frozen checkpoint and 128 episodes. Student entries are mean final return over ten seeds, each evaluated for 128 episodes. These complete-policy scores do not determine individual advice utility.

Task	Teacher	RL	Quality-only	Combined
RWARE-small	4.820	18.220	23.179	23.627
RWARE-tiny	2.008	15.020	2.034	2.030

The two historical gate-by-divergence cohorts in Appendix B use 128 final evaluation episodes and a 50-step LBF horizon; the primary study uses 320 episodes and a 100-step horizon. Results are not pooled across these cohorts. The original historical training intervals are pointwise; the primary study adjusts over 35 comparisons. Other supplementary studies state their evaluation and multiplicity procedures separately. The prospective experiments resample student blocks and condition on fixed fitted predictors. Local interventions, scaling results, and failed validation checks retain their respective interpretation limits.

A.2 What the theoretical comparison adds

Local prediction and coordination motivate the analysis but answer different questions. Table 5 makes the distinction explicit. We use the reproducibility heuristic as a hypothesis about allocating an imitation budget; we do not attribute a universal local-return guarantee to ADVISOR or to work on coordination. The new comparison holds teacher quality, auxiliary accuracy, and update size fixed while changing the direction.

Table 5: **Antecedents and the scope of the present comparisons.** The rows distinguish questions and assumptions; they are not claims of exhaustive novelty or of a guarantee made by earlier methods.

Question	Setting or quantity	Relevant conclusion
Local teacher prediction (Weihs et al., 2021)	Privileged teacher, restricted student information, imitation objective	The imitating policy averages over unavailable information; ADVISOR uses an auxiliary to guide imitation/RL weighting.
Coordination (Zhao et al., 2024)	Cooperative policy learning with changing partner behavior	Relative overgeneralization motivates optimistic policy-gradient updates.
This paper: local direction	Optimal teacher, exact auxiliary, additive JSD; equal logit norm or an intrinsic equal-KL construction	The discrepancy-weighted direction can yield strictly lower return than constant weighting after a small step.
This paper: limiting policy	The same two-agent game; positive fixed weights and continuous-time pure JSD imitation	The gated limiting policy has higher return. Local and endpoint orderings differ.

The contribution is the comparison between the two update directions after removing three alternative explanations, together with the reversed endpoint ordering. It does not require the gate to perform poorly at every horizon. The calibration study supplies an exact measurement reference for the fitted auxiliary, and the neural study evaluates selected procedures. Neither study establishes that this game's local mechanism explains PPO performance.

B SUPPLEMENTARY GATE-BY-DIVERGENCE RESULTS

The allocation-control study supplies Table 2. The separate policy-diagnostic study supplies the checkpoints for the local utility and advice experiments. Both use a 50-step LBF horizon and 128 final evaluation episodes. They retain all 19 methods and ten training seeds per method; their estimates are not pooled with the primary study.

Table 6: **All original historical headline comparisons are retained.** Selected combined-JSD minus each independently selected comparator in final return. Entries give mean differences and unadjusted per-contrast 95% bootstrap intervals from ten training seeds per method. Positive favors combined. The two supplementary studies use different teachers and training protocols; each contrast compares methods within its own study.

Task	RL	Quality-only JSD	ADVISOR
Allocation-control study			
LBF-full	0.0000 [0.0000, 0.0000]	0.0016 [0.0000, 0.0039]	0.0000 [0.0000, 0.0000]
LBF-partial	-0.0262 [-0.0355, -0.0164]	0.0047 [-0.0070, 0.0164]	0.0145 [0.0004, 0.0285]
RWARE-small	6.491 [4.652, 8.442]	0.227 [-0.252, 0.720]	-2.613 [-3.126, -2.088]
RWARE-tiny	0.741 [-2.820, 4.828]	0.636 [0.196, 1.082]	-6.314 [-6.806, -5.881]
Policy-diagnostic study			
LBF-full	-0.0020 [-0.0039, 0.0000]	-0.0020 [-0.0039, 0.0000]	0.0090 [0.0039, 0.0141]
LBF-partial	-0.0625 [-0.0730, -0.0512]	-0.0363 [-0.0473, -0.0250]	0.0078 [-0.0105, 0.0258]
RWARE-small	5.407 [3.996, 6.726]	0.448 [-1.085, 1.950]	13.170 [11.944, 14.259]
RWARE-tiny	-12.990 [-13.473, -12.503]	-0.003 [-0.018, 0.012]	0.006 [-0.012, 0.027]

Table 7: Allocation-control study, LBF: final return, mean (sample SD), over ten training seeds and 128 episodes per policy.

Method	LBF full	LBF partial
RL	1.000 (0.000)	0.982 (0.009)
ADVISOR code-style	1.000 (0.000)	0.942 (0.020)
ADVISOR additive FKL	1.000 (0.000)	0.948 (0.020)
ADVISOR additive JSD	1.000 (0.000)	0.918 (0.023)
Raw FKL	1.000 (0.000)	0.934 (0.019)
Raw RKL	1.000 (0.000)	0.899 (0.038)
Raw JSD	0.999 (0.002)	0.952 (0.011)
Dec FKL	1.000 (0.000)	0.894 (0.034)
Dec RKL	1.000 (0.000)	0.919 (0.025)
Dec JSD	1.000 (0.000)	0.952 (0.018)
Quality FKL	1.000 (0.000)	0.938 (0.022)
Quality RKL	1.000 (0.000)	0.946 (0.016)
Quality JSD	0.998 (0.003)	0.952 (0.015)
Combined FKL	1.000 (0.000)	0.921 (0.029)
Combined RKL	0.999 (0.002)	0.950 (0.018)
Combined JSD	1.000 (0.000)	0.956 (0.013)
Constant JSD	1.000 (0.000)	0.952 (0.020)
Shuffled JSD	0.999 (0.002)	0.957 (0.014)
Linear JSD	1.000 (0.000)	0.945 (0.016)

Table 8: Allocation-control study, RWARE: final return, mean (sample SD), over ten training seeds and 128 episodes per policy.

Method	RWARE small	RWARE tiny
RL	4.870 (3.205)	10.902 (6.449)
ADVISOR code-style	13.975 (0.593)	17.958 (0.584)
ADVISOR additive FKL	10.472 (0.569)	10.249 (0.322)
ADVISOR additive JSD	11.591 (0.532)	11.784 (0.636)
Raw FKL	9.490 (0.515)	9.582 (0.304)
Raw RKL	10.386 (0.370)	10.531 (0.386)
Raw JSD	10.884 (0.642)	10.965 (0.476)
Dec FKL	9.682 (0.366)	9.658 (0.535)
Dec RKL	11.173 (0.403)	10.689 (0.383)
Dec JSD	11.311 (0.595)	11.344 (0.457)
Quality FKL	9.716 (0.270)	9.424 (0.398)
Quality RKL	10.888 (0.454)	10.535 (0.368)
Quality JSD	11.134 (0.510)	11.008 (0.531)
Combined FKL	9.946 (0.339)	9.672 (0.215)
Combined RKL	11.020 (0.446)	10.550 (0.299)
Combined JSD	11.362 (0.657)	11.644 (0.538)
Constant JSD	11.338 (0.471)	11.637 (0.579)
Shuffled JSD	11.339 (0.504)	11.372 (0.521)
Linear JSD	12.280 (0.465)	13.675 (0.724)

Table 9: Policy-diagnostic study, LBF: final return, mean (sample SD), over ten training seeds and 128 episodes per policy. Section D examines evaluation stopping.

Method	LBF full	LBF partial
RL	1.000 (0.000)	0.992 (0.006)
ADVISOR code-style	0.989 (0.008)	0.921 (0.026)
ADVISOR additive FKL	1.000 (0.000)	0.979 (0.006)
ADVISOR additive JSD	1.000 (0.000)	0.921 (0.019)
Raw FKL	1.000 (0.000)	0.972 (0.016)
Raw RKL	1.000 (0.000)	0.861 (0.016)
Raw JSD	1.000 (0.000)	0.880 (0.018)
Dec FKL	1.000 (0.000)	0.875 (0.022)
Dec RKL	1.000 (0.000)	0.877 (0.020)
Dec JSD	1.000 (0.000)	0.894 (0.021)
Quality FKL	1.000 (0.000)	0.859 (0.017)
Quality RKL	1.000 (0.000)	0.949 (0.007)
Quality JSD	1.000 (0.000)	0.966 (0.006)
Combined FKL	1.000 (0.000)	0.854 (0.023)
Combined RKL	1.000 (0.000)	0.886 (0.030)
Combined JSD	0.998 (0.003)	0.929 (0.018)
Constant JSD	1.000 (0.000)	0.922 (0.026)
Shuffled JSD	1.000 (0.000)	0.926 (0.020)
Linear JSD	1.000 (0.001)	0.983 (0.010)

Table 10: Policy-diagnostic study, RWARE: final return, mean (sample SD), over ten training seeds and 128 episodes per policy.

Method	RWARE small	RWARE tiny
RL	18.220 (1.463)	15.020 (0.829)
ADVISOR code-style	10.457 (0.630)	2.024 (0.028)
ADVISOR additive FKL	9.228 (0.309)	2.035 (0.014)
ADVISOR additive JSD	7.870 (0.326)	2.032 (0.014)
Raw FKL	8.102 (0.303)	2.034 (0.014)
Raw RKL	13.838 (1.436)	2.026 (0.025)
Raw JSD	9.219 (0.725)	2.038 (0.015)
Dec FKL	6.602 (0.273)	2.036 (0.016)
Dec RKL	14.528 (1.043)	2.034 (0.016)
Dec JSD	7.131 (0.275)	2.027 (0.025)
Quality FKL	22.948 (2.179)	2.034 (0.012)
Quality RKL	22.434 (1.649)	2.033 (0.016)
Quality JSD	23.179 (1.788)	2.034 (0.019)
Combined FKL	23.312 (1.829)	2.030 (0.014)
Combined RKL	24.152 (1.153)	2.030 (0.015)
Combined JSD	23.627 (1.856)	2.030 (0.017)
Constant JSD	23.379 (1.225)	3.811 (5.606)
Shuffled JSD	23.983 (1.557)	3.687 (5.227)
Linear JSD	21.912 (2.005)	2.035 (0.015)

C CONTROLLED LEARNING OUTCOMES

The independent-seed experiment completed all 10,260 runs: 54 fixed conditions, 19 methods and ten training seeds per method and condition. Configuration selection used the separate tuning seeds specified in Section C.1. These conditions vary parameters of one exogenous-history coordination model; they are not 54 independent benchmark tasks. Returns are evaluated exactly for each stochastic policy, while training randomness remains.

C.1 CONTROLLED PROTOCOL

Two agents observe independent noisy emissions from a shared binary hidden-state sequence and act only at its final step. Reward is one exactly when both actions equal the current hidden state. Enumeration preserves correlations between their histories induced by shared past states. Training uses sampled REINFORCE updates, while evaluation computes exact stochastic-policy return. The grid varies observation noise, state-flip probability, history length and teacher error. Its quality weights use exact counterfactual joint values. State histories are exogenous, so this block isolates information and coordination effects rather than action-dependent visitation effects. Conditional on the realized final hidden label z and local histories h_1, h_2, teacher error e gives

$$V_T = (1-e)^2, \qquad V_S = \prod_{i=1}^{2} \pi_i(z \mid h_i), \qquad w^{\text{quality}} = \sigma\left(4\frac{V_T - V_S}{\max(V_T, 0.1)}\right).$$

These are exact last-step joint success probabilities under the all-teacher and all-student product policies. The same scalar quality weight is supplied to both agents, with privileged access to z; it is not a per-action advantage or a learned critic prediction. Controlled learning uses teacher errors $\{0.001, 0.25, 0.999\}$, whereas exact calibration includes endpoints zero and one.

The 54 conditions cross observation noise $\{0, 0.25, 0.5\}$, hidden-state flip probability $\{0, 0.2, 0.5\}$, history length $\{1, 3\}$, and teacher error $\{0.001, 0.25, 0.999\}$. Each run uses 3,000 updates with 256 joint samples per update (768,000 sampled joint histories). Each of 19 methods receives four candidate configurations and tuning seeds 100–102, totaling 12,312 tuning runs. Additive methods vary imitation coefficient $\lambda \in \{0.03, 0.1, 0.3, 1\}$ at learning rate 0.5. RL and code-style ADVISOR vary learning rate over $\{0.1, 0.3, 0.5, 1\}$; code-style ADVISOR retains imitation coefficient one. Selection maximizes mean exact final return over the three tuning seeds, with lower candidate index breaking ties. Confirmation uses fresh seeds 0–9.

Environment and action-uniform random streams are shared across methods within each seed block, while expert-action and gate-shuffling streams are separate. Since histories are exogenous, policy changes do not alter their distribution or consume a different environment stream. Actions can differ under the same uniform draws because policies differ. This common-random-number design justifies resampling whole training-seed blocks for paired intervals; exact policy evaluation removes evaluation noise, not training variation.

C.2 RESULTS AND INTERVENTIONS

The tuned RL comparison is negative. Combined discrepancy and oracle-quality weighting with JSD produces lower final return than tuned RL in every condition. Differences range from -0.008610 to -0.000638 on the unit-return scale. Step-normalized AUC is also lower in every condition, with differences from -0.043777 to -0.002995. Every corresponding pointwise paired 95% interval lies below zero. These intervals use 10,000 bootstrap resamples of the ten common training-seed blocks; they are not simultaneous family-wise intervals.

The search spaces constrain interpretation: RL selects learning rate 1.0 in every condition, whereas additive methods tune their imitation coefficient at learning rate 0.5. Thus the result rejects a practical benefit under this specified search protocol, but does not isolate an adverse effect of the discrepancy factor. Equal trial counts do not make the optimization search spaces identical.

Incremental gate effects are small and mixed. Table 11 reports the range of condition-level mean differences, preserving both signs. The additional discrepancy factor changes final return relative to quality-only JSD by at most 0.000467 in absolute value. Combined and quality-only choose

Table 11: Combined JSD minus each baseline in the controlled confirmation. Entries are the minimum and maximum of 54 condition-level mean-return differences, each estimated from ten training seeds. These ranges are not confidence intervals or averages over independent tasks. Full seed scores and pointwise intervals accompany the released analysis tables.

Baseline	Minimum difference	Maximum difference
Tuned RL	-0.008610	-0.000638
Ungated JSD	-0.000502	$+0.001136$
Quality-only JSD	-0.000467	$+0.000404$
Mean-matched constant	-0.000453	$+0.000709$
Shuffled gate	-0.000460	$+0.000709$

the same imitation coefficient in all 54 conditions. Their comparison therefore provides a cleaner test of the extra factor than the RL comparison. Differences against constant and shuffled controls are also mixed. These controls preserve supervision mass within their own batches; they do not replay identical weights along the combined learner's trajectory, and their independently selected coefficients differ in six conditions.

The ADVISOR contrast depends on the objective and teacher. With inverted teachers, combined JSD exceeds code-style ADVISOR in all 18 conditions, by 0.060105–0.998618. This compares oracle-quality additive weighting with a sampled-action, complementary-RL adaptation of ADVISOR to REINFORCE. It does not show that a learned critic detects harmful teachers, or that ADVISOR is generally inferior. Differences against additive ADVISOR-JSD are small and mixed; with strong teachers, combined JSD loses in all 18 conditions. The large inverted-teacher contrast cannot therefore be attributed to divergence replacement alone.

A separate optimization diagnostic. A completed post-hoc intervention fixes all methods to learning rate 0.5, retains previously selected imitation coefficients, and uses fresh seeds 20–29. RL at learning rate 1.0 supplies an additional anchor on the same new seed blocks. All 10,800 runs complete. At the common rate, combined-minus-RL final-return differences range from -0.001104 to $+0.000888$, with 12 positive and 39 negative pointwise intervals. Increasing RL's rate from 0.5 to 1.0 improves return by 0.000681–0.005370; its return and AUC intervals are positive in every condition. Combined remains below the rate-1 RL anchor in all conditions.

The same-seed intervention identifies an effect of RL's learning rate at this fixed training budget. In every condition, this rate effect exceeds the absolute common-rate combined-method effect for both final return and AUC. It therefore explains a major part of the comparison gap without establishing that rate accounts for every residual difference. The intervention does not test increasing the combined method's rate, establish equivalence, or replace the original tuned comparison. No additional tuning is used to seek a favorable combined-method result.

Discrepancy does not determine the local supervision direction. A post-hoc diagnostic inspects the final RL, ungated-JSD and combined-JSD policies from all 54 conditions and ten seeds, totaling 1,620 checkpoints. For local logits $\ell_{i,h}$, define

$$D_i(h, z) = -\left\langle \nabla_{\ell_{i,h}} J, \nabla_{\ell_{i,h}} \mathrm{JSD}(T_z, \pi_i(\cdot \mid h)) \right\rangle . \tag{9}$$

This is the exact return derivative for an infinitesimal, ungated, unit raw-JSD descent step in one agent's local logit row, with other rows fixed. The reward gradient retains the history probability. We weight records by $P(h_i = h, z)/2$ and bin teacher/local discrepancy divided by $\log 2$, using either the saved auxiliary or its ideal conditional-mean reference. Negative D below -10^{-10} is labeled harmful and positive D above 10^{-10} beneficial; this is a numerical convention, not a practical effect threshold. Central-difference witnesses have maximum directional error below 1.36×10^{-12}.

Table 12 displays a post-hoc counterexample within one condition and method. Beneficial directions occupy the highest discrepancy bin, while harmful directions occupy the lowest bin, for all ten seeds and both score definitions. Thus high discrepancy is neither necessary nor sufficient for the specified negative local derivative in these saved observations. The highest bin has only 1.5625%

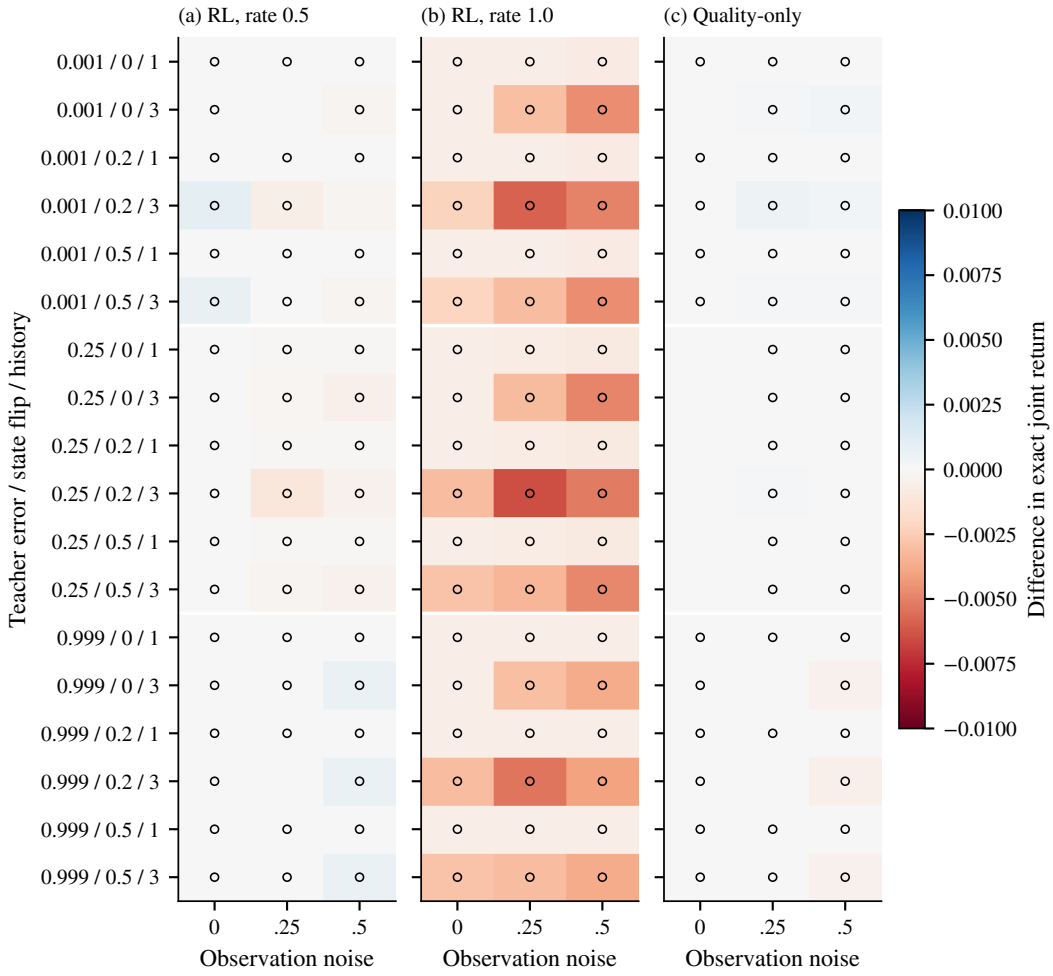


Figure 4: **Optimization rate matters more than the small common-rate effects in this diagnostic.** Each cell is one of the 54 fixed controlled conditions, averaged over ten fresh training seeds. Colors show combined JSD minus (a) RL at rate 0.5, (b) RL at rate 1.0, and (c) quality-only JSD in exact final joint return; no unmeasured cells are interpolated. All distillation methods use learning rate 0.5 and previously selected coefficients. The RL anchors use rates 0.5 and 1.0 on the same seed blocks. Open circles indicate pointwise paired 95% bootstrap intervals excluding zero, without multiplicity adjustment; small effects can still have such markers. Quality values are oracle quantities. This post-hoc diagnostic supplements the original tuned comparison.

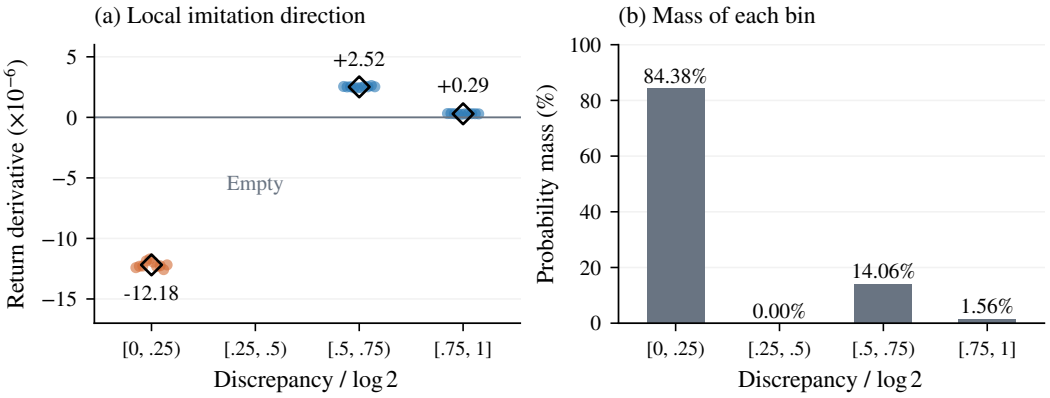


Figure 5: **Exact-reference discrepancy need not order helpful imitation directions.** A post-hoc example from ungated-JSD policies: observation noise 0.25, flip probability 0, history length 3 and teacher error 0.999. (a) Each dot is one training seed's conditional bin-mean return derivative for an infinitesimal raw-JSD descent step in one local logit row; diamonds and labels are means over ten seeds. Negative derivatives indicate harmful directions. (b) Mean probability mass of each bin. The empty bin has zero mass, not zero measured effect. Binning by fitted-auxiliary and exact conditional-mean discrepancy gives identical displayed summaries here. The highest bin contains only 1.5625% of mass. These are local derivatives at saved policies, not finite-update return changes or held-out prediction. Appendix C defines the derivative and probability weighting and retains the full diagnostic coverage.

Table 12: **A post-hoc counterexample to a deterministic discrepancy–harm implication.** Ungated-JSD policies with observation noise 0.25, flip probability 0, history length 3 and teacher error 0.999. Entries agree for saved-auxiliary and ideal-reference score bins. Mass averages all ten training seeds; harmful fractions and directions average conditional seed-bin quantities. Every populated bin contains all ten seeds. Dashes denote an empty bin, not missing runs. Negative/positive directions use the numerical convention defined in the text; magnitudes do not measure finite-training effect sizes.

Normalized JSD bin	Mean mass	Harmful fraction	Mean D $(\times 10^{-6})$
$[0, 0.25)$	0.843750	1	-12.1844
$[0.25, 0.5)$	0.000000	—	—
$[0.5, 0.75)$	0.140625	0	$+2.5154$
$[0.75, 1]$	0.015625	0	$+0.2925$

probability mass in this example. These counterexamples do not rule out statistical association or useful prediction, and do not establish finite-update harm, gate benefit or benchmark behavior. All 25,920 raw directions and 1,296 condition-bin summaries, including empty bins, remain available; no cross-condition significance claim is made.

Table 13: Selected partial-LBF benchmark effects, combined JSD minus reference, under fixed per-slot final evaluation. Each effect uses ten training seeds per method and 10,000 independent-method bootstrap draws; intervals are pointwise 95%. This post-hoc sensitivity does not change original tuning or monitoring AUC.

Reference	Difference	95% interval
RL	-0.067188	$[-0.078906, -0.054297]$
Raw JSD	0.046484	$[0.030078, 0.062891]$
Quality-only JSD	-0.039453	$[-0.051172, -0.026172]$
Constant JSD	0.003906	$[-0.014062, 0.023047]$
Shuffled JSD	0.002734	$[-0.015234, 0.021494]$
Linear JSD	-0.057422	$[-0.070703, -0.043740]$
Additive-FKL ADVISOR	-0.054297	$[-0.065625, -0.041406]$

D SENSITIVITY TO EVALUATION COMPLETION RULES

The policy-diagnostic evaluator steps eight environments and retains the first N completed episode returns. LBF can terminate upon success before its 50-step limit, so finite stopping can favor shorter episodes. Original logs lack episode lengths and slot identities. RWARE uses a fixed 500-step horizon with no inactivity termination; its evaluation counts are multiples of eight, so this particular stopping issue does not affect those records.

We declare a post-hoc final-checkpoint sensitivity covering every affected LBF cohort: 380 benchmark students, 60 pure-imitation students, 90 teacher-stage students, two final teacher references and three stage-teacher evaluation streams. These 535 evaluations contain 530 student checkpoints and five teacher streams, rather than 535 independent training seeds. Each evaluation retains exactly 16 episodes from each of eight slots, totaling 128. All slots continue stepping and sampling after their individual quota to preserve the original batched RNG sequence. Simultaneously recorded first-128 pooled completions exactly reproduce every original return array. Checkpoints, evaluation seeds and source identities remain fixed; no policy is retrained or selected again.

Across the sensitivity analysis's 73 declared method/stage contrasts, observed effect signs and interval zero-exclusion categories remain unchanged between its original and quota calculations. This does not mean numerical effects are unchanged: the partial-LBF combined-minus-RL effect shifts by -0.004688 $[-0.007422, -0.001563]$, using paired original/quota observations within each checkpoint. Seven contrast-shift intervals exclude zero, all negatively on partial-LBF benchmark comparisons. Full-LBF estimates often sit at discrete zero boundaries, so threshold categories should not be read as equivalence or safety evidence.

The stability statement also excludes student-versus-teacher ordering. Partial-LBF combined-JSD mean changes from 0.929297 to 0.923828, while the final teacher reference remains 0.925781. The student-minus-teacher point difference changes from positive to negative. Thus a descriptive claim that this student exceeds its evaluated teacher does not survive. Teacher references come from one training seed and a finite evaluation stream.

E SUPPLEMENTARY EXPERIMENTAL PROTOCOLS

The benchmark tasks use cooperative team return (Papoudakis et al., 2021; Bonnet et al., 2024; de Kock et al., 2023). LBF normalizes complete collection to return one; RWARE sums deliveries. Table 14 specifies the four conditions used in the supplementary training studies. The primary five-task comparison is described in Appendix J.

Training and selection. Student actors and auxiliaries have two hidden layers of width 128. Actors share parameters across agents and receive local observations and agent identity. Each student uses 19,988,480 joint environment steps, 128-step rollouts, four PPO epochs, and four minibatches per epoch. Discount and GAE parameters are 0.99 and 0.95; entropy and value coefficients are 0.01 and 0.5. Monitoring uses 32 episodes per point, and final evaluation uses 128 episodes. The dis-

Table 14: Supplementary task conditions. RWARE has three shelf columns, column height eight, and sensor range one; tiny/small use one/two shelf rows. LBF has an 8×8 grid, two food items, player levels one–two, and forced cooperation. Warehouse size changes are not isolated agent-count interventions.

Condition	Agents	Horizon	LBF sight	Requests
LBF full	2	50	8	N/A
LBF partial	2	50	2	N/A
RWARE small	4	500	N/A	4
RWARE tiny	2	500	N/A	2

crepancy exponent is 2, the quality scale is 4, and the ADVISOR exponent is 4. The update rules are given in Appendix F.

Each task/method receives four candidates and three tuning seeds (100–102). Additive candidates use imitation coefficients 0.03, 0.1, 0.3, and 1 at learning rate 3×10^{-4}. RL and ADVISOR use learning rates 10^{-4}, 3×10^{-4}, 10^{-3}, and 3×10^{-3}, with imitation coefficients zero and one. Selection maximizes mean final tuning return, breaking ties by candidate order. Tuning uses 2,097,152 steps and 64 final episodes; confirmation uses ten separate seeds (0–9). Candidate counts are equal, but search spaces and training budgets differ between tuning and confirmation.

Allocation-control study. A frozen autoregressive MAT teacher supplies targets conditioned on student action prefixes. Local actors use ReLU activations and valid-action masks; critics use local observations. Each update collects 16,384 joint transitions. PPO clips are 0.1, 0.1, 0.2, and 0.05 for LBF-full, LBF-partial, RWARE-small, and RWARE-tiny; the corresponding gradient clips are 5, 5, 10, and 5. Each teacher is trained for 19,988,480 steps with seed 100.

Policy-diagnostic study. These checkpoints use frozen privileged MLP teachers, tanh actors, and centralized critics. Actors do not use action masks. Each update collects 2,048 joint transitions, with PPO clip 0.2 and gradient clip 10. Teacher training uses seed 100 and 4,999,168 steps on each LBF task, 7,999,488 on RWARE-small, and 6,000,640 on RWARE-tiny. The warehouse teachers' limited returns constrain the advice experiments (Table 4).

For normalized supplementary metrics, student return is divided by the fixed teacher-reference mean over 128 episodes, with lower anchor zero and no clipping above one. Student bootstraps hold this reference fixed and exclude teacher-population uncertainty. Timing is not used to rank methods: the local RL arm retains teacher and auxiliary computations despite its zero imitation coefficient.

E.1 Exact calibration: grid and fitting protocol

The grid uses $p \in \{0, 0.25, 0.5\}$, $f \in \{0, 0.2, 0.5\}$, and $e \in \{0, 0.25, 1\}$. Auxiliary classes retain no observation, the latest observation, or the full history, with duplicate classes removed when $H = 1$. Training sizes are 64, 256, 1,024 and 4,096; optimization budgets are 0, 10, 100 and 1,000 updates. Ten data replicates yield 21,600 calibration cases across 54 information/teacher conditions. Each case uses 8,192 held-out samples drawn independently of training. A direct empirical conditional mean provides a finite-data control. Cross-entropies clip logarithm arguments at 10^{-12}.

Auxiliary logits start at zero, giving a uniform policy in every table row. Each observed suffix row receives a count-normalized cross-entropy gradient step with learning rate 0.2: its logit vector is updated by subtracting $0.2(q - \widehat{q})$, where $\widehat{q}$ is that row's empirical mean teacher distribution. Unobserved rows remain uniform. This gives equally scaled row updates rather than visitation-weighted global gradient descent. Capacity and update controls at a fixed condition and sample size share training draws; different sample sizes use separate streams. Held-out draws are independent of fitting draws and shared across capacity and update comparisons.

F UPDATE ORDER AND GRADIENT BOUNDARIES

The auxiliary has its own optimizer and does not share actor parameters. The quality proxy uses rollout-time student values even when the discrepancy weight is refreshed. Teacher targets and weights are detached from the student gradient.

Additive training. Freeze the teacher and collect student rollouts. Compute advantages and teacher targets on the visited states, using student action prefixes for autoregressive targets. Reserve 20% of transitions for auxiliary evaluation. At each minibatch, form detached weights from current auxiliary predictions, update the student with policy, entropy, and weighted imitation losses, and fit the auxiliary on the training split. The critic uses ungated value regression. Repeat to the prescribed budget and evaluate on a separate stream; deployment uses only the local actor.

The policy-diagnostic checkpoints hold weights fixed through a rollout's PPO epochs, then take four auxiliary updates. ADVISOR instead fits all labelled rows and refreshes sampled-action weights per minibatch, with complementary policy and entropy weights. Pure imitation omits PPO and value regression.

Additive adaptations and allocation controls. Additive ADVISOR uses $w_i = \exp[-4\,\mathrm{KL}(T_i\|q_i)]$ with natural logarithms for both forward-KL and JSD imitation. Linear decay uses $w = 1 - u/U$ at rollout update $u = 0, \ldots, U-1$. Constant and shuffled controls replace the current combined-weight array by its mean or permute its transition–agent entries. Both preserve that array's mean at each weight refresh, not a common mass across independently trained trajectories.

F.1 AN ILLUSTRATIVE CONTINUATION-POLICY EXAMPLE

This analytic construction illustrates why complete-policy value and one-action utility are distinct. It is not a benchmark measurement, a new gating algorithm, or a finite-update result.

Consider two agents with binary actions in a fully observed cooperative game, with undiscounted team return and at most two decisions. At the initial state s_0, joint action $(0,0)$ terminates with reward $1/2$, joint action $(1,1)$ gives reward zero and moves to u, and either mixed joint action terminates with reward zero. At u, every action terminates: $(1,1)$ gives reward one, $(0,0)$ gives reward r, and mixed actions give zero. The two games differ only in $r \in \{1/4, 3/4\}$; all rewards lie in $[0,1]$.

Fix the teacher to choose $(1,1)$ at both states and the student to choose $(0,0)$ at both states. Both policies factorize across agents. Give each local auxiliary the exact teacher distribution, so teacher–auxiliary discrepancy is zero, including at s_0. The teacher is optimal in both games. Exact root-state values are

$$V_T(s_0) = 1, \qquad V_S(s_0) = 1/2.$$

Thus teacher and student action distributions, exact policy-value gap, and discrepancy at s_0 agree across the two games. Following the teacher for the first joint action instead reaches u, after which the fixed student chooses $(0,0)$ and receives r. Consequently

$$U(s_0) = Q_S(s_0,(1,1)) - V_S(s_0) = r - 1/2 = \begin{cases} -1/4, & r = 1/4, \\ +1/4, & r = 3/4. \end{cases}$$

The same positive complete-policy value gap therefore coexists with either sign of the teacher-first intervention. Exact critics and an exact auxiliary do not remove this ambiguity for a rule using only those root-state quantities. This is a pointwise counterexample across two games, not a claim that discrepancy lacks statistical predictive value in a fixed task. It also says nothing about updating the student's continuation policy: changing both decisions could recover the teacher's higher return.

G LARGER-TASK AND LARGER-MODEL OUTCOMES

These completed historical extensions test transferred procedures and execution at larger scale. They are separate from the primary confirmations in Appendices J–K and from language-model training. Their smaller seed counts limit uncertainty coverage.

G.1 TRANSFER TO AN EIGHT-AGENT WAREHOUSE

The official Jumanji large-8ag task uses three shelf rows, five shelf columns, column height eight, request queue eight, sensor radius one and horizon 500. Nine small-task recipes transfer without retuning, with five student seeds each, 19,988,480 joint steps and 128 final evaluation episodes. All 45 runs, including the four full-budget pilot runs counted once, are retained. A single seed-100 teacher is shared. Geometry, team size and request queue change together, so this is not an isolated agent-count intervention.

Table 15: All nine transferred large-8ag procedures. Mean (sample SD) over five student training seeds. AUC integrates 62 monitoring points and divides by the joint-step budget. Teacher and recipes remain fixed.

Method	Final return	AUC
RL	5.848 (0.193)	2.649 (0.289)
ADVISOR code-style	12.620 (0.742)	11.222 (0.189)
Raw JSD	10.592 (0.451)	8.213 (0.089)
Dec JSD	10.269 (0.324)	8.075 (0.352)
Quality JSD	10.213 (0.630)	8.247 (0.202)
Combined JSD	10.644 (0.226)	8.036 (0.162)
Constant JSD	10.205 (0.603)	8.067 (0.195)
Shuffled JSD	10.363 (0.499)	8.093 (0.111)
Linear JSD	9.177 (0.923)	8.288 (0.238)

Combined JSD improves over RL by 4.795 in final return, with pointwise 95% interval $[4.564, 5.023]$, but trails code-style ADVISOR by 1.977 (difference interval $[-2.544, -1.350]$). Its final-return comparisons with raw and quality-only JSD have intervals including zero. Combined-minus-quality AUC is -0.211 $[-0.419, -0.009]$. Thus a gain over RL does not establish added gate value or superiority to the stronger imitation baseline. Intervals use 10,000 independent training-run bootstrap resamples per method, without multiplicity adjustment; five-seed coverage is limited. ADVISOR retains learning rate 0.003, while JSD recipes retain rate 0.0003 and coefficient one. These are transferred system comparisons, not interventions on the gate alone.

G.2 LARGER NETWORKS

A separate large-8ag experiment uses a MAT teacher with embedding width 256, four blocks and eight heads, and student modules with two hidden layers of width 1024. The models have 4,376,986 teacher parameters and 1,131,525 actor parameters; actor, critic and auxiliary together contain 3,390,475 trainable parameters. RL, quality-only JSD and combined JSD share one larger frozen teacher, with three student seeds each and fixed learning rate 0.0003. Imitation coefficient is one for the distillation arms and zero for RL.

Each learner replicates parameters across four GB300 GPUs within one node; nine independent learners use nine nodes. This is synchronous data parallelism, not one model sharded across nodes. Sixteen environments per device and update batch, two update batches and rollout length 128 give 16,384 global joint steps per update. Each run retains 19,988,480 joint steps, 62 monitoring points and 128 final episodes. Final parameter/optimizer replicas match, and evaluations use restored checkpoints. Feasibility runs are excluded from the nine-run comparison.

Combined-minus-quality final return is -0.057 $[-0.648, 0.534]$, while its AUC difference is -0.097 $[-0.161, -0.012]$. These pointwise 95% intervals use 10,000 independent method-level training-run resamples; three seeds provide severely limited coverage. The study establishes execution of

Table 16: All larger-model procedures, mean (sample SD) over three training seeds. GPU replicas and evaluation episodes are not independent training seeds. Two RL runs have zero final return; both remain in the summary.

Method	Final return	AUC
RL	1.810 (3.135)	0.377 (0.653)
Quality JSD	6.047 (0.444)	5.336 (0.008)
Combined JSD	5.990 (0.485)	5.240 (0.079)

the specified larger networks, without evidence of an additional combined-gate advantage. Teacher, architecture and distributed batching change relative to the smaller-model study, preventing a causal capacity or efficiency claim. Ungated JSD is absent from this three-method grid, so the distillation-versus-RL comparison cannot identify the contribution of quality weighting. These results do not support a hardware-efficiency comparison.

H HELD-OUT TEACHER-ACTION UTILITY PREDICTION

The diagnostic uses 120 saved final policy-diagnostic checkpoints from four tasks, three student procedures (RL, quality-JSD and combined-JSD) and ten student seeds. Each checkpoint uses eight reset streams. Reset and continuation seeds are shared across tasks and student procedures within a training seed. The student takes 13 steps before the environment is copied. Each copied state supplies 32 paired continuations: one teacher joint action followed by the unchanged student, versus the student throughout. Returns use discount 0.99 and the remaining episode horizon. Common random streams pair the two arms; the first pair is replayed exactly, and equal first actions must yield identical trajectories. The 960 visits produce 30,720 pairs, or 61,440 arm trajectories before replay witnesses. All 120 checkpoints are included.

Fitting and evaluation. Each task has 120 fit visits from student seeds 0–4 and 120 test visits from seeds 5–9, equally weighting the three procedures. Five predictors are fixed: a fit-set constant, value gap alone, teacher–auxiliary JSD alone, teacher and student entropy, and value gap plus JSD. Feature means and population standard deviations come only from the fit split. Ridge minimizes summed squared error plus the squared coefficients with penalty 1 and an unpenalized intercept. No interactions, feature clipping, tuning or test refitting are performed. Parameters are durably saved before test scoring. Each eight-visit checkpoint MSE is averaged within its training-seed block across the three procedures; five blocks determine each test MSE. All 3,125 ordered paired bootstrap samples of these five blocks yield linear-interpolated percentile 95% intervals. The primary comparison is value-plus-discrepancy minus value-only. All other contrasts are exploratory, and the intervals are not multiplicity-adjusted.

Table 17: All prespecified utility predictors. MSE change is relative to value-only, in percent; positive is worse. The last column is the unadjusted pointwise 95% interval for the raw MSE difference. Scientific notation uses ∈ for powers of ten. No raw MSE is pooled across tasks.

Task	Predictor	MSE	Change(%)	Difference interval
LBF-full	Constant	4.52133e-05	+2.086	[-1.0786e-06, 3.3457e-06]
	Value	4.42895e-05	0.000	[0, 0]
	JSD	4.53676e-05	+2.434	[-9.6855e-07, 3.6384e-06]
	Entropy	4.51312e-05	+1.900	[-2.7243e-06, 4.8545e-06]
	Value+JSD	4.43784e-05	+0.201	[-2.2146e-07, 3.964e-07]
LBF-partial	Constant	0.00128224	-4.688	[-0.00013128, 1.0626e-06]
	Value	0.00134531	0.000	[0, 0]
	JSD	0.00127554	-5.186	[-0.00013457, -1.1189e-05]
	Entropy	0.00125966	-6.367	[-0.00019217, 2.1394e-05]
	Value+JSD	0.00134311	-0.163	[-6.0677e-06, 1.9006e-06]
RWARE-small	Constant	0.0192978	+0.209	[-0.00039733, 0.00058052]
	Value	0.0192576	0.000	[0, 0]
	JSD	0.0201494	+4.631	[8.0983e-05, 0.0014808]
	Entropy	0.0199594	+3.644	[-0.00049274, 0.0018963]
	Value+JSD	0.0200065	+3.889	[7.6566e-05, 0.0014001]
RWARE-tiny	Constant	0.00329774	-1.511	[-0.00034696, 0.00016552]
	Value	0.00334834	0.000	[0, 0]
	JSD	0.00329465	-1.603	[-0.00034934, 0.00016381]
	Entropy	0.00314018	-6.217	[-0.00063928, 0.00019375]
	Value+JSD	0.00335301	+0.140	[1.9152e-06, 7.5099e-06]

Interpretation and coverage. The primary predictor shows no supported reduction across the four tasks. Its RWARE-small and RWARE-tiny increases are respectively 3.89% and 0.14%, so interval zero exclusion should not be confused with a uniformly large practical effect. This is a fixed linear-predictor protocol under one frozen teacher per task, with new student seeds but no held-out tasks or

Table 18: Test-label diagnostics over 120 visits per task. MC SE is the estimated standard error of a 32-pair mean. All-zero denotes that all 32 paired differences were zero, not proof that the population effect is zero.

Task	Mean utility	Median MC SE	All-zero (%)
LBF-full	-0.00251189	0.000273788	44.17
LBF-partial	0.00275514	0.00486901	0.00
RWARE-small	-0.0447926	0.0903962	0.00
RWARE-tiny	-0.0117854	0.00829715	2.50

teachers. The labels have Monte Carlo error; zero empirical variance is not proof of an exact target. We do not subtract estimated noise variance or select visits by observed effect. The result does not establish a general inability to predict utility, nor determine the effect of a finite neural update.

I FINITE NEURAL UPDATES WITH MATCHED GLOBAL IMITATION MASS

This study distinguishes teacher-action utility (Appendix H) from the effect of a finite neural update. We use 120 saved final checkpoints: four policy-diagnostic tasks, three historical training procedures (RL, quality-JSD and combined-JSD), and ten training seeds each. One fresh 2,048-transition joint rollout per checkpoint supplies four updates: PPO alone, quality, combined, and globally mass-matched quality. The updates share initial actor, critic and Adam states, teacher distributions and ordered minibatches (four epochs, four minibatches). Each arm uses its own imitation weights computed once from common pre-update signals. The learning rate is reset to 3×10^{-4} for every arm while retaining the stored Adam moments; the imitation coefficient is 1 for quality, combined and mass-matched quality, and 0 for PPO. An unchanged actor is the fifth evaluation arm. Each actor receives 128 retained stochastic-policy episodes under common RNG initialization and equal per-slot quotas, for 76,800 retained episodes overall.

The primary contrast is combined minus mass-matched quality in undiscounted return. Within each training seed, we equally average the three historical procedures, then average the ten seed blocks. The frozen analysis uses 50,000 block-bootstrap draws and 98.75% percentile intervals for the four primary contrasts (Bonferroni family coverage target 95%). These finite-sample intervals are approximate. Global mean imitation weight is matched up to floating-point tolerance (absolute 10^{-7}, relative 10^{-6}); minibatch mass, gradient norm and post-update occupancy need not be. Table 19 gives the primary results and Table 20 the absolute returns.

No primary adjusted interval excludes zero. LBF-full and RWARE-tiny have identical combined and matched-quality return vectors at all 30 checkpoints, and all combined actor displacements are positive. A post-hoc read-only comparison of saved actor tensors separately confirms that combined and mass-matched-quality actor parameters differ in all 120 conditions. Whole-state fingerprints also include critic and optimizer state, so fingerprint inequality alone would not establish this actor difference. Zero empirical return intervals do not prove policy or population-return equivalence. The RWARE-small seed-block contrast ranges from -0.2109375 to $+0.109375$, much larger than its mean.

Table 19: Primary finite-update contrasts. Positive differences favor combined. Adjusted intervals include zero in every task.

Task	Difference	Seed SD	98.75% interval
LBF-full	0.000000	0.000000	[0.000000, 0.000000]
LBF-partial	0.001432	0.002413	[-0.000391, 0.003125]
RWARE-small	0.002865	0.084478	[-0.071875, 0.053906]
RWARE-tiny	0.000000	0.000000	[0.000000, 0.000000]

Table 20: Mean return after one update, or without updating. Each entry averages 30 checkpoint means with equal procedure weights.

Task	Unchanged	PPO	Quality	Combined	Matched quality
LBF-full	0.998568	0.998568	0.998568	0.998568	0.998568
LBF-partial	0.963672	0.962240	0.960026	0.961719	0.960286
RWARE-small	21.701563	21.859896	21.867969	21.877344	21.874479
RWARE-tiny	6.316146	6.377865	6.369010	6.369010	6.369010

Numerical correction. An initial campaign encountered nine mass-matching failures caused by FP32 reduction error. All 120 conditions were rerun after computing global scaling in FP64 and casting back, with the original tolerance unchanged. Only the complete corrected campaign enters the reported results.

Scope. One rollout and evaluation realization per final checkpoint does not separately estimate within-checkpoint intervention variance. Episodes are not additional training replicates. There is one

fixed teacher per task; no inference over teacher populations, earlier training, or complete training runs follows. Secondary contrasts below use unadjusted pointwise intervals and do not replace the primary result. Per-procedure contrasts and update diagnostics are retained in the accompanying machine-readable report.

Table 21: All secondary finite-update contrasts with exploratory pointwise 95% intervals. C: combined; M: mass-matched quality; Q: quality; P: PPO; U: unchanged.

Task	Contrast	Difference	95% interval
LBF-full	$C - Q$	0.000000	[0.000000, 0.000000]
LBF-full	$C - P$	0.000000	[0.000000, 0.000000]
LBF-full	$M - P$	0.000000	[0.000000, 0.000000]
LBF-full	$Q - P$	0.000000	[0.000000, 0.000000]
LBF-full	$P - U$	0.000000	[0.000000, 0.000000]
LBF-full	$Q - U$	0.000000	[0.000000, 0.000000]
LBF-full	$C - U$	0.000000	[0.000000, 0.000000]
LBF-full	$M - U$	0.000000	[0.000000, 0.000000]
LBF-partial	$C - Q$	0.001693	[-0.000521, 0.004557]
LBF-partial	$C - P$	-0.000521	[-0.003255, 0.002474]
LBF-partial	$M - P$	-0.001953	[-0.004818, 0.001302]
LBF-partial	$Q - P$	-0.002214	[-0.005469, 0.001172]
LBF-partial	$P - U$	-0.001432	[-0.005859, 0.002604]
LBF-partial	$Q - U$	-0.003646	[-0.009635, 0.002995]
LBF-partial	$C - U$	-0.001953	[-0.008073, 0.004427]
LBF-partial	$M - U$	-0.003385	[-0.009896, 0.003255]
RWARE-small	$C - Q$	0.009375	[-0.029687, 0.038542]
RWARE-small	$C - P$	0.017448	[-0.038281, 0.068490]
RWARE-small	$M - P$	0.014583	[-0.036458, 0.065104]
RWARE-small	$Q - P$	0.008073	[-0.062760, 0.072917]
RWARE-small	$P - U$	0.158333	[0.013021, 0.289583]
RWARE-small	$Q - U$	0.166406	[-0.015892, 0.343490]
RWARE-small	$C - U$	0.175781	[0.008594, 0.327865]
RWARE-small	$M - U$	0.172917	[0.004688, 0.323704]
RWARE-tiny	$C - Q$	0.000000	[0.000000, 0.000000]
RWARE-tiny	$C - P$	-0.008854	[-0.025781, 0.000260]
RWARE-tiny	$M - P$	-0.008854	[-0.025781, 0.000260]
RWARE-tiny	$Q - P$	-0.008854	[-0.025781, 0.000260]
RWARE-tiny	$P - U$	0.061719	[-0.023698, 0.129948]
RWARE-tiny	$Q - U$	0.052865	[-0.031250, 0.122656]
RWARE-tiny	$C - U$	0.052865	[-0.031250, 0.122656]
RWARE-tiny	$M - U$	0.052865	[-0.031250, 0.122656]

J COMPLETE PRIMARY CONFIRMATION

This separate confirmation grid contains all 400 runs: five tasks, eight procedures and ten fresh training seeds (7400–7409). Each selected configuration receives 19,988,480 joint steps and 320 final-policy evaluation episodes. The LBF horizon is 100; these results are not pooled with historical 50-step LBF or supplementary policy-diagnostic results. The warehouse tasks include the eight-agent large task. Tuning choices are locked before confirmation. The analysis specification was fixed after launch but before complete-grid aggregation; it is not a prelaunch preregistration. Data availability is described in Appendix J.2.

Comparison scope. Algorithms share task definitions and student transition budgets, but their architectures, optimizers and teacher-training costs differ. The reproduced MAGPO/MAPPO actors are recurrent, whereas local actors are feedforward; matching sensor ranges therefore does not match policy memory. MAPPO uses a centralized critic, while its actor uses local observations and recurrent state. Local methods use one fixed teacher per task; student-seed intervals exclude teacher-population uncertainty. The local RL arm retains auxiliary computation despite zero imitation coefficient. MAPPO and MAGPO refer to the pinned reproduced implementations, not published scores or evidence of exhaustive SOTA coverage. ADVISOR denotes the code-style baseline used throughout this paper. These are independently selected-recipe comparisons: combined and quality select different imitation coefficients on RWARE-small (0.3 versus 0.1), and different learning rates on LBF-full (10^{-4} versus 3×10^{-4}) and LBF-partial (3×10^{-4} versus 10^{-4}). The large and tiny warehouse comparisons match both settings; the five-task grid therefore does not uniformly isolate the discrepancy factor.

Table 1 reports all final-policy means and standard deviations.

Uncertainty. We independently resample ten training runs within each task and method, using 100,000 bootstrap draws. Evaluation episodes are averaged within each trained policy. Table 22 retains all 35 primary combined-minus-comparator contrasts, with Bonferroni percentile bounds across that family (nominal per-contrast level 99.857%). These approximate intervals have limited small-sample coverage; including zero does not establish equivalence. The supplementary data also retain all 35 secondary best-policy contrasts; secondary endpoints do not replace final-policy conclusions.

Observed comparisons. Combined exceeds local RL on RWARE-small and RWARE-large, and reproduced MAPPO on all three warehouse tasks. It falls below MAGPO on LBF-partial, RWARE-large and RWARE-tiny, and below ADVISOR on RWARE-small and RWARE-tiny, under the stated adjusted intervals. Every primary combined-minus-quality interval includes zero. Thus gains over RL do not demonstrate additional discrepancy-gate value or general superiority over stronger baselines. LBF-full is near its return ceiling.

Table 22: All primary contrasts from the complete primary grid. Positive values favor combined. Adjusted bounds cover the full 35-contrast family.

Task	Comparator	Difference	Adjusted interval
LBF-full	Advisor	-0.0006	$[-0.0031, 0.0000]$
	Linear	-0.0006	$[-0.0031, 0.0000]$
	MAGPO	-0.0006	$[-0.0031, 0.0000]$
	MAPPO	-0.0006	$[-0.0031, 0.0000]$
	Quality	-0.0006	$[-0.0031, 0.0000]$
	Raw	-0.0006	$[-0.0031, 0.0000]$
	RL	-0.0003	$[-0.0028, 0.0013]$
LBF-partial	Advisor	0.0053	$[-0.0011, 0.0109]$
	Linear	0.0025	$[-0.0047, 0.0095]$
	MAGPO	-0.0091	$[-0.0147, -0.0045]$
	MAPPO	-0.0097	$[-0.0152, -0.0052]$
	Quality	0.0052	$[-0.0032, 0.0134]$
	Raw	0.0109	$[0.0008, 0.0208]$
	RL	0.0031	$[-0.0067, 0.0130]$
RWARE-large	Advisor	-0.6294	$[-1.2546, 0.0122]$
	Linear	0.3762	$[-0.2019, 0.9342]$
	MAGPO	-1.6863	$[-2.3549, -1.1467]$
	MAPPO	10.9191	$[10.6058, 11.2044]$
	Quality	0.1837	$[-0.2086, 0.5883]$
	Raw	0.0938	$[-0.3914, 0.5711]$
	RL	6.9650	$[5.1080, 9.7091]$
RWARE-small	Advisor	-2.2125	$[-2.5666, -1.8718]$
	Linear	-0.3147	$[-0.8344, 0.1432]$
	MAGPO	-1.6569	$[-2.6363, 0.0414]$
	MAPPO	12.1784	$[11.4622, 12.6219]$
	Quality	-0.0544	$[-0.6777, 0.5877]$
	Raw	-0.0472	$[-0.5117, 0.4225]$
	RL	7.0772	$[4.7475, 10.3149]$
RWARE-tiny	Advisor	-2.7434	$[-3.6389, -1.8075]$
	Linear	-0.4309	$[-1.2389, 0.4994]$
	MAGPO	-3.4922	$[-4.6919, -2.5508]$
	MAPPO	10.9078	$[3.7279, 14.5828]$
	Quality	-0.0078	$[-0.6523, 0.5757]$
	Raw	0.5988	$[0.0267, 1.0858]$
	RL	0.9784	$[-1.5356, 6.9842]$

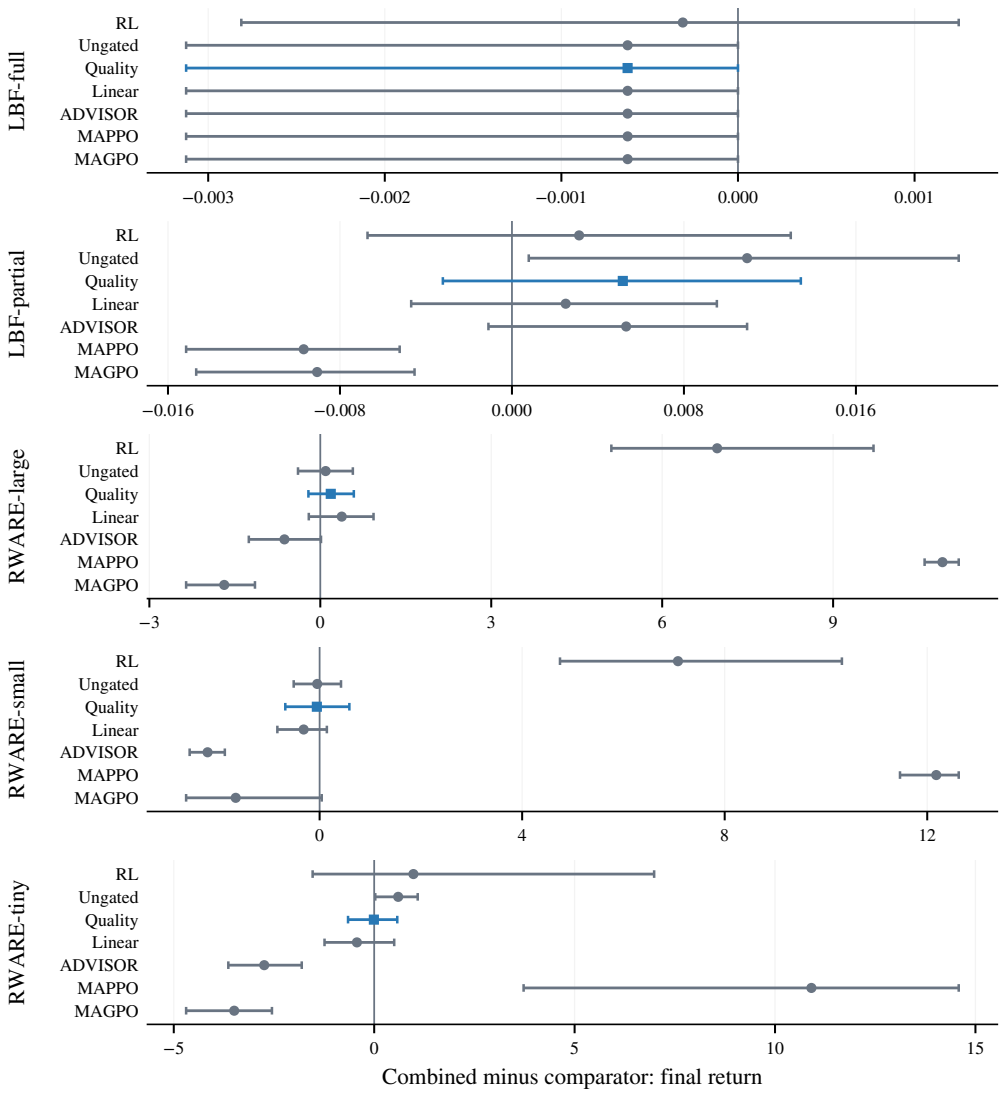


Figure 6: All 35 primary final-policy contrasts from the complete primary grid. Positive differences favor combined; negative differences favor the named comparator. Points are differences between means over ten independent training runs per method and task, with 320 evaluation episodes per trained policy. Intervals resample training runs. Local distillation arms condition on one fixed teacher per task and exclude teacher-population uncertainty; MAGPO co-trains its guider. Bars are the stored approximate 99.857% percentile bootstrap intervals, targeting 95% Bonferroni coverage across all 35 comparisons. Axes use task-specific return scales; horizontal distances must not be compared across panels. Blue squares identify quality-only comparisons; zero-containing intervals do not establish equivalence. All methods share task definitions and student transition budgets, but architectures, optimizers, selected recipes and teacher-training costs differ. ADVISOR is the code-style adaptation; MAGPO/MAPPO are the pinned reproduced implementations. These comparisons neither isolate the discrepancy factor uniformly nor establish general SOTA superiority.

J.1 Effect ranges and fixed-teacher references

Table 23: **Pointwise and family-adjusted uncertainty for the added factor.** Combined minus quality-only final return. Both columns reuse the original 100,000-draw bootstrap; the primary family remains all 35 contrasts. † indicates matching selected coefficients and learning rates.

Task	Difference	Pointwise 95%	Adjusted 99.857%
LBF-full	-0.0006	$[-0.0019, 0.0000]$	$[-0.0031, 0.0000]$
LBF-partial	+0.0052	$[0.0000, 0.0103]$	$[-0.0032, 0.0134]$
RWARE-large†	+0.1837	$[-0.0625, 0.4366]$	$[-0.2086, 0.5883]$
RWARE-small	-0.0544	$[-0.4478, 0.3459]$	$[-0.6777, 0.5877]$
RWARE-tiny†	-0.0078	$[-0.3956, 0.3653]$	$[-0.6523, 0.5757]$

Both the pointwise and adjusted intervals include zero for the two comparisons that match coefficient and learning rate. Their widths permit effects in either direction. The pointwise columns are a descriptive sensitivity view of the same stored bootstrap draws, not a smaller replacement hypothesis family. Ten runs per method limit coverage accuracy, and increasing the number of bootstrap draws does not add independent training evidence. No equivalence margin was specified.

Table 24: **The five teachers fixed by the primary confirmation study.** Descriptive final-policy means from retained teacher completion records. Episode counts differ from one another and from student confirmation.

Task	Teacher seed	Evaluation episodes	Mean return
LBF-full	100	320	1.0000
LBF-partial	100	320	1.0000
RWARE-large	100	128	15.9453
RWARE-small	100	128	15.1641
RWARE-tiny	100	128	15.5000

Each local task uses one fixed seed-100 MAT teacher trained for 19,988,480 joint steps. All local methods on a task share that teacher; references are descriptive complete-policy evaluations, not confidence bounds or action-level advice utilities. The 100-step LBF teachers are separate from the historical 50-step LBF teachers. The warehouse teachers are inherited from the allocation-control study. Student-seed intervals condition on these choices and do not estimate variation across teacher seeds.

J.2 Selection and data availability

Each task/method is tuned over six candidates with two seeds (7300–7301). Selection maximizes mean final return and breaks ties by candidate order. Confirmation uses ten fresh seeds (7400–7409). The analysis specification was fixed after confirmation began but before complete-grid aggregation; we do not claim prelaunch preregistration or blindness to all partial results. Final-policy return is primary, with all 35 comparisons retained. Best-policy results are secondary.

Episode-level records are available for all 300 local-method runs and reproduce their stored final-return means. Aggregate results remain available for all 400 runs. The raw evaluations for the 100 MAGPO/MAPPO runs and the five teacher parameter files are absent from the current data copy, so their full replay requires the original files. This limits independent verification without showing that the reported measurements are incorrect.

K COMPLETE WIDTH-4096 MATCHED CONFIRMATION

This separate grid contains all 120 runs: RWARE-small and RWARE-large, six local procedures and ten fresh seeds per procedure. Student hidden width is 4096; each run uses four GPUs, 19,988,480 joint steps and 320 final evaluation episodes. It has no capacity-matched native MAGPO or MAPPO arm. The fixed teacher, selected recipes and total-compute limitations remain; GPU replicas are not independent training seeds. One run exceeded its two-hour allocation and was replayed with the identical seed and configuration under a four-hour cap. The incomplete original is retained for provenance and cost accounting.

Table 25: Width-4096 final return: mean (sample SD) over ten training runs.

Method	RWARE-large	RWARE-small
Advisor	12.279 (0.515)	14.682 (0.400)
Combined	11.155 (0.498)	12.312 (0.779)
Linear	10.819 (0.500)	12.695 (0.473)
Quality	10.992 (0.418)	12.425 (0.537)
Raw	8.371 (4.403)	12.645 (0.419)
RL	3.192 (3.372)	8.092 (4.629)

Table 26: All ten primary wide-model contrasts. Positive favors combined. Bounds use 100,000 independent training-run bootstrap resamples, adjusted across ten contrasts (99.5% nominal per-contrast level).

Task	Comparator	Difference	Adjusted interval
RWARE-large	Advisor	-1.1241	$[-1.7313, -0.5250]$
	Linear	0.3359	$[-0.3072, 0.8688]$
	Quality	0.1631	$[-0.3981, 0.6809]$
	Raw	2.7844	$[0.2006, 7.0181]$
	RL	7.9634	$[5.2381, 10.6806]$
RWARE-small	Advisor	-2.3694	$[-3.1950, -1.7306]$
	Linear	-0.3825	$[-1.2369, 0.2803]$
	Quality	-0.1128	$[-0.9834, 0.5916]$
	Raw	-0.3325	$[-1.1538, 0.3113]$
	RL	4.2206	$[1.0544, 8.6531]$

Combined exceeds RL but trails code-style ADVISOR on both tasks. Both combined-minus-quality intervals include zero, as do comparisons with linear decay. Combined exceeds raw distillation on large. These results leave added discrepancy-gate value unresolved and do not establish equivalence or general SOTA superiority. Comparing interval signs between widths is not a test of a width-by-method interaction. The supplementary analysis retains all twenty final- and best-policy contrasts; the latter are secondary. Small-sample coverage and fixed-teacher uncertainty limits remain as in Appendix J.

For the secondary best-policy endpoint on RWARE-small, combined exceeds quality by 0.4375 (99.5% interval $[0.0341, 0.8728]$, adjusted across the ten secondary contrasts). This endpoint-specific selected-policy gain does not change the unresolved primary final-policy contrast, establish equivalence, or distinguish state-specific weighting from reduced imitation mass.

Table 27: Exploratory as-run RWARE-small contrasts. All 60 policies are retained despite the original mass-diagnostic acceptance failure. Values are differences of ten-run mean team returns with nominal 99.1667% Bonferroni percentile intervals, adjusted over six contrasts separately for each endpoint. These are not confirmatory results. C denotes combined; Q, quality-only; M, the attempted own-minibatch mass-matched quality arm.

Endpoint	Coefficient / contrast	Difference	Interval
Final	0.1: C–Q	+0.0466	$[-0.5812, 0.6667]$
Final	0.1: C–M	+0.4550	$[-0.1180, 1.0522]$
Final	0.1: M–Q	-0.4084	$[-1.0367, 0.2094]$
Final	0.3: C–Q	+0.0881	$[-0.3354, 0.5148]$
Final	0.3: C–M	+0.0047	$[-0.3694, 0.3775]$
Final	0.3: M–Q	+0.0834	$[-0.3606, 0.5325]$
Selected best	0.1: C–Q	+0.0488	$[-0.5145, 0.7189]$
Selected best	0.1: C–M	+0.0681	$[-0.5701, 0.7348]$
Selected best	0.1: M–Q	-0.0194	$[-0.6387, 0.6811]$
Selected best	0.3: C–Q	+0.2050	$[-0.2579, 0.7300]$
Selected best	0.3: C–M	+0.2881	$[-0.1362, 0.7506]$
Selected best	0.3: M–Q	-0.0831	$[-0.4544, 0.3491]$

L EXPLORATORY FIXED-COEFFICIENT RWARE-SMALL EXTENSION

Question and design. The basic selected-recipe comparison uses imitation coefficients 0.3 for combined and 0.1 for quality-only. This post-hoc extension compares combined, quality-only, and an attempted mass-matched quality control at each coefficient, fixing hidden width 128, learning rate 0.001, and the original RWARE-small configuration. Each cell has ten training seeds, 19,988,480 joint steps, and 320 evaluation episodes per endpoint. Twenty previously observed policies are reused and forty new policies complete the grid; there is no additional hyperparameter search. The mass control rescales quality weights to the counterfactual combined mean on its own current global minibatch. This does not match trajectories, gradient norms, optimizer updates, or cumulative imitation dose across learners.

Failed numerical acceptance condition. All 60 training runs completed, with matching saved recipes and source archives. However, the original analysis required every recorded global mean-weight mass error to be at most 2×10^{-7}. One global minibatch event exceeded this threshold, with error $2.384185791015625 \times 10^{-7}$, stored identically in all eight device/batch replicas. It occurred in the coefficient-0.3 mass-control run at seed 7404, update 92. The other 390,399 recorded global events did not exceed the threshold. The original complete-grid analysis therefore remains failed. A separate instrumented 100-update diagnostic retained the configured 1,220-update schedule but did not reproduce the original metric prefix; it cannot identify the cause of the original breach as roundoff.

Separate as-run reporting. A separately recorded amendment retains all 60 policies, both endpoints, both coefficients and all declared contrasts. Its pre-effect chronology is an author declaration rather than an independently registered protocol. Table 27 reports the actual trained policies descriptively. It preserves the original independent-run bootstrap and separate six-contrast endpoint families, but every estimate and interval is exploratory. The failed criterion is neither relaxed nor retrospectively passed; no policy is dropped, replaced, or rerun for this report.

Interpretation boundary. At either coefficient, every final-policy and selected-best interval includes zero. Additional gate benefit remains unresolved. This is not equivalence: several upper bounds exceed $+0.5$ return, a post-hoc descriptive reference without a validated equivalence margin. The attempted mass-control comparisons describe as-run policies, not an intervention certifying exact supervision mass or its causal effect. One task and teacher do not establish transfer, early utility prediction, or general MARL superiority.

M ADDITIONAL AUXILIARY FITTING AT NONFINAL CHECKPOINTS

Question and reused cohort. This intervention asks whether additional auxiliary fitting changes finite-update return when the original global imitation weight is preserved. It uses the four policy-diagnostic tasks, three student-training procedures (RL, quality-only JSD, and combined JSD), ten training seeds, and saved checkpoints at 4,997,120 and 9,994,240 steps: 240 checkpoint cells. Each cell has two independently seeded realizations of the entire fitting-and-update package. Final outcomes from these training trajectories were already inspected and motivated the question. The new intervention was specified before full-study collection, conditional on this reused cohort; it is not held-out student-seed, task, or teacher validation.

Fixed fitting package. For each realization, the unchanged student collects three separate datasets, respectively for auxiliary fitting, auxiliary validation, and the common PPO update. Each contains 2,048 joint transitions from 16 environments over 128 steps. The original auxiliary q_0 is cloned to q_1, which receives exactly 64 additional cross-entropy Adam updates on teacher distributions. Each update takes the first 256 joint-transition rows of a newly sampled random permutation of the fitting dataset; these are not partitioned epochs. Stored auxiliary Adam state is restored, learning rate is reset to 3×10^{-4}, and gradient norm is clipped at 10. No validation-based stopping or return-based choice of fitting budget is used. All cases are retained, including worsened validation cross-entropy. This treatment bundles fresh-data fitting and occupancy adaptation with additional optimization; it does not isolate pure optimization error or recover an oracle auxiliary.

Five common-rollout interventions. On the common PPO batch, let v denote the fixed value-based quality weight and $d(q)$ the discrepancy weight. Define $w_0 = vd(q_0)$, $w_1 = vd(q_1)$ and $m_j = \mathrm{mean}(w_j)$. The five weight arrays are

$$C_0 : w_0, \quad M_0 : v\frac{m_0}{\mathrm{mean}(v)}, \quad C_1 : w_1, \quad M_1 : v\frac{m_1}{\mathrm{mean}(v)}, \quad C_{1c} : w_1\frac{m_0}{m_1}.$$

Means and scaling use FP64 before casting to the original dtype. Zero target mass produces zero weights; positive target mass with zero source mass is an integrity failure, without an epsilon repair. Normalized weights are not clipped to one. The matching tolerance is absolute 10^{-7} plus relative 10^{-6}. Every arm restores the same student actor, critic, and Adam state, resets learning rate to 3×10^{-4}, and performs one PPO rollout update with four epochs, four minibatches, imitation coefficient one, and the same minibatch permutations. This fixes global mass for C_{1c} versus C_0, not minibatch mass, gradient norms, optimizer trajectories, or subsequent occupancy.

Evaluation and analysis. Five updated policies and an unchanged-policy anchor each receive 128 evaluation episodes with eight slots and 16 episodes per slot. Arms share an evaluation seed; fitting, validation, PPO collection, update permutations, and evaluation have separate assigned seeds. Two realizations repeat the whole package rather than isolating evaluation noise. The primary contrast is return(C_{1c})−return(C_0). For each task and stage, both realizations and all three training procedures are equally averaged within each of ten seed blocks. Using 100,000 common-label seed resamples, nominal 99.375% percentile intervals target a Bonferroni 95% family across the eight task/stage contrasts. These approximate intervals are not power or equivalence tests. No raw returns are pooled across tasks.

Secondary pointwise 95% intervals retain $C_0 - M_0$, $C_{1c} - M_0$, $C_1 - M_1$, and the own-mass interaction $(C_1 - M_1) - (C_0 - M_0)$. The latter changes both allocation and operating mass and cannot replace the primary contrast. A positive primary effect can mean less harmful selection while both arms remain below M_0. Validation cross-entropy changes are manipulation checks, not post-treatment inclusion criteria or causal mediation evidence. The retained 480 realizations yield 307,200 updated-policy and 61,440 unchanged-policy episodes; replay witnesses are additional. A separate four-task, 16-episode engineering pilot is excluded from these outcomes. Saved-state checks do not independently verify the original execution; the probability-normalization failure below limits interpretation.

Retained audit failure and reporting scope. All 480 realizations completed. A separate saved-tensor audit checked all cases and failed one additional probability-normalization criterion: one vector among 4,096 teacher distributions in an LBF-full quarter-stage realization had absolute sum error, computed in FP64 from stored FP32 entries, of $2.0171118109146846 \times 10^{-7}$, exceeding its fixed 2×10^{-7} tolerance. This was an audit-added simplex check, not a failed original mean-weight matching assertion. A separate precision diagnostic exactly reproduced all stored FP32 target elements; FP64 softmax on the same FP32 logits reduced maximum sum error to 4.44×10^{-16}. This supports precision sensitivity of this discrepancy, not harmlessness or an experiment pass. The original audit remains failed, and confidence in efficacy interpretation is low. The following results are descriptive as-run estimates retaining every case and every planned contrast, without relaxed tolerances, outcome-based exclusions, replacement runs or revised intervals.

Additional fitting does not establish improved selection. Every one of the eight primary return intervals includes zero (Figure 7); this does not establish equivalence. A degenerate $[0, 0]$ interval results when all observed paired seed-block differences are zero, not from proof of an exactly zero population effect. Exploratory validation-CE intervals indicate improved teacher prediction in three task/stage groups, worsening in two, and unresolved direction in three. Across individual realizations, validation CE worsens in 207 of 480 cases. Thus extra fitting is not uniformly improved teacher prediction, and neither lower CE nor its treatment effect establishes better supervision selection. Of 32 secondary pointwise intervals, two exclude zero: early LBF-partial has a positive own-mass interaction, while early RWARE-tiny has a negative original combined-minus-matched-quality contrast. These exploratory findings neither replace the primary family nor establish a common-mass refitting benefit. The experiment tests a fixed package at reused nonfinal checkpoints, not an oracle, a causal mediation model, or end-to-end training improvement.

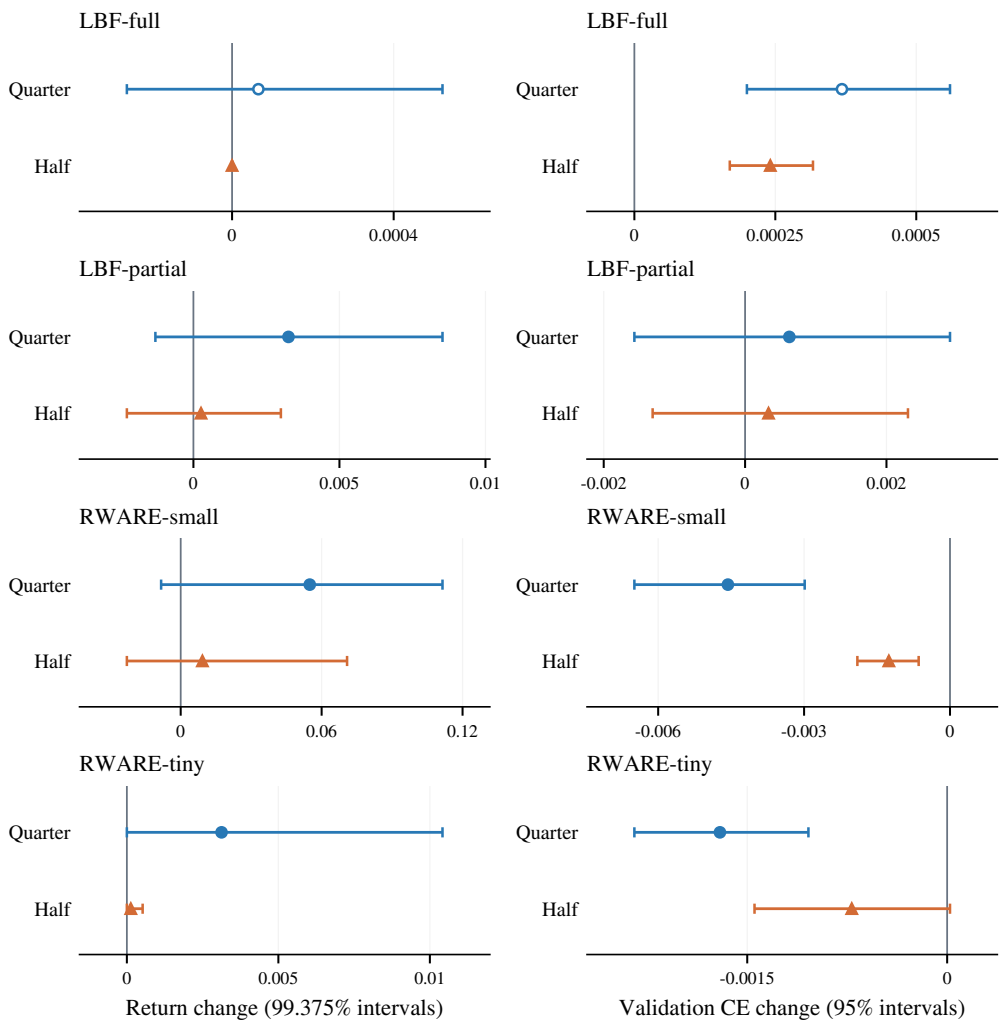


Figure 7: **Additional auxiliary fitting does not resolve a finite-update return gain.** Left: refitted minus original combined weighting at the original global mean imitation weight; positive favors refitting. Right: change in auxiliary validation cross-entropy; negative means better teacher prediction. Each point averages ten seed blocks. Paired bootstrap intervals are 99.375% for the eight primary return contrasts and exploratory 95% for cross-entropy. Quarter and half denote 4,997,120 and 9,994,240 historical steps. Open circles mark the group containing the retained probability-audit failure. Axes are task-specific; all 480 cases are included.

Interpretation. These are descriptive as-run results. The probability-normalization audit remains failed, and confidence in efficacy interpretation is low; the preceding audit paragraph reports the fixed tolerance and separate precision diagnostic. Intervals use 100,000 paired seed-block resamples, retaining three historical procedures and two realizations per block. Zero-containing intervals do not establish equivalence. The reused cohort and one fixed teacher per task do not support generalization to new student cohorts or teacher populations, or claims about full-training improvement.

Table 28: All 32 secondary nonfinal-fitting contrasts. Values are mean return differences with exploratory pointwise 95% percentile intervals from 100,000 resamples of ten training-seed blocks. Quarter/half refer to 4,997,120/9,994,240 historical steps. All 480 cases are retained; the original tensor-audit failure remains and confidence is low. These contrasts do not replace the eight primary comparisons in Figure 7. C denotes combined weighting, M matched quality weighting, and subscript $1c$ fixes the original global mean weight.

Task	Stage	Contrast	Mean [interval]
LBF-full	Quarter	$C_0 - M_0$	0.000000 [0.000000, 0.000000]
		$C_{1c} - M_0$	0.000065 [-0.000195, 0.000391]
		$C_1 - M_1$	0.000065 [-0.000195, 0.000391]
		$(C_1 - M_1) - (C_0 - M_0)$	0.000065 [-0.000195, 0.000391]
LBF-full	Half	$C_0 - M_0$	0.000000 [0.000000, 0.000000]
		$C_{1c} - M_0$	0.000000 [0.000000, 0.000000]
		$C_1 - M_1$	0.000000 [0.000000, 0.000000]
		$(C_1 - M_1) - (C_0 - M_0)$	0.000000 [0.000000, 0.000000]
LBF-partial	Quarter	$C_0 - M_0$	0.000260 [-0.004297, 0.004167]
		$C_{1c} - M_0$	0.003516 [-0.000521, 0.007682]
		$C_1 - M_1$	0.004232 [-0.000260, 0.008854]
		$(C_1 - M_1) - (C_0 - M_0)$	0.003971 [0.000195, 0.008008]
LBF-partial	Half	$C_0 - M_0$	-0.000326 [-0.003516, 0.002865]
		$C_{1c} - M_0$	-0.000065 [-0.002148, 0.002018]
		$C_1 - M_1$	-0.000846 [-0.002930, 0.001237]
		$(C_1 - M_1) - (C_0 - M_0)$	-0.000521 [-0.002214, 0.001107]
RWARE-small	Quarter	$C_0 - M_0$	-0.027083 [-0.089193, 0.038932]
		$C_{1c} - M_0$	0.027865 [-0.000391, 0.053125]
		$C_1 - M_1$	0.023177 [-0.017578, 0.061458]
		$(C_1 - M_1) - (C_0 - M_0)$	0.050260 [-0.037760, 0.131644]
RWARE-small	Half	$C_0 - M_0$	-0.002604 [-0.023307, 0.020052]
		$C_{1c} - M_0$	0.006641 [-0.027865, 0.049609]
		$C_1 - M_1$	0.008594 [-0.024609, 0.048437]
		$(C_1 - M_1) - (C_0 - M_0)$	0.011198 [-0.017318, 0.048177]
RWARE-tiny	Quarter	$C_0 - M_0$	-0.003125 [-0.007943, -0.000260]
		$C_{1c} - M_0$	0.000000 [-0.001172, 0.001042]
		$C_1 - M_1$	-0.001953 [-0.007292, 0.001693]
		$(C_1 - M_1) - (C_0 - M_0)$	0.001172 [0.000000, 0.002865]
RWARE-tiny	Half	$C_0 - M_0$	0.000000 [0.000000, 0.000000]
		$C_{1c} - M_0$	0.000130 [0.000000, 0.000391]
		$C_1 - M_1$	0.000000 [0.000000, 0.000000]
		$(C_1 - M_1) - (C_0 - M_0)$	0.000000 [0.000000, 0.000000]

N AN OPTIMAL JOINT TEACHER CAN GIVE HARMFUL LOCAL UPDATE DIRECTIONS

This construction separates the teacher's joint optimality from the usefulness of a unilateral student update. It uses an exact auxiliary and no teacher error. It is not a benchmark result or a statement about simultaneous PPO updates.

Game and policies. Let $Z \sim \mathrm{Bernoulli}(3/4)$ be hidden from both students. Each receives a constant observation and independently chooses $A_i \in \{0, 1\}$; the one-step team reward is one if $A_1 = A_2 = Z$ and zero otherwise. A privileged teacher sees Z and chooses (Z, Z), achieving the maximum possible return of one. The exact conditional teacher distribution for either constant observation is $q = (1/4, 3/4)$. Student 1 chooses one with probability $p = \sigma(\theta)$, initially $p = 1/2$; student 2 chooses one with probability $1/10$ and is held fixed. Thus

$$R(\theta) = \frac{3}{4}\frac{p}{10} + \frac{1}{4}\frac{9(1-p)}{10} = \frac{9}{40} - \frac{3}{20}p, \qquad R'(0) = -\frac{3}{80}.$$

Discrepancy and local utility have opposite orderings. Write $H(u) = -u \log u - (1-u)\log(1-u)$ and $j(u) = H((1+u)/2) - H(u)/2$, the symmetric JSD from a point mass to a policy assigning its action probability u. Natural logarithms give

$$j'(u) = \tfrac{1}{2} \log \frac{u}{1+u} < 0 \quad (0 < u < 1).$$

Consequently $d_1 = \mathrm{JSD}(\delta_1, q) = j(3/4) < d_0 = \mathrm{JSD}(\delta_0, q) = j(1/4)$. For $L_z(\theta) = \mathrm{JSD}(\delta_z, \mathrm{Bernoulli}(p))$,

$$L_1'(0) = -\frac{\log 3}{8}, \qquad L_0'(0) = \frac{\log 3}{8}.$$

The derivative of team return along the corresponding gradient-descent direction is $-3 \log 3/640$ for the lower-discrepancy signal $z = 1$ and $+3 \log 3/640$ for the higher-discrepancy signal $z = 0$. Evaluation here uses the original state distribution, not only the selected state.

A mean-matched gate comparison. For any $\alpha > 0$, let $w_z = e^{-\alpha d_z}$ and $m = 3w_1/4 + w_0/4$. Compare $L_g = 3w_1 L_1/4 + w_0 L_0/4$ with the constant-weight objective $L_m = m(3L_1/4 + L_0/4)$. Both have population mean weight m, while $w_1 > w_0$. Their negative logit gradients u_g, u_m at $\theta = 0$ satisfy

$$u_g - u_m = \frac{3 \log 3}{64}(w_1 - w_0) > 0.$$

After one scalar update with any finite $\eta > 0$ and shared additive term $c \in \mathbb{R}$, the logits are $\theta_g = \eta(c + u_g)$ and $\theta_m = \eta(c + u_m)$. Since σ increases strictly and R decreases strictly in p, $R(\theta_g) < R(\theta_m)$. At $\alpha = 0$ or $\eta = 0$ the comparison becomes equality. All gradients are evaluated at the same initial policy; this argument does not cover arbitrary optimizer transformations.

Changing only the fixed partner reverses the effect. For partner probability $b \in (0, 1)$, the same construction gives $R_b(p) = (1 - b)/4 + (b - 1/4)p$. The gate/control ordering $p_g > p_m$ is unchanged, so $R_b(p_g) - R_b(p_m) = (b - 1/4)(p_g - p_m)$ has the sign of $b - 1/4$. The same discrepancy scores therefore yield harm, equality or benefit as b crosses $1/4$. This does not identify causes of benchmark outcomes.

Interpretation. The teacher coordinates both agents optimally, but the student's partner usually chooses zero. Moving only student 1 toward the teacher's more probable action therefore worsens coordination. An exact local conditional-mean auxiliary does not remove this dependence on the current partner. The construction uses a deliberately mismatched partner and constant observations; it does not measure the prevalence of this failure or establish harm from joint training, a quality-weighted method, or the complete benchmark algorithm.

O A SIMULTANEOUS-UPDATE COORDINATION COUNTEREXAMPLE

The fixed-partner construction is not the only possible failure. In the same one-step game, both students can update simultaneously and the discrepancy gate can still lose to its mean-matched constant control. This result uses ordinary population-gradient steps and does not concern PPO or long-run learning.

Setup and gradients. Keep the hidden-bit distribution, optimal joint teacher, constant observations, independent actions and exact conditional auxiliary of Appendix N. Now both students start with $p_0 = 1/10$, so $\theta_0 = -\log 9$, and both logits update. Take finite nonnegative weights $w_1 > w_0$ and $m = (3w_1 + w_0)/4$, all held constant with respect to the student logits when differentiating. The exponential discrepancy gate with $\alpha > 0$ is one such choice. For either student, the raw-JSD loss derivatives at p_0 are

$$L_1' = -\frac{9}{200}\log 11, \qquad L_0' = \frac{9}{200}\log(19/9).$$

For the negative gradients $u_g = -(3w_1L_1' + w_0L_0')/4$ and $u_m = -m(3L_1' + L_0')/4$,

$$u_g - u_m = \frac{27}{3200}(w_1 - w_0)\log(209/9) > 0.$$

A finite interval of simultaneous steps. Let a common finite scalar c represent the same additional logit-gradient term in both arms and both agents. After step size $\eta > 0$, each agent's logit is $\theta_g = -\log 9 + \eta(c + u_g)$ in the gated arm and $\theta_m = -\log 9 + \eta(c + u_m)$ in the control. Hence $p_g > p_m$. Along this symmetric policy line, the original team return is

$$R(p, p) = \tfrac{3}{4}p^2 + \tfrac{1}{4}(1 - p)^2 = p^2 - \tfrac{1}{2}p + \tfrac{1}{4},$$

which strictly decreases on $(0, 1/4)$. If $c + u_g > 0$, any $0 < \eta < \log 3/(c + u_g)$ keeps $p_g < 1/4$; if $c + u_g \le 0$, every finite positive η does so. Thus each case has a nonempty step-size interval with $R(p_g, p_g) < R(p_m, p_m)$, although both students update. Tied weights give equality. The exact per-agent policy-gradient term at the common initial policy is $c = p_0(1 - p_0)(p_0 - 1/4) = -27/2000$, so the construction includes a specified RL-plus-imitation gradient step.

Interpretation and limits. The students initially coordinate mostly on zero. Moving a little toward the teacher's more frequent action can reduce their current joint return before a higher-return coordinated policy is reached. The result compares two ordinary updates from identical initial policies. It does not imply that the gated policy is worse than the initial policy, that the effect persists under repeated updates, or that the gate cannot help cross this coordination valley later. Symmetric initialization, exact auxiliary access, population gradients, and fixed weights are explicit assumptions; clipped or adaptive optimizers and benchmark prevalence remain outside the claim.

Equal population mean weight does not equalize the gradient norm. The two students have equal coordinates within each arm, and their imitation-update vectors are collinear. This construction therefore does not isolate directional selectivity from a change in update magnitude.

P A SIMULTANEOUS COMPARISON AT EQUAL JOINT LOGIT-STEP NORM

The symmetric example does not match gradient magnitudes. Here an open family retains a disadvantage after normalizing both joint imitation updates to the same Euclidean logit norm. We first prove a one-step result, then characterize cumulative return along normalized imitation flow. Neither result includes an RL gradient or establishes a PPO effect.

Setup and claim. Use the one-step game, optimal joint teacher and exact auxiliary $q = (1/4, 3/4)$ from Appendix N. For every $0 < p_1 < p_2 < 1/4$ and finite fixed $w_1 > w_0 \geq 0$, set $m = (3w_1 + w_0)/4$. Let g and v be the negative separate-logit gradients of the sums of the two students' gated and constant-weighted population raw-JSD losses at the initial policy. Define $d_g = g/\|g\|_2$ and $d_m = v/\|v\|_2$, fixed during the update. Both students update simultaneously by $\theta_g = \theta + \eta d_g$ or $\theta_m = \theta + \eta d_m$. We show that $\delta = \nabla R(\theta) \cdot (d_g - d_m) < 0$ and

$$R(\theta_g) < R(\theta_m) \qquad \text{for every } 0 < \eta < -4\delta.$$

Both joint logit displacements have norm exactly η. Since $m > 0$, $v = mu$ for the unweighted imitation direction u, so $d_m = u/\|u\|_2$: mean matching adds no further constraint after normalization. The step interval depends on the initial policy and weights; no uniform positive bound is claimed.

Ordering the directions. Write $s(p) = p(1-p)$, $a(p) = \log((1+p)/p)$ and $b(p) = \log((2-p)/(1-p))$. Direct differentiation of the finite raw-JSD sums gives derivatives $-a(p)/2$ and $b(p)/2$ for targets one and zero. Thus, with $K(p) = 3w_1 a(p) - w_0 b(p)$,

$$g_i = s(p_i)K(p_i)/8, \qquad v_i = ms(p_i)[3a(p_i) - b(p_i)]/8.$$

For $p < 1/4$, $a(p) > \log 5 > \log(7/3) > b(p) > 0$, $a'(p) = -1/[p(1+p)] < 0$ and $b'(p) = 1/[(1-p)(2-p)] > 0$. Consequently all gradient components are positive and $K'(p) < 0$. The ratio $x(p) = b(p)/a(p)$ is strictly increasing and lies in $(0, 1)$. Moreover,

$$\frac{g_i}{v_i} = f(x(p_i)), \qquad f(x) = \frac{3w_1 - w_0 x}{m(3-x)}, \qquad f'(x) = \frac{3(w_1 - w_0)}{m(3-x)^2} > 0.$$

Hence the slopes satisfy $0 < t_m = v_2/v_1 < t_g = g_2/g_1$, proving noncollinearity. Since K decreases, $t_g = [s(p_2)/s(p_1)][K(p_2)/K(p_1)] < s(p_2)/s(p_1)$.

Return ordering and finite steps. The team return is $R = p_1 p_2 - (p_1 + p_2)/4 + 1/4$. Its initial logit gradient has entries $r_1 = s(p_1)(p_2 - 1/4) < 0$ and $r_2 = s(p_2)(p_1 - 1/4) < 0$, so

$$\frac{r_2}{r_1} = \frac{s(p_2)}{s(p_1)} \frac{1/4 - p_1}{1/4 - p_2} > \frac{s(p_2)}{s(p_1)} > t_g > t_m > 0.$$

A positive unit direction of slope t has return derivative $h(t) = (r_1 + r_2 t)/\sqrt{1+t^2}$ and $h'(t) = (r_2 - r_1 t)/(1+t^2)^{3/2} < 0$ for $0 < t < r_2/r_1$. Therefore $\delta = h(t_g) - h(t_m) < 0$. Globally, the logit Hessian has diagonal entries $s(p_i)(1-2p_i)(p_j - 1/4)$ bounded in absolute value by $3/16$, and off-diagonal entries $s(p_1)s(p_2) \leq 1/16$. Its operator norm is at most $1/4$ by symmetry and the row-sum bound. For $F(\eta) = R(\theta + \eta d_g) - R(\theta + \eta d_m)$, this implies $F(0) = 0$, $F'(0) = \delta$ and $|F''| \leq 1/2$. Smoothness and Taylor's integral remainder give $F(\eta) \leq \eta\delta + \eta^2/4 < 0$ on the stated interval.

Gate interpretation and scope. Define $d_z = \mathrm{JSD}(e_z, q)$, where e_z is the point mass at action z. The exact auxiliary gives $d_1 < d_0$: the divergence from target one is strictly decreasing in its assigned probability, and label symmetry compares probabilities $3/4$ and $1/4$. Thus every finite exponential gate $w_z = \exp(-\alpha d_z)$ with $\alpha > 0$ satisfies the assumptions. Any positive ordered pair is proportional to such a gate with $\alpha = \log(w_1/w_0)/(d_0 - d_1)$; the common scale cancels under normalization. The case $w_0 = 0$ is an ordered-weight extension, not a finite exponential gate. Both directions initially reduce return. At $p_1 = p_2$ or $w_1 = w_0 > 0$, normalized directions coincide, so strict ordering disappears. Equal joint logit norm does not match per-agent norms, policy distances or KL. This result does not establish a practical decision rule, benchmark prevalence, coordinate-invariant behavior or long-run harm. Known coordination phenomena remain the setting of this more specific comparator test.

Local disadvantage need not persist along imitation flow. In the same game, suppose $w_1 > w_0 > 0$ remain fixed and consider continuous-time population-JSD gradient flow with no RL term. Put $K_w(p) = 3w_1a(p) - w_0b(p)$, using the definitions above. Each coordinate obeys $\dot{p}_i = s(p_i)^2K_w(p_i)/8$. The function K_w is strictly decreasing from $+\infty$ to $-\infty$, so has exactly one root p_w. The constant comparator has root p_c defined by $3a(p_c) = b(p_c)$. Since $a(1/2) = b(1/2) = \log 3$, $p_c > 1/2$, and $K_w(p_c) = 3(w_1 - w_0)a(p_c) > 0$ gives $p_w > p_c$.

From any interior probability, the scalar flow stays between its start and its root: the field points toward the root, and smooth ODE uniqueness prevents crossing it. This compact interior interval gives global existence. Monotonicity then gives convergence to the root; any other limit would have speed bounded away from zero. An initial root stays fixed. Thus both coordinates converge to p_w in the gated arm and p_c in the control. Since $R(p, p) = p^2 - p/2 + 1/4$ increases for $p > 1/4$, the gated limiting return is strictly greater, including from the initial-policy region above.

Transient cumulative cost at equal joint logit speed. Continue to assume $w_1 > w_0 > 0$ and the unequal initial probabilities $0 < p_1^0 < p_2^0 < 1/4$. Write $G_g(\theta)$ and $G_m(\theta)$ for the joint imitation vector fields whose initial values were g and v above, and $d_X(\theta) = G_X(\theta)/\|G_X(\theta)\|_2$ for $X \in \{g, m\}$. Consider repeated normalized flow $d\theta_X/d\tau = d_X(\theta_X)$, with no RL gradient. At an equilibrium the normalized field is undefined; we explicitly keep the trajectory stationary once it reaches that endpoint. Until stopping, τ measures joint Euclidean logit path length, not training steps, compute, or a coordinate-invariant policy distance.

Let $H(\theta) = R(\sigma(\theta_1), \sigma(\theta_2))$ and retain $\delta = \nabla H(\theta^0) \cdot [d_g(\theta^0) - d_m(\theta^0)] < 0$ from the initial-direction proof. Define

$$\rho = \tfrac{1}{2} \min_i\{\mathrm{logit}(1/4) - \theta_i^0\} > 0, \qquad B_X = \max_{\|\theta - \theta^0\|_2 \le \rho} \|Dd_X(\theta)\|_{\mathrm{op}}.$$

The ball lies strictly below probability $1/4$ in both coordinates, where both gradient components are positive. Thus B_X is finite. Unit speed keeps both flows inside it for $0 \le \tau \le \rho/2$, and smoothness permits continuation throughout this interval. Globally, $\|\nabla H\|_2 \le M = 3\sqrt{2}/16$ and the previously established Hessian bound is $\|D^2H\|_{\mathrm{op}} \le 1/4$. Therefore, for $D(\tau) = H(\theta_g(\tau)) - H(\theta_m(\tau))$,

$$|D''(\tau)| \le A = \tfrac{1}{2} + M(B_g + B_m).$$

Indeed, each arm has second derivative $d_X^T D^2 H d_X + \nabla H \cdot (Dd_X)d_X$. With $\tau_0 = \min\{\rho/2, -\delta/A\} > 0$, Taylor's integral remainder gives

$$D(\tau) \le \delta\tau + \tfrac{1}{2}A\tau^2 \le \tfrac{1}{2}\delta\tau, \qquad C(T) := \int_0^T D(\tau)\, d\tau \le \frac{\delta T^2}{4} < 0 \quad (0 < T \le \tau_0).$$

Thus the disadvantage persists over a positive interval of repeated normalized flow and cumulative return. This bound depends on the initial policy and weights and may describe a short interval.

Eventual cumulative benefit under the stopping convention. The unnormalized imitation trajectories proved above increase coordinatewise from p_i^0 to p_X^*, where $p_g^* = p_w > p_m^* = p_c > 1/2$. Their logits therefore increase to finite limits $\theta_X^* = \mathrm{logit}(p_X^*)$. Their total arc lengths satisfy

$$0 < S_X = \int_0^\infty \|G_X(\theta_X(t))\|_2\, dt \le \int_0^\infty \sum_i G_{X,i}(\theta_X(t))\, dt = 2\theta_X^* - \theta_1^0 - \theta_2^0 < \infty.$$

No coordinate reaches its scalar equilibrium in finite unnormalized time by smooth ODE uniqueness. Accumulated arc length is consequently strictly increasing, and its inverse reparameterizes the trajectory into the unit-speed flow on $[0, S_X)$. At S_X it converges to the equilibrium and is extended constantly. This defines the trajectory without evaluating the normalized field at a zero gradient.

Let $S = \max_X S_X$ and $\Delta = R(p_w, p_w) - R(p_c, p_c) > 0$. For all $T \ge S$,

$$C(T) = C(S) + \Delta(T - S).$$

It follows that $C(T) > 0$ whenever $T > S + \max\{0, -C(S)\}/\Delta$. Hence the early cumulative disadvantage and eventual benefit coexist in this exact game, even though the moving trajectories have the same joint logit speed. The two arms need not travel the same total distance before stopping. This result concerns pure-imitation continuous flow with a stated stationary extension; it proves neither convergence of discrete normalized updates nor a finite-step RL or PPO effect. It does not identify a benchmark mechanism or a generally useful teaching rule.

Q FRESH-VISIT UTILITY PREDICTION REPLICATION

The original RWARE observations motivate a prospective measurement replication, not an untouched hypothesis. We freeze the original five predictors fitted on student seeds 0–4 and reuse final checkpoints from seeds 5–9 for RL, quality-JSD and combined-JSD in small and tiny. Each of these 30 cells supplies eight fresh reset visits after 13 student actions and 128 paired remaining-horizon continuations per visit, with discount 0.99. The teacher-first intervention is unchanged. All 240 visits and 30,720 pairs are retained. No predictor is refitted. Freshness concerns reset and continuation streams, not tasks, teachers or student training seeds. Streams are shared across methods within a task/seed/visit and separated across tasks and training seeds.

The frozen primary estimand is MSE(value+JSD) minus MSE(value). We average eight visits per checkpoint, three methods per training-seed block, and five blocks per task. All 3,125 ordered bootstrap resamples give linear-interpolated percentile bounds, Bonferroni-adjusted across two tasks to 97.5% per interval. At five blocks this is an approximate uncertainty summary, not guaranteed coverage. No task or old/new-label pooling is performed.

Table 29: Prespecified replication contrasts. Positive means worse prediction with JSD. Relative changes divide the difference by value-only MSE.

Task	MSE difference	Change(%)	Adjusted interval
RWARE-small	$-2.55042\text{e}{-5}$	-0.3044	$[-1.15943\text{e}{-3}, 1.01875\text{e}{-3}]$
RWARE-tiny	$1.31312\text{e}{-5}$	$+0.4356$	$[5.64394\text{e}{-6}, 1.94652\text{e}{-5}]$

Table 30: All frozen predictor MSEs on the new visits.

Predictor	RWARE-small	RWARE-tiny
Constant	0.00795070	0.00294135
Value	0.00837771	0.00301478
JSD	0.00817965	0.00293070
Entropy	0.00730793	0.00282732
Value+JSD	0.00835220	0.00302791

Small's original adverse observation does not repeat: the new point estimate changes sign and its primary interval includes zero. Tiny retains a positive MSE difference under this fixed predictor comparison. The studies change both sampled visits and continuation precision (32 versus 128 pairs); these data do not isolate whether the change reflects reset sampling, Monte Carlo variability, or both. Neither zero inclusion nor cross-study changes in interval sign establish equivalence or a statistically resolved between-study difference. These results do not evaluate a practical advice-selection rule, other predictor families or unseen-teacher generalization.

Four disjoint 32-pair batches retain secondary MSE differences of $2.20514\text{e}{-4}, -6.58705\text{e}{-4}, 5.34846\text{e}{-4}, -1.98673\text{e}{-4}$ for small and $1.37402\text{e}{-5}, 1.24414\text{e}{-5}, 1.30561\text{e}{-5}, 1.32873\text{e}{-5}$ for tiny. These batches are descriptive, not independent training replicates.

Table 31: Historical combined-JSD minus adapted native MAPPO at the upper budget bracket. Means use ten training seeds. Intervals are nominal four-comparison 98.75% percentile bootstrap intervals; finite-sample coverage is approximate. Full-LBF non-resolution is not equivalence.

Task	Combined	MAPPO	Difference	98.75% interval
LBF-full	0.9980	0.9992	-0.0012	$[-0.0043, 0.0020]$
LBF-partial	0.9293	1.0000	-0.0707	$[-0.0832, -0.0570]$
RWARE-small	23.6273	1.1305	$+22.4969$	$[20.8359, 23.9508]$
RWARE-tiny	2.0305	1.1281	$+0.9023$	$[0.3445, 1.3289]$

R A NATIVE MAPPO COMPARISON ON THE HISTORICAL TASKS

This comparison tests sensitivity to a separately implemented RL baseline on the policy-diagnostic tasks. Forty adapted EPyMARL MAPPO confirmations use ten training seeds per task. We compare them with the 760 historical confirmations covering 19 selected methods. The selected combined system exceeds this MAPPO configuration on both warehouse tasks, trails it on partial LBF, and has an unresolved difference on full LBF (Table 31). The warehouse comparator also trails historical RL, so these results do not identify an additional gate benefit.

Selection and evaluation. The native search evaluates four learning rates with three tuning seeds per task, for 48 trials at a requested 2,097,152 transitions and 64 sampled evaluation episodes. Selection maximizes mean upper-bracket return, with the lowest candidate index breaking ties. Confirmation seeds are new to native tuning; the historical methods' outcomes had already been inspected before this native analysis was specified. The selected learning rate is 3×10^{-4} on both LBF tasks and 10^{-4} on both warehouse tasks. Warehouse tuning means are 0.1042 on small and 0.1198 on tiny; this limited learning-rate search does not establish a broadly optimized baseline.

The target student budget is 19,988,480 transitions. Native episode batches bracket that target, and both policies are evaluated with 128 sampled episodes, retaining the first completions from eight vector slots. The LBF horizon is 50 and the RWARE horizon is 500. This comparison retains the historical LBF stopping-rule limitation and is separate from the fixed-slot sensitivity in Appendix D. Lower/upper native steps are 19,985,000/19,990,000 on both warehouse tasks. LBF-full upper steps range from 19,988,487 to 19,988,590; partial-LBF upper steps range from 19,988,509 to 19,988,581. These are close, not identical, transition budgets.

Inference and budget sensitivity. The four primary contrasts compare combined-JSD with the upper bracket. We use 100,000 independent-method training-seed bootstrap draws; matching numeric seed labels does not define a paired algorithm experiment. All 19 historical methods and both brackets remain in the 152-contrast analysis. The other 148 intervals are exploratory 95% intervals. Lower-bracket combined-minus-MAPPO differences are -0.0012 on full LBF, -0.0707 on partial LBF, 22.4813 on small, and 0.9125 on tiny. Their exploratory intervals preserve the primary pattern; the lower bracket never replaces the primary endpoint. Native partial LBF returns 1 on every saved evaluation episode, giving zero empirical seed variance. This observed saturation does not establish deterministic reliability.

Baseline health and interpretation. Historical RL scores 18.2203 on small and 15.0203 on tiny, while quality-only JSD scores 23.1789 and 2.0336. Thus the native warehouse deficit is not specific to combined-JSD: on tiny, combined-JSD itself is far below historical RL. The native actor is recurrent with hidden width 128, whereas the historical actor is feedforward with hidden width 128. Native episode batches, reward standardization, five-step return targets and separate optimizers differ from historical GAE and rollout updates; training time-limit target semantics also differ. These differences are possible explanations, not identified causes. The comparison concerns selected systems, without matching total compute or varying the fixed teacher per task.

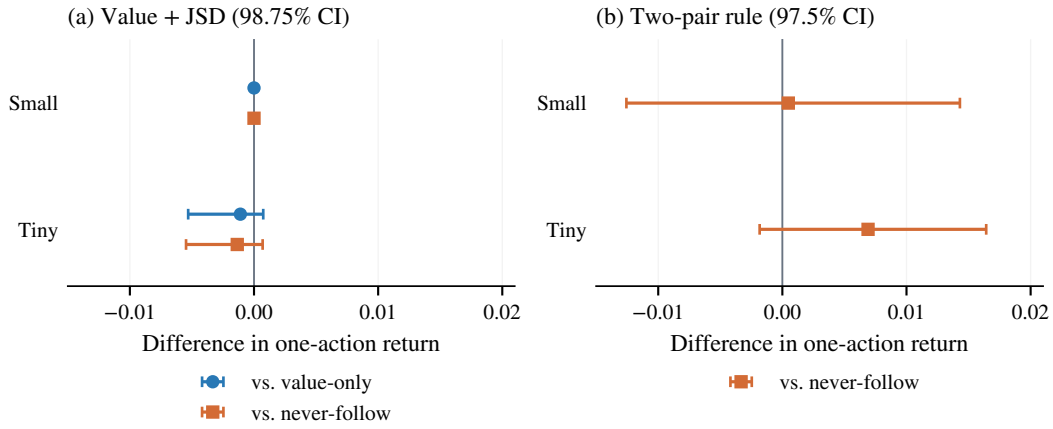


Figure 8: **Neither prospective test establishes the proposed advice benefit.** Distinct cohorts of ten new student seeds per task, conditional on fixed teachers. (a) Value+JSD minus both controls required by the usefulness criterion, with nominal 98.75% student-block bootstrap intervals adjusted over four contrasts. Both rules reject every small visit, giving exact sample zeros rather than evidence of population equivalence. (b) Two-pair simulator-assisted selection minus never-follow, with nominal 97.5% intervals adjusted over two contrasts. Selection and evaluation continuations are separate. Points measure one teacher action followed by the unchanged student, not full-training gains or net benefit after query costs.

S PROSPECTIVE TEACHER-ADVICE DECISIONS ON NEW STUDENTS

Question and frozen design. Prediction error alone does not establish a useful advice decision. We test whether adding teacher–auxiliary JSD to a value-gap predictor improves the choice to take one teacher action. The protocol fixes the predictors, decision threshold, controls and inference before collecting outcomes from new student training seeds. The two tasks are the policy-diagnostic RWARE-small and RWARE-tiny, each with its original frozen privileged teacher. Existing RL students with seeds 0–4 provide fit data at 4,997,120 and 9,994,240 joint environment steps. Twenty new RL students provide test data at the same checkpoints: ten per task, seeds 938100–938109 for small and 938200–938209 for tiny. The pilot student is excluded.

Students retain the original RL recipe: hidden width 128, 16 environments, 128-step rollouts, four epochs and four minibatches, zero imitation coefficient, four auxiliary updates, discount 0.99 and episode horizon 500. Learning rates are 0.001 for small and 0.003 for tiny. Each checkpoint supplies eight reset visits after 13 student actions. At a visit, paired remaining-horizon continuations compare one joint teacher action followed by the unchanged student against the unchanged student throughout. Fit visits use 64 pairs; test visits use 128. The complete fit cohort has 20 cells, 160 visits and 10,240 pairs; the test cohort has 40 cells, 320 visits and 40,960 pairs. Collection stream inputs are disjoint across the frozen manifest; this alone does not prove statistical independence.

Rules and estimand. For each task and stage, fit three ridge predictors to 40 visits: value gap, value gap plus teacher–auxiliary JSD, and teacher/student entropies. Each fit minimizes summed squared error plus unit coefficient penalty, with an unpenalized intercept. Features use fit-only means and population standard deviations, replacing zero scales by one. No hyperparameter selection is used. Coefficients are sealed before new student training. A rule follows the teacher if and only if predicted utility is strictly positive; ties reject. The other controls always follow or never follow. For rule r and visit v, its measured gain over never following is

$$G_r(v) = I_r(v)\,\overline{\Delta}(v),$$

where $I_r(v)$ is one if the rule follows and zero otherwise, and $\overline{\Delta}(v)$ is the mean paired discounted teacher-first minus student-only return. The measured intervention changes one joint action, not the student parameters. Features use privileged teaching-time information.

Primary inference. Average eight visits per stage and both stages within each student seed, giving ten blocks per task. The primary family contains four contrasts: value+JSD minus value and value+JSD minus never-follow on each task. We use 100,000 paired seed-block bootstrap draws per task, shared across its two contrasts, with seeds 205093001 and 205093002. Linear percentile bounds at 0.00625 and 0.99375 give nominal Bonferroni-adjusted 98.75% intervals. Useful added decision support requires both lower bounds to exceed zero within a task. No cross-task pooling, post-test threshold tuning or promotion of secondary endpoints is used. With ten blocks, interval coverage is approximate and conditional on the fitted predictors and fixed teachers.

Table 32: All prespecified primary advice-decision contrasts. Neither task meets the useful-added-decision criterion. Small's degenerate intervals reflect identical observed reject decisions, not population equivalence.

Task	Contrast	Mean	Nominal 98.75% interval
RWARE-small	Value+JSD − value	0.000000	$[0.000000, 0.000000]$
RWARE-small	Value+JSD − never	0.000000	$[0.000000, 0.000000]$
RWARE-tiny	Value+JSD − value	-0.001098	$[-0.005303, 0.000744]$
RWARE-tiny	Value+JSD − never	-0.001349	$[-0.005475, 0.000696]$

Table 33: All frozen decision controls. Gains are discounted one-action utility relative to never-follow, not full-training returns. Follow counts are out of 160 visits per task. MSE measures utility prediction; fixed-action controls have no predictor.

Task	Rule	Mean gain	Follow count	MSE
RWARE-small	Value	0.00000000	0	0.00794197
RWARE-small	Value+JSD	0.00000000	0	0.00837894
RWARE-small	Entropy	-0.00223372	12	0.00791260
RWARE-small	Always	-0.04783361	160	–
RWARE-small	Never	0.00000000	0	–
RWARE-tiny	Value	-0.00025094	19	0.01209480
RWARE-tiny	Value+JSD	-0.00134897	14	0.01216473
RWARE-tiny	Entropy	0.00090051	5	0.01281436
RWARE-tiny	Always	-0.00186607	160	–
RWARE-tiny	Never	0.00000000	0	–

Small's value rules reject all observed visits. In tiny, value+JSD follows at 14 visits versus 19 for value-only; both primary point estimates are negative, but their intervals include zero. The tested added-JSD rule therefore fails the prespecified usefulness criterion on both tasks. This conditional result is neither an equivalence finding nor demonstrated harm. It does not establish that JSD lacks predictive information for other rules or that teacher advice is always harmful.

Scope. The 320 visits and 40,960 pairs are not independent training replicates. Inference conditions on two teachers and four fit samples of 40 visits; fit-sample and teacher-population variation are not included. Predictors were fixed before test collection, without external preregistration. Repeated advice, full-training benefits, and decentralized deployment require separate experiments.

Post-hoc outcome-assisted opportunity diagnostic. Failure of the sealed feature rules need not imply an absence of useful advice opportunities. We probe this distinction with a secondary analysis fixed after the primary results were known. At every one of the 320 saved visits, we split the 128 paired continuations by their prescribed replicate order into two halves of 64. Let their mean intervention differences be A_s and B_s. We select teacher advice when one half's mean is strictly positive, evaluate that decision on the other half, then swap and average:

$$Z_s = \tfrac{1}{2}\{\mathbf{1}[A_s > 0]B_s + \mathbf{1}[B_s > 0]A_s\}.$$

Table 34: Descriptive stage-specific controls. Each task/stage has 80 visits. These secondary summaries are not additional confirmatory tests.

Task / steps	Rule	Mean gain	Follow count	MSE
Small / 4,997,120	Value	0.00000000	0	0.00760954
Small / 4,997,120	Value+JSD	0.00000000	0	0.00805881
Small / 4,997,120	Entropy	0.00000000	0	0.00731886
Small / 4,997,120	Always	-0.05477091	80	–
Small / 4,997,120	Never	0.00000000	0	–
Small / 9,994,240	Value	0.00000000	0	0.00827439
Small / 9,994,240	Value+JSD	0.00000000	0	0.00869906
Small / 9,994,240	Entropy	-0.00446745	12	0.00850635
Small / 9,994,240	Always	-0.04089631	80	–
Small / 9,994,240	Never	0.00000000	0	–
Tiny / 4,997,120	Value	-0.00050188	19	0.00723655
Tiny / 4,997,120	Value+JSD	-0.00269793	14	0.00743759
Tiny / 4,997,120	Entropy	0.00181183	3	0.00722665
Tiny / 4,997,120	Always	-0.00785180	80	–
Tiny / 4,997,120	Never	0.00000000	0	–
Tiny / 9,994,240	Value	0.00000000	0	0.01695305
Tiny / 9,994,240	Value+JSD	0.00000000	0	0.01689186
Tiny / 9,994,240	Entropy	-0.00001080	2	0.01840207
Tiny / 9,994,240	Always	0.00411965	80	–
Tiny / 9,994,240	Never	0.00000000	0	–

Under conditionally independent, identically distributed halves at a fixed saved state and policy, $\mathbb{E}[Z_s \mid s] = \Pr(A_s > 0 \mid s)\mu_s$, where μ_s is the true mean one-action intervention difference. Distinct replicate seeds separate selection outcomes from evaluation outcomes; the two arms within each pair intentionally share random streams. We do not treat the two swapped terms as independent samples. The diagnostic requires repeated outcomes at the same state and is not a feature-only deployable rule or an oracle bound. This is a restricted Monte Carlo rollout comparison: a candidate first action is followed by a base policy, as in Bertsekas & Castañon (1999). Multiagent rollout is also established (Bertsekas, 2021); we do not claim a new selection principle or transfer exact policy-improvement guarantees to finite-sample decisions.

We average the 16 visits within each student and bootstrap ten student blocks per task 100,000 times. Table 35 gives exploratory nominal 98.75% percentile intervals with a four-contrast Bonferroni adjustment. The adjustment does not make the post-hoc analysis confirmatory or guarantee coverage with ten blocks. Teachers, visited-state protocol and the chosen half-split remain fixed; acquisition cost is not matched to the frozen rules. No predictor or threshold was retuned, and the original primary result is unchanged.

Table 35: Post-hoc repeat-state outcome-assisted control. All 160 visits per task are retained. Differences use the original discounted one-joint-action utility, not full-training return. Small's two contrasts coincide because the frozen value+JSD rule rejects all visits.

Task	Outcome-assisted minus	Mean	Exploratory 98.75% interval
Small	Never	-0.000553	[-0.006776, 0.006412]
Small	Frozen value+JSD	-0.000553	[-0.006776, 0.006412]
Tiny	Never	0.016680	[0.004759, 0.031852]
Tiny	Frozen value+JSD	0.018029	[0.005243, 0.033462]

Tiny's two exploratory intervals are positive. This is evidence for useful repeat-state choices under outcome-assisted selection despite the frozen rules' failed usefulness criterion. It does not establish

that JSD identifies those choices, that another observable-feature predictor can do so, or that this sampling-intensive control is a practical teaching method. Small remains unresolved: its intervals include zero, which neither establishes equivalence nor rules out useful opportunities. Selecting and evaluating positive utility on the same 128 pairs would instead give naive means of 0.011342 on small and 0.022633 on tiny; these reuse outcomes for selection and are not independent performance estimates.

New-student confirmation of a fixed two-pair decision. A subsequent exploratory query-budget analysis motivated a separate confirmation: a two-pair outcome-assisted decision was expected to improve tiny and harm small. Before collecting new outcomes, we fixed the rule, both directional expectations, new student seeds, and the inference procedure. Ten new RL-only students per task were evaluated at two training stages, with eight visits per stage (320 visits). The teachers and training recipes were unchanged. At each visit, paired continuations 0–1 select advice only when their mean difference is strictly positive; separate pairs 64–127 evaluate that decision. There is no half swapping, threshold fitting, or budget selection in this confirmation.

Averaging 16 visits within each student, we bootstrap ten student blocks per task 100,000 times. The two prespecified primary comparisons against never-follow use nominal 97.5% percentile intervals (Bonferroni adjustment for two contrasts). Coverage remains approximate with ten student blocks per task. Small has mean 0.000480 with interval $[-0.012563, 0.014302]$; tiny has mean 0.006904 with interval $[-0.001829, 0.016421]$. Neither expected direction is confirmed, so the joint directional pattern is not reproduced. These zero-containing intervals do not establish equivalence, absence of useful advice, or a between-cohort difference.

The retained 64-pair reference has descriptive mean gain 0.016680 on small and 0.012005 on tiny; these are not additional confirmatory tests. The primary decision requires 1,948 recorded joint transitions per visit; evaluation requires 62,336. Reference acquisition includes primary acquisition and must not be counted as disjoint additional volume. This fixed-teacher, simulator-assisted one-action experiment establishes neither full-training improvement nor practical net benefit.

T AN INTRINSIC COMPARISON AT EQUAL FISHER DISTANCE AND EXACT JOINT KL

Setting. Let $Z \sim \mathrm{Bernoulli}(3/4)$ be hidden from two independent, constant-observation students with action-one probabilities p_1, p_2. Reward is $\mathbf{1}\{A_1 = A_2 = Z\}$. The privileged teacher chooses (Z, Z), attains the maximal return one, and has exact conditional local law $q = (1/4, 3/4)$. Student team return is $R(p) = p_1 p_2 - (p_1 + p_2)/4 + 1/4$. Fix finite weights $w_1 > w_0 \geq 0$ and $m = (3w_1 + w_0)/4 > 0$. With $\ell_z(p) = \mathrm{JSD}(\delta_z, \mathrm{Bernoulli}(p))$ in natural logs, compare population imitation losses

$$L_g = \sum_i \{3w_1 \ell_1(p_i)/4 + w_0 \ell_0(p_i)/4\}, \qquad L_m = m \sum_i \{3\ell_1(p_i)/4 + \ell_0(p_i)/4\}.$$

Weights remain fixed during differentiation. There is no RL gradient.

The interior product-Bernoulli manifold has Fisher metric $ds_F^2 = \sum_i dp_i^2/[p_i(1-p_i)]$. Its chart $\phi_i = 2 \arcsin \sqrt{p_i} \in (0, \pi)$ is isometric to the open Euclidean square. Write G, V for the negative natural gradients of L_g, L_m in this chart, and $D_g = G/\|G\|_2, D_m = V/\|V\|_2$, evaluated at the same initial policy. The updates are the Fisher-geodesic rays $\phi_X(t) = \phi^0 + tD_X$ with fixed initial tangent, not finite logit Euler steps.

Claim. For every fixed $0 < p_1 < p_2 < 1/4$ and finite $w_1 > w_0 \geq 0$, put

$$H(\phi) = R(\sin^2(\phi_1/2), \sin^2(\phi_2/2)), \qquad \delta_F = \nabla H(\phi^0) \cdot (D_g - D_m).$$

Then $\delta_F < 0$. With $B = \min_i(\pi - \phi_i^0) > 0$,

$$H(\phi_g(t)) < H(\phi_m(t)) \quad \text{for all } 0 < t < \min\{B, -8\delta_F/5\}.$$

Both endpoints have intrinsic product-manifold Fisher distance exactly t from the start. Separately, there exists $\kappa_0 > 0$ such that for every $0 < \kappa < \kappa_0$ the local points on the two rays satisfying $\mathrm{KL}(P_0 \| P_X) = \kappa$ exist uniquely and have strictly smaller return in the gated arm. Neither positive radius is asserted uniform over policies or weights.

Proof. Set $s = p(1-p)$, $a(p) = \log((1+p)/p)$, $b(p) = \log((2-p)/(1-p))$, and $K(p) = 3w_1 a(p) - w_0 b(p)$. Differentiating the finite entropy expression for JSD gives $\ell_1' = -a/2$ and $\ell_0' = b/2$. Since $dp/d\phi = \sqrt{s}$ and the chart metric is identity,

$$G_i = \sqrt{s(p_i)} K(p_i)/8, \qquad V_i = m\sqrt{s(p_i)}[3a(p_i) - b(p_i)]/8.$$

For $p < 1/4$, $a > b > 0$, $a' = -1/[p(1+p)] < 0$, and $b' = 1/[(1-p)(2-p)] > 0$. Thus $K > 0$, $K' < 0$, and $x = b/a$ increases strictly. The ratio is $G_i/V_i = f(x(p_i))$ with

$$f(x) = \frac{3w_1 - w_0 x}{m(3-x)}, \qquad f'(x) = \frac{3(w_1 - w_0)}{m(3-x)^2} > 0.$$

Consequently the positive direction slopes satisfy $0 < \tau_m = V_2/V_1 < \tau_g = G_2/G_1 < \sqrt{s(p_2)/s(p_1)}$. The chart return gradient has components $r_1 = \sqrt{s(p_1)}(p_2 - 1/4) < 0$ and $r_2 = \sqrt{s(p_2)}(p_1 - 1/4) < 0$; hence

$$r_2/r_1 = \sqrt{s(p_2)/s(p_1)}\, \frac{1/4 - p_1}{1/4 - p_2} > \sqrt{s(p_2)/s(p_1)} > \tau_g.$$

For a positive unit vector of slope τ, its return derivative is $h(\tau) = (r_1 + r_2\tau)/\sqrt{1+\tau^2}$. Since $h'(\tau) = (r_2 - r_1\tau)/(1+\tau^2)^{3/2} < 0$ on $0 < \tau < r_2/r_1$, we obtain $\delta_F = h(\tau_g) - h(\tau_m) < 0$. Both individual derivatives are negative.

The chart is convex and flat. Positive unit tangents remain interior for $t < B$, so the rays realize intrinsic distance t. The derivatives of $p(\phi) = \sin^2(\phi/2)$ satisfy $|p'| \leq 1/2$ and $|p''| \leq 1/2$. Hessian diagonal entries of H have absolute value at most $3/8$, and off-diagonal entries at most $1/4$. Symmetry and the row-sum bound give $\|\nabla^2 H\|_{\mathrm{op}} \leq 5/8$. For $F(t) = H(\phi_g(t)) - H(\phi_m(t))$, we have $F(0) = 0$, $F'(0) = \delta_F$, and $|F''| \leq 5/4$. The integral Taylor remainder yields $F(t) \leq$

$\delta_F t + 5t^2/8 < 0$ on the stated interval. Its guaranteed upper endpoint is positive for every fixed admissible tuple.

For exact KL, write $q_{X,i}(t) = \sin^2((\phi_i^0 + tD_{X,i})/2)$ and

$$J_X(t) = \mathrm{KL}(P_0 \| P_X(t)) = \sum_i \left[p_i \log \frac{p_i}{q_{X,i}(t)} + (1-p_i) \log \frac{1-p_i}{1-q_{X,i}(t)} \right].$$

Product independence gives this equality. Interior smoothness gives $J_X(0) = J'_X(0) = 0$ and $J''_X(0) = \sum_i q'_{X,i}(0)^2/s(p_i) = \sum_i D_{X,i}^2 = 1$. Thus $J_X(t) = t^2/2 + O(t^3)$ for these fixed rays. For $0 < t < B$, each $q_{X,i}(t) > p_i$ and $q'_{X,i}(t) > 0$. Since the derivative of the Bernoulli KL with respect to q is $(q-p)/[q(1-q)] > 0$, we have $J'_X(t) > 0$. Fix a sufficiently small $b_0 \in (0, B)$. Continuity and strict monotonicity give unique local inverse lengths $\ell_X(\kappa) \in (0, b_0)$ whenever $0 < \kappa < \min_X J_X(b_0)$. The quadratic expansion first implies $\ell_X = O(\sqrt{\kappa})$ and then

$$|\ell_X - \sqrt{2\kappa}| = \frac{|\ell_X^2 - 2\kappa|}{\ell_X + \sqrt{2\kappa}} = O(\kappa).$$

A return expansion along each ray therefore gives

$$H(\phi^0 + \ell_g D_g) - H(\phi^0 + \ell_m D_m) = \sqrt{2\kappa}\, \delta_F + O(\kappa) < 0$$

for sufficiently small positive κ. All remainder constants and neighborhoods depend on the fixed policy and weights; only a finite pair of rays is involved. KL equality is exact by construction.

Gate interpretation and boundaries. Point-mass JSD decreases strictly with the probability assigned to its target, so $d_1 = \mathrm{JSD}(\delta_1, q) < d_0 = \mathrm{JSD}(\delta_0, q)$. Every finite exponential gate $w_z = e^{-\alpha d_z}$, $\alpha > 0$, satisfies the weight assumptions. The case $w_0 = 0$ is an ordered-weight extension. Within $0 < p_1 = p_2 < 1/4$, the normalized directions coincide by symmetry. For equal positive weights and probabilities in the stated below-$1/4$ region, they also coincide. Boundary probabilities are excluded. Parameter sharing that forces equal probabilities is outside the strict-order family.

Scope. Intrinsic distance here belongs to the product-policy manifold, not the ambient simplex allowing correlated policies. The specified natural gradient and Riemannian exponential are invariant under smooth coordinate changes of this manifold. This does not make the earlier Euclidean-logit algorithm invariant. Equal KL does not mean the step optimizes return under a KL constraint. No PPO claim, benchmark mechanism, useful decision rule, operational effect size, or long-run harm follows. The finite grid in the accompanying numeric check is illustrative, not a proof or a general practical-effect estimate.