# WHAT TEAMMATES LEARN NEXT CHANGES WHICH UPDATES ARE WORTH MAKING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Off-policy cooperative multi-agent reinforcement learning (MARL) reuses past interactions through experience replay, but limited update budgets permit learning from only a fraction of candidates, while each update's team utility depends on teammates' subsequent learning and responses. However, existing replay and data-attribution criteria largely use current policies and error signals, without accounting for subsequent teammate learning. The most valuable update with teammates held fixed need not remain optimal under co-adaptation, a phenomenon we call Co-Adaptive Replay-Value Misalignment (CoRVM). To address CoRVM, we propose ExSpectra, which estimates co-adaptive update utility before replay allocation. Co-Adaptive Utility Decomposition (CAUD) exactly decomposes co-adaptation-induced utility changes under fixed-data matched interventions into baseline teammate-learning and update-induced teammate-response effects. Based on CAUD, we construct a first-order local influence estimator that augments frozen-teammate utility with co-adaptive corrections using cross-agent evaluation curvature and update sensitivity. Shared Hessian–vector products and cross-agent curvature estimates amortize second-order computation across candidates, while sparse paired probes calibrate residual error. Held-out matched audits across five benchmark families confirm that CoRVM changes finite-budget replay decisions. Extensive experiments show that ExSpectra outperforms strong replay-based and resource-matched baselines on most tasks. Code is anonymously available at https://anonymous.4open.science/r/ExSpectra-0DC0/

## 1 INTRODUCTION

Off-policy cooperative MARL reuses past interactions through experience replay to improve sample efficiency (Yang et al., 2024). Under a limited per-round update budget, replay candidates compete for a finite number of updates, making their relative utility central to selection. This utility, however, depends on teammates' subsequent learning and responses. Beyond the familiar mismatch between historical experience and evolving teammate policies (Foerster et al., 2017a), the most valuable replay update with teammates held fixed need not remain optimal under co-adaptation.

Existing replay methods prioritize experiences using signals. PER uses TD error (Schaul et al., 2015), ReLo uses reducible loss (Sujit et al., 2023), and ReaPER adjusts TD-based priority by target-value reliability (Pleiss et al., 2026). In cooperative MARL, DIFFER derives individual replay priorities from decomposed team rewards (Hu et al., 2023), while MAC-PO optimizes sampling weights by minimizing multi-agent policy regret (Mei et al., 2023). Training data affect subsequent optimization, as studied by trajectory-dependent attribution (Wang et al., 2025), local online-RL attribution (Hu et al., 2026), and replay-based non-local attribution (Liu et al.). When learning changes the distribution of future data, replay-side information alone need not identify the corresponding interventional effect (Rege, 2026). These approaches account for richer replay signals or temporal training effects, but they do not account for how a replay update's marginal team utility changes when agents subsequently continue learning and respond.

We use a simple collaborative door-opening scenario as an example: one agent holds down a pressure plate while the other passes through the resulting open doorway. An update on a plate-holding sequence can become more useful than an alternative update as other the agent learns to cross, even when that learning path is unaffected by the focal replay update. The focal update can also alter that



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

agent's subsequent learning through the coupled training objective, even on the same continuation data. Either effect can reverse the replay ranking obtained with teammates held fixed.

To make this distinction precise, we define $u_k^{\mathrm{co}}(e)$ as the expected team-return change from replacing a reference replay unit with $e$ and running a short fixed-data continuation with teammates learning, and $u_k^{\mathrm{frz}}(e)$ as the same replacement effect with subsequent teammate learning frozen. Their difference, $\Gamma_k(e) = u_k^{\mathrm{co}}(e) - u_k^{\mathrm{frz}}(e)$, isolates the teammate-learning contribution to replay utility. A ranking reversal occurs when $u_k^{\mathrm{frz}}(e) > u_k^{\mathrm{frz}}(e')$ but $u_k^{\mathrm{co}}(e) < u_k^{\mathrm{co}}(e')$. Under top-$B$ selection, CoRVM occurs when such a reversal crosses the selection boundary and changes the selected set.

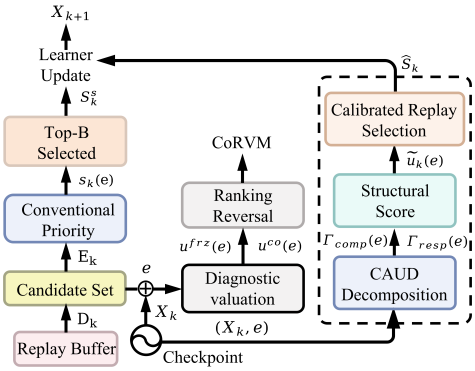


Figure 1: The ExSpectra model architecture.

To address this mismatch, we propose ExSpectra (**Figure 1**), which replaces conventional replay scoring with a structural estimate of co-adaptive update utility before allocation. CAUD exactly decomposes the utility shift $\Gamma_k(e)$ into baseline-learning complementarity and update-induced teammate response. Their distinct derivative channels yield candidate-shared coefficients, so scoring each candidate requires only one focal gradient and a few inner products rather than a separate continuation. Sparse matched probes then calibrate the resulting structural score $\widetilde{u}_k(e)$ using only earlier checkpoints, after which budgeted sampling produces $\widehat{S}_k$ for the focal learner update and the next replay round.

In summary, the contributions of this work are as follows:

- We formalize Co-Adaptive Replay-Value Misalignment (CoRVM), show that identical current replay summaries can imply opposite optimal replay choices and derive matched-intervention decomposition with a first-order characterization of co-adaptive mechanisms.
- We propose ExSpectra, whose allocation-time estimator is dictated by two CAUD channels; shared derivative computation amortizes their cross-agent influence terms, while sparse paired probes calibrate the remaining approximation error.
- We conduct extensive experiments across five benchmark families. We evaluate decision-level CoRVM, mechanism fidelity, estimation quality, and learning efficiency, demonstrating that ExSpectra achieves superior performance over strong baselines.

## 2 Preliminaries

We begin by formalizing finite-budget replay allocation in cooperative MARL and its co-adaptive utility objective, then characterize Co-Adaptive Replay-Value Misalignment (CoRVM) as the decision-level failure induced by ignoring subsequent teammate learning.

### 2.1 Cooperative MARL with Experience Replay

We consider a cooperative Dec-POMDP with $N$ agents, local action–observation histories $\tau_t^i$, and shared rewards $|r_t| \le r_{\max}$ at environment step $t$. Our main instantiation uses non-shared QMIX (Rashid et al., 2020) under centralized training with decentralized execution (CTDE). Online parameters $\mathbf{\Theta} = (\boldsymbol{\theta}^{1:N}, \boldsymbol{\psi})$ are partitioned into $\mathcal{R}$ (focal agent and mixer) and $\mathcal{T}$ (remaining agents), enabling separate interventions on teammate learning. The full checkpoint $X_k$ also includes targets and optimizer states with matching ownership, update counters, and RNG state. For local valuation with discount $\gamma \in [0, 1)$, define

$$J(\mathbf{\Theta}) \triangleq \mathbb{E}_{\boldsymbol{\pi}_{\mathbf{\Theta}}^{\mathrm{eval}}}\left[\sum_{t=0}^{T_{\mathrm{eval}}-1} \gamma^t r_t\right], \tag{1}$$

where $\pi_{\boldsymbol{\theta}^i}^{\mathrm{eval}}(a \mid \tau) \propto \exp(Q_i(\tau, a; \boldsymbol{\theta}^i)/\beta_{\mathrm{eval}})$ over available actions, with fixed $\beta_{\mathrm{eval}} > 0$. The initial-state distribution and dynamics are fixed and parameter-independent. Each replay unit $e$ is a

fixed-length joint sequence with loss $\ell(e; X_k) = \frac{1}{2|\mathcal{I}(e)|} \sum_{t \in \mathcal{I}(e)} (Q_{\mathrm{tot},t} - y_t)^2$, where $\mathcal{I}(e)$ excludes burn-in and $y_t$ is the target-network TD target, detached in the inner update (Tirumala et al., 2024; Pleiss et al., 2026). At learner update $k$, select $B < M_k = |\mathcal{E}_k|$ units under a fixed update rule and a candidate-independent round-robin focal schedule (Hu et al., 2023).

### 2.2 Matched Local Replay Utility

Draw a candidate-independent reference batch $\mathcal{B}_k^0 = (e_1^0, \ldots, e_B^0)$ from $q_{\mathrm{ref}}$ (Hu et al., 2026), and let $\mathcal{B}_{k,\iota \to e}^0$ replace a uniformly sampled slot $\iota \in \{1, \ldots, B\}$ with $e$. The initial update changes only $\mathcal{R}$-owned state, apart from candidate-independent bookkeeping. Before selection, fix a horizon of $H$ learner updates and $\mathcal{P}_k$, comprising the $H - 1$ continuation minibatches, update masks, learning rates, target-copy schedule, and random keys. Under co, teammates follow the prescribed updates; under frz, their complete owned state is fixed. The active block uses the same update rule in both. Let $\Phi_k^p(\mathcal{B})$ return terminal online parameters after the initial update on $\mathcal{B}$ and continuation. Conditioning on $\mathcal{F}_k$, containing $X_k$, $\mathcal{E}_k$, $\mathcal{B}_k^0$, $\mathcal{P}_k$, and completed historical calibration information, define

$$u_k^p(e) \triangleq \mathbb{E}_\iota \left[ J\big(\Phi_k^p(\mathcal{B}_{k,\iota \to e}^0)\big) - J\big(\Phi_k^p(\mathcal{B}_k^0)\big) \mid \mathcal{F}_k \right]. \tag{2}$$

Uniform slot sampling and paired evaluation with correct branch marginals yield unbiased estimates, including under common random numbers. Fixed continuation data exclude data-collection effects.

### 2.3 Co-Adaptive Replay-Value Misalignment

Let $\mathcal{S}_k = \{S \subseteq \mathcal{E}_k : |S| = B\}$. With $S$ as a fixed-order replacement batch, define $U_k^p(S) = J(\Phi_k^p(S)) - J(\Phi_k^p(\mathcal{B}_k^0))$. Generally, $U_k^p(S) \neq \sum_{e \in S} u_k^p(e)$. For $p \in \{\mathrm{frz}, \mathrm{co}\}$, let $S_k^{p,*} \in \arg\max_{S \in \mathcal{S}_k} U_k^p(S)$ with fixed tie-breaking, and $\Delta U_{\mathrm{sel},k} \triangleq U_k^{\mathrm{co}}(S_k^{\mathrm{co},*}) - U_k^{\mathrm{co}}(S_k^{\mathrm{frz},*}) \geq 0$. CoRVM occurs when $\Delta U_{\mathrm{sel},k} > 0$. The exact frozen-utility maximizer is suboptimal under co-adaptation. This diagnostic is unavailable during allocation and independent of proxy scores.

**Proposition 1.** *For $B = 1$ and $H = 2$, there exist two equally likely continuation protocols for a two-agent monotone value-mixing learner under fixed-target SGD that share the checkpoint (hence Q-values, policies, candidates, and TD errors) but yield opposite co-adaptive rankings. Any selector using only these summaries, including the frozen-utility maximizer, incurs strictly positive expected co-adaptive regret, whereas the exact-derivative structural score in Eq. equation 6 does not.*

**Remark 1.** *The appendix gives a construction with zero induced teammate response, where teammate drift is determined by continuation data in $\mathcal{P}_k$. Thus, the information restriction concerns the listed candidate-level summaries, not the full pre-allocation information $\mathcal{F}_k$, which includes $\mathcal{P}_k$.*

## 3 Method

Motivated by CoRVM, ExSpectra estimates co-adaptive replay utility through three components (**Figure 2**). CAUD decomposes the utility shift into complementarity and induced response; shared influence estimation and sparse calibration turn these components into efficient replay decisions.

### 3.1 Co-Adaptive Utility Decomposition

Under scr, overwrite teammate-owned states at every continuation boundary, including the terminal boundary, with their reference-co values, independently of the candidate. The active block follows its prescribed updates using its own branch state. Reference-scr therefore coincides with reference-co. Defining $u^{\mathrm{scr}}$ by **Eq. 2** gives the exact decomposition

$$u^{\mathrm{co}}(e) = u^{\mathrm{frz}}(e) + \underbrace{u^{\mathrm{scr}}(e) - u^{\mathrm{frz}}(e)}_{\Gamma_{\mathrm{comp}}(e)} + \underbrace{u^{\mathrm{co}}(e) - u^{\mathrm{scr}}(e)}_{\Gamma_{\mathrm{resp}}(e)}. \tag{3}$$

These corrections separate reference-learning effects from candidate-induced responses; without scr, the matched co and frz branches identify only their sum.

For the local analysis, specialize all matched branches to unclipped plain SGD on averaged batch losses, with fixed TD targets and no optimizer memory. The initial replacement displacement is

$$\mathbf{d}_{e,\iota}^{\mathcal{R}} \triangleq -\frac{\eta_0}{B} \left[ \nabla_{\mathcal{R}} \ell(e; X_k) - \nabla_{\mathcal{R}} \ell(e_\iota^0; X_k) \right], \tag{4}$$

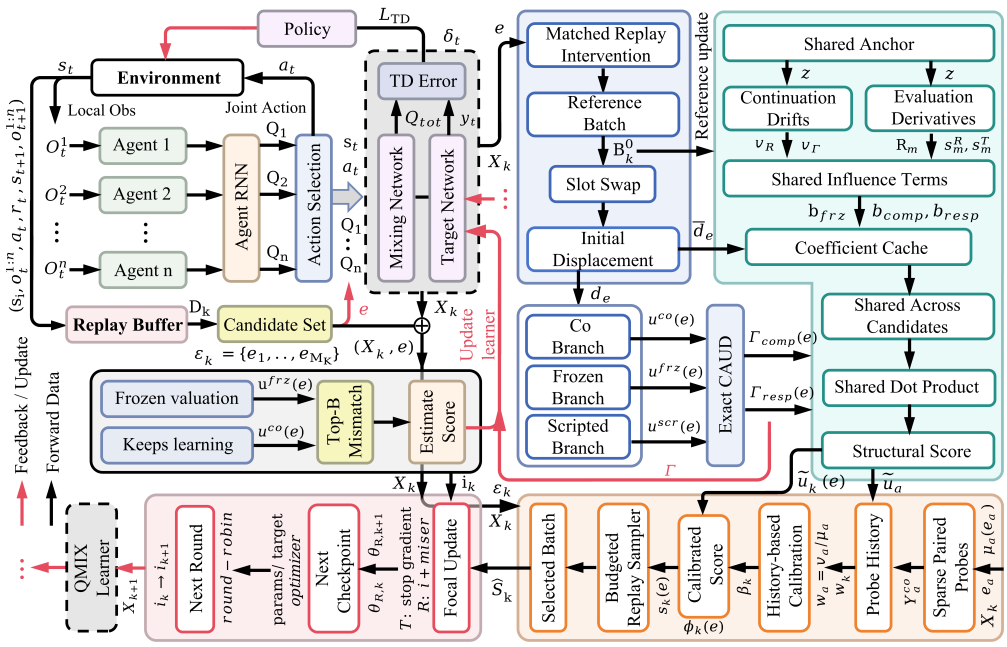



Figure 2: **ExSpectra framework architecture.** ExSpectra addresses Co-Adaptive Replay-Value Misalignment (CoRVM), where frozen-teammate replay rankings can differ from utilities realized after team co-adaptation. Matched co-adaptive, frozen, and scripted continuations enable CAUD to decompose replay utility into frozen, complementarity, and induced-response components. Shared Influence Estimation converts these mechanisms into candidate-shared coefficients, scoring each candidate by its focal gradient and inexpensive dot products. Sparse paired probes calibrate these structural scores, after which budgeted replay selects the update batch. The focal agent and mixer are updated to produce the next checkpoint, closing the replay–learning loop.

with $\mathbf{d}_{e,\iota} = (\mathbf{d}^{\mathcal{R}}_{e,\iota}, \mathbf{0})$. Let $\mathbf{z}$ be the post-reference parameters, so the treatment starts at $\mathbf{z} + \mathbf{d}_{e,\iota}$. For $H = 2$, continuation is synchronous SGD, $\mathbf{z}^+ = \mathbf{z} - \eta \mathbf{G}(\mathbf{z})$. Write $D_{\mathcal{R}}$ for the Jacobian with respect to $\mathcal{R}$, and define $\mathbf{A}^{\mathcal{T}\mathcal{R}} = D_{\mathcal{R}}\mathbf{G}_{\mathcal{T}}$, $\mathbf{C}_{\mathcal{R}\mathcal{T}} = \nabla^2_{\mathcal{R}\mathcal{T}} J$, $\bar{\mathbf{d}}_e = \mathbb{E}_\iota[\mathbf{d}^{\mathcal{R}}_{e,\iota}]$, and $D_e = \max_\iota \|\mathbf{d}_{e,\iota}\|$.

**Theorem 1.** *Under fixed-target SGD specialization, suppose $J$ and $G$ have bounded derivatives through orders three and two, near the matched paths. For $H = 2$, with all terms evaluated at $\mathbf{z}$,*

$$\begin{cases} \Gamma_{\text{comp}}(e) = -\eta\, \bar{\mathbf{d}}_e^\top \mathbf{C}_{\mathcal{R}\mathcal{T}} \mathbf{G}_{\mathcal{T}} + r_c(e), \\ \Gamma_{\text{resp}}(e) = -\eta\, \bar{\mathbf{d}}_e^\top (\mathbf{A}^{\mathcal{T}\mathcal{R}})^\top \nabla_{\mathcal{T}} J + r_r(e). \end{cases} \tag{5}$$

*Here $|r_c(e)| + |r_r(e)| = O(\eta D_e^2 + \eta^2 D_e)$. If $D_e = O(\eta)$, the leading terms are $O(\eta^2)$ and the remainders are $O(\eta^3)$.*

*Proof sketch.* A mixed Taylor expansion in the focal perturbation and reference teammate displacement yields the first term. Linearizing the teammate update field yields the second. Bounded derivatives and slot averaging give the remainders (A detailed proof is provided in the appendix.). □

**Remark 2.** *Zero reference drift need not imply zero response; longer continuations involve update-Jacobian products (A detailed analysis is provided in the appendix.).*

### 3.2 Shared Influence Estimation

Under the same SGD specialization, let $\Psi^p_k$ map post-initial-update parameters to terminal online parameters. For scr, hold the reference teammate path fixed when differentiating. Local linearization at $\mathbf{z}$ gives $u^p_k(e) = \bar{\mathbf{d}}_e^\top \mathbf{a}_p + O(\mathbb{E}_\iota \|\mathbf{d}_{e,\iota}\|^2)$, where $\mathbf{a}_p = [D_{\mathcal{R}} \Psi^p_k(\mathbf{z})]^\top \nabla J(\Psi^p_k(\mathbf{z}))$, following trajectory-influence propagation (Wang et al., 2025).

Evaluate each scheduled, masked gradient field at $\mathbf{z}$ with fixed targets, and set $\mathbf{v}_{\mathcal{A}} = \sum_{h=1}^{H-1} \eta_h \mathbf{G}_{\mathcal{A},h}(\mathbf{z})$. For $\mathcal{A}, \mathcal{B} \in \{\mathcal{R}, \mathcal{T}\}$, let $\widehat{\mathbf{g}}_{\mathcal{A}}$ and $\widehat{\mathbf{C}}_{\mathcal{A}\mathcal{B}}$ estimate $\nabla_{\mathcal{A}} J(\mathbf{z})$ and $\nabla^2_{\mathcal{A}\mathcal{B}} J(\mathbf{z})$. Truncating $\mathbf{a}_{\mathrm{co}}$ to terms linear in continuation step sizes gives the structural score

$$\widetilde{u}_k(e) = \bar{\mathbf{d}}_e^\top \Big( \underbrace{\widehat{\mathbf{g}}_{\mathcal{R}} - (D_{\mathcal{R}} \mathbf{v}_{\mathcal{R}})^\top \widehat{\mathbf{g}}_{\mathcal{R}} - \widehat{\mathbf{C}}_{\mathcal{R}\mathcal{R}} \mathbf{v}_{\mathcal{R}}}_{\mathbf{b}_{\mathrm{frz}}} \underbrace{- \widehat{\mathbf{C}}_{\mathcal{R}\mathcal{T}} \mathbf{v}_{\mathcal{T}}}_{\mathbf{b}_{\mathrm{comp}}} \underbrace{- (D_{\mathcal{R}} \mathbf{v}_{\mathcal{T}})^\top \widehat{\mathbf{g}}_{\mathcal{T}}}_{\mathbf{b}_{\mathrm{resp}}} \Big). \tag{6}$$

The three coefficients are candidate-independent. For $H = 2$, $\bar{\mathbf{d}}_e^\top \mathbf{b}_{\mathrm{comp}}$ and $\bar{\mathbf{d}}_e^\top \mathbf{b}_{\mathrm{resp}}$ are the leading terms of $\Gamma_{\mathrm{comp}}(e)$ and $\Gamma_{\mathrm{resp}}(e)$ in Theorem 1.

#### 3.2.1 Return derivatives and the cross-agent product.

At each refresh, collect $n_v$ independent evaluation trajectories $\xi_m$ at $\mathbf{z}$, with discounted returns $R_m$ and scores $\mathbf{s}_m^{\mathcal{A}} = \nabla_{\mathcal{A}} \log p_{\mathbf{z}}(\xi_m)$. Then $\widehat{\mathbf{g}}_{\mathcal{A}} = n_v^{-1} \sum_{m=1}^{n_v} R_m \mathbf{s}_m^{\mathcal{A}}$. Non-shared decentralized policies and parameter-independent environment laws imply $\nabla^2_{\mathcal{R}\mathcal{T}} \log p_{\boldsymbol{\Theta}}(\xi_m) = 0$ on fixed trajectories. Hence $\widehat{\mathbf{C}}_{\mathcal{R}\mathcal{T}} \mathbf{v}_{\mathcal{T}} = n_v^{-1} \sum_{m=1}^{n_v} R_m \langle \mathbf{s}_m^{\mathcal{T}}, \mathbf{v}_{\mathcal{T}} \rangle \mathbf{s}_m^{\mathcal{R}}$, avoiding a cross-agent evaluation HVP. Within-block curvature products use DiCE (Foerster et al., 2018), including the log-likelihood Hessian term. Update directions are held fixed in evaluation-curvature products.

#### 3.2.2 Update sensitivity and mixer propagation.

Automatic differentiation gives fixed-target HVPs. sg fixes the evaluation vector, avoiding duplicate curvature terms. Although $\nabla_{\boldsymbol{\psi}} J = 0$, mixer perturbations affect subsequent TD gradients through update Jacobians.

$$\begin{cases} \mathbf{h}_{\mathrm{resp}} = \nabla_{\mathcal{R}} \left[ \mathbf{v}_{\mathcal{T}}^\top \mathrm{sg}(\widehat{\mathbf{g}}_{\mathcal{T}}) \right], \\ \mathbf{h}_{\mathrm{own}} = \nabla_{\mathcal{R}} \left[ \mathbf{v}_{\mathcal{R}}^\top \mathrm{sg}(\widehat{\mathbf{g}}_{\mathcal{R}}) \right]. \end{cases} \tag{7}$$

The reference-batch gradient is shared; each candidate needs one gradient and shared dot products. The number of shared derivative products is independent of $M_k$, but all candidate gradients count toward compute. Coefficients refresh every $K_v$ updates and are reused only for matching focal blocks. For the configured learner, optimizer-state, clipping, and target-copy discrepancies enter the total score error in **Proposition 2**, not the SGD remainder. Linearization, continuation truncation, derivative sampling, and cache staleness also contribute.

### 3.3 Calibration and Budgeted Replay

To correct predictable residual errors, run matched configured-learner probes every $K_p$ updates within its resource budget. Draw probe candidates with recorded propensity $\mu_k(e) = \epsilon_p / M_k + (1 - \epsilon_p) \mu_k^{\mathrm{audit}}(e)$, where $0 < \epsilon_p \leq 1$ and $\mu_k^{\mathrm{audit}}$ is fixed before outcomes. The co, frz, and scr branches yield $Y^{\mathrm{co}}$ for calibration and component labels for mechanism audits. Define $s_k(e) = \widetilde{u}_k(e) + \boldsymbol{\beta}_k^\top \boldsymbol{\varphi}_k(e)$, using pre-probe features with their specification fixed before held-out auditing.

The window $\mathcal{W}_k$ contains only completed probes from earlier checkpoints. Historical features and structural predictions are frozen at their source checkpoints. Fit $\boldsymbol{\beta}_k$ by minimizing $\sum_{a \in \mathcal{W}_k} w_a [Y_a^{\mathrm{co}} - \widetilde{u}_a(e_a) - \boldsymbol{\beta}^\top \boldsymbol{\varphi}_a(e_a)]^2 + \lambda_{\mathrm{ridge}} \|\boldsymbol{\beta}\|^2$, where $\lambda_{\mathrm{ridge}} > 0$ and $w_a = \nu_a(e_a) / \mu_a(e_a)$ reweights recorded probe draws to the target distribution $\nu_a$ on the same source-checkpoint support.

Sample exactly $B$ units without replacement. At draw $j$, for $e$ in the remaining set $\mathcal{A}_j \subseteq \mathcal{E}_k$, use

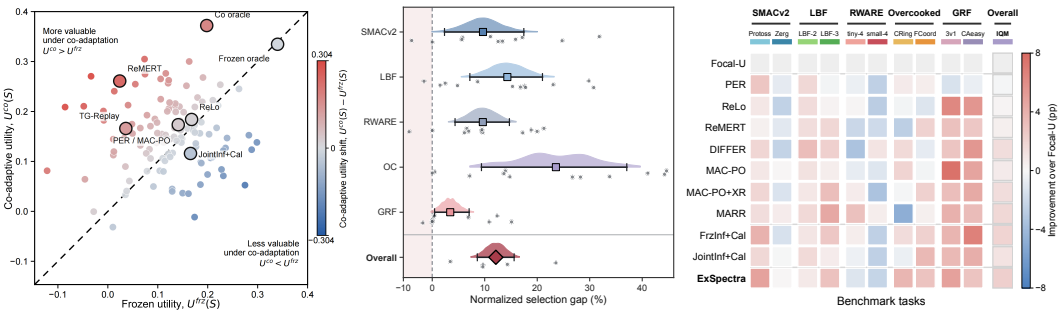
$$q_{k,j}(e) = (1 - \epsilon_s) \frac{\exp(s_k(e)/\alpha_r)}{\sum_{f \in \mathcal{A}_j} \exp(s_k(f)/\alpha_r)} + \frac{\epsilon_s}{|\mathcal{A}_j|}, \tag{8}$$

with fixed temperature $\alpha_r > 0$ and $0 \leq \epsilon_s \leq 1$. The selected set uses the intervention's fixed ordering, loss normalization, focal mask, and update rule.

**Proposition 2.** *Fix a checkpoint and the realized scores. For the configured learner, suppose* $\max_{e \in \mathcal{E}_k} |s_k(e) - u_k^{\mathrm{co}}(e)| \leq \varepsilon_k$ *and* $\max_{S \in \mathcal{S}_k} |U_k^{\mathrm{co}}(S) - \sum_{e \in S} u_k^{\mathrm{co}}(e)| \leq \kappa_k$. *Write* $\widehat{V}_k(S) = \sum_{e \in S} s_k(e)$ *and define the sampling shortfall* $\Lambda_k = \max_{S \in \mathcal{S}_k} \widehat{V}_k(S) - \mathbb{E}[\widehat{V}_k(\widehat{S}_k)]$. *Then*

$$\max_{S \in \mathcal{S}_k} U_k^{\mathrm{co}}(S) - \mathbb{E}[U_k^{\mathrm{co}}(\widehat{S}_k)] \leq 2B\varepsilon_k + 2\kappa_k + \Lambda_k. \tag{9}$$

*Expectations are over replay sampling conditional on the checkpoint and scores.*




(a) Frozen vs. co-adaptive utility. (b) Co-adaptive selection gap. (c) nAUC gains over Focal-U.

Figure 3: **CoRVM changes replay decisions.** (a) Frozen and co-adaptive utilities disagree on replay batches. (b) Frozen-optimal selection incurs a positive held-out co-adaptive selection gap. (c) nAUC differences relative to Focal-U are heterogeneous across tasks.

*Proof sketch.* For every feasible $S$, $|U_k^{\mathrm{co}}(S) - \widehat{V}_k(S)| \le B\varepsilon_k + \kappa_k$. Apply this at the true oracle and sampled batch, then insert $\Lambda_k$ (A detailed proof is provided in the appendix.). □

**Remark 3.** *Under* **Proposition 2**, *the expected advantage over the frozen oracle satisfies* $\mathbb{E}[U_k^{\mathrm{co}}(\widehat{S}_k)] - U_k^{\mathrm{co}}(S_k^{\mathrm{frz},*}) \ge \Delta U_{\mathrm{sel},k} - (2B\varepsilon_k + 2\kappa_k + \Lambda_k)$. *With scores fixed, sorting and independent sampler runs estimate* $\Lambda_k$ *(A detailed analysis is provided in the appendix.).*

## 4 Experiments

In the experimental section, we primarily investigate the following four questions. **RQ1**: Does CoRVM change finite-budget replay decisions? **RQ2**: Do the complementarity and response mechanisms in CAUD explain this change? **RQ3**: Does ExSpectra select replay closer to the co-adaptive oracle? **RQ4**: Do these selection gains improve learning under matched total resources?

**Benchmarks.** We evaluate eleven tasks from five cooperative MARL benchmark families: SMACv2 (Ellis et al., 2023) (Terran°, Protoss, and Zerg 5-vs-5), Level-Based Foraging (LBF) (Christianos et al., 2020) (LBF-2 and LBF-3), Multi-Robot Warehouse (RWARE) (Yu et al., 2020) (tiny-4 and small-4), Overcooked (Carroll et al., 2019) (Coordination Ring and Forced Coordination), and Google Research Football (GRF) (Kurach et al., 2020) (3-vs-1 and Counterattack Easy). Terran° is the development task, while a two-agent pressure-plate task is used only for controlled exhaustive replay audits. All end-to-end experiments use QMIX without parameter sharing as the common learning backbone. Detailed environment configurations, replay parameters, audit construction, and hyperparameters are provided in Appendix D.

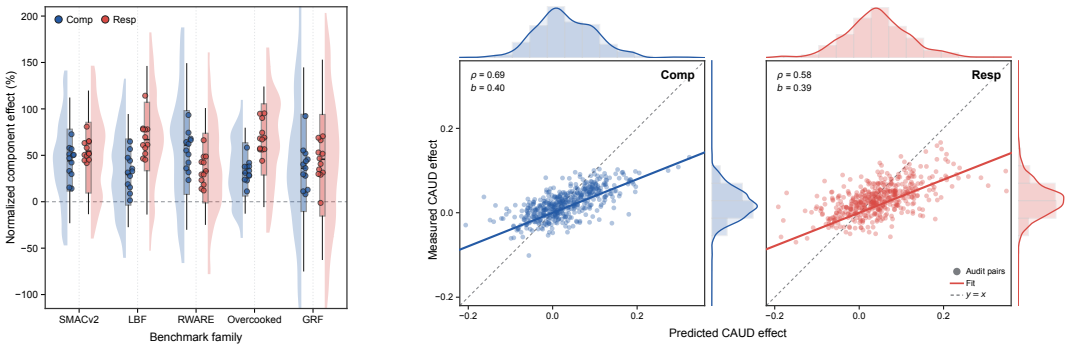
**Baselines.** We compare native QMIX and Focal-U, which uses ExSpectra's candidate-independent focal schedule with uniform replay, together with TG-Replay (Korniak et al., 2026), PER (Schaul et al., 2015), ReaPER (Pleiss et al., 2026), ReLo (Sujit et al., 2023), ReMERT (Liu et al., 2021), DIFFER (Hu et al., 2023), and MAC-PO (Mei et al., 2023). LocalInf+Cal and JointInf+Cal use immediate and full matched-continuation influence, respectively, under the same calibration interface (Hu et al., 2026; Wang et al., 2025). PER+XR, MAC-PO+XR, and MARR (Yang et al., 2024) provide resource-matched controls. Structural ablations remove complementarity, response, calibration, or both CAUD corrections; full definitions are given in Appendix D.2.

**Metrics.** Our primary metric is the normalized area under the learning curve (nAUC). Replay-selection quality is measured by Oracle Gap Closed (Gap), where $0\%$ corresponds to uniform replay. All scored selectors are evaluated on common audit checkpoints, candidate pools, and held-out utility labels. Detailed information is provided in the appendix.

Table 1: Oracle Gap Closed (Gap, %) under a common cross-fitted co-adaptive replay audit.

| Method | SMACv2 | | | LBF | | RWARE | | Overcooked | | GRF | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Terran° | Protoss | Zerg | LBF-2 | LBF-3 | tiny-4 | small-4 | CRing | FCoord | 3v1 | CAeasy | |
| ***Replay-priority methods*** | | | | | | | | | | | | |
| **TG-Replay** | 27.1 $\pm$2.3 | 25.7 $\pm$2.0 | 27.9 $\pm$2.9 | 28.4 $\pm$2.3 | 26.6 $\pm$2.5 | 28.8 $\pm$2.5 | 27.6 $\pm$1.7 | 20.9 $\pm$1.3 | 18.7 $\pm$1.7 | 25.6 $\pm$1.8 | 19.5 $\pm$2.5 | 25.0 $\pm$0.8 |
| **PER** | 33.6 $\pm$1.5 | 32.1 $\pm$2.2 | 28.2 $\pm$1.4 | 32.7 $\pm$3.1 | 28.4 $\pm$0.6 | 38.5 $\pm$3.7 | 34.9 $\pm$1.4 | 22.8 $\pm$0.9 | 22.5 $\pm$2.6 | 28.5 $\pm$2.5 | 30.0 $\pm$2.2 | 30.1 $\pm$0.9 |
| **ReaPER** | 33.8 $\pm$2.8 | 32.1 $\pm$1.6 | 31.6 $\pm$2.0 | 35.0 $\pm$2.6 | 31.9 $\pm$3.2 | 35.3 $\pm$2.5 | 34.7 $\pm$2.1 | 29.2 $\pm$1.7 | 26.0 $\pm$2.4 | 28.7 $\pm$1.7 | 29.9 $\pm$2.3 | 31.6 $\pm$0.8 |
| **ReLo** | 32.7 $\pm$2.4 | 30.8 $\pm$2.2 | 31.0 $\pm$1.7 | 36.3 $\pm$2.2 | 33.5 $\pm$2.8 | 36.2 $\pm$2.7 | 34.2 $\pm$1.5 | 30.4 $\pm$1.1 | 27.8 $\pm$1.8 | 28.9 $\pm$2.1 | 27.9 $\pm$2.0 | 31.8 $\pm$0.6 |
| **ReMERT** | 42.7 $\pm$4.7 | 39.3 $\pm$3.1 | 37.2 $\pm$2.7 | 46.3 $\pm$2.5 | 43.7 $\pm$2.6 | 39.9 $\pm$2.8 | 37.5 $\pm$1.0 | 31.8 $\pm$3.3 | 28.4 $\pm$0.6 | 32.5 $\pm$2.9 | 29.8 $\pm$0.9 | 36.9 $\pm$0.3 |
| **DIFFER** | 48.5 $\pm$2.0 | 44.7 $\pm$1.8 | 40.5 $\pm$1.6 | 45.9 $\pm$2.0 | 45.1 $\pm$2.1 | 42.7 $\pm$1.7 | 41.6 $\pm$2.3 | 42.8 $\pm$1.5 | 39.7 $\pm$1.6 | 43.2 $\pm$2.2 | 40.0 $\pm$2.1 | 43.0 $\pm$0.6 |
| **MAC-PO** | 53.0 $\pm$3.4 | 49.9 $\pm$1.1 | 39.0 $\pm$1.2 | 47.2 $\pm$1.4 | 48.7 $\pm$2.0 | 41.7 $\pm$2.2 | 40.3 $\pm$1.3 | 43.1 $\pm$0.6 | 41.9 $\pm$1.9 | 45.6 $\pm$2.3 | 46.5 $\pm$3.0 | 45.0 $\pm$1.0 |
| ***Influence-based methods*** | | | | | | | | | | | | |
| **LocalInf+Cal** | 62.9 $\pm$1.7 | 63.9 $\pm$1.3 | 58.5 $\pm$1.4 | 62.9 $\pm$1.9 | 60.0 $\pm$1.8 | 64.1 $\pm$1.8 | 59.7 $\pm$2.1 | 60.4 $\pm$1.5 | 55.5 $\pm$1.7 | 53.8 $\pm$1.9 | 50.0 $\pm$1.6 | 59.0 $\pm$0.5 |
| **JointInf+Cal** | 71.3 $\pm$0.7 | 69.6 $\pm$0.4 | 62.2 $\pm$1.0 | 73.6 $\pm$0.4 | 68.5 $\pm$1.6 | 62.9 $\pm$1.7 | 60.2 $\pm$3.3 | **74.6** $\pm$3.3 | **64.8** $\pm$2.3 | 60.5 $\pm$2.8 | 55.8 $\pm$2.3 | 65.6 $\pm$0.7 |
| **ExSpectra** | **72.2** $\pm$2.0 | **72.2** $\pm$1.8 | **66.1** $\pm$0.8 | **79.0** $\pm$1.9 | **74.8** $\pm$2.7 | **70.4** $\pm$1.6 | **69.6** $\pm$1.9 | 67.3 $\pm$0.9 | 64.7 $\pm$1.9 | **61.2** $\pm$0.2 | **60.0** $\pm$1.1 | **68.7** $\pm$0.3 |
| **w/o Comp.** | **68.1** $\pm$1.5 | **66.5** $\pm$1.5 | **59.4** $\pm$3.4 | **69.7** $\pm$0.8 | **67.5** $\pm$1.3 | **63.9** $\pm$2.6 | **62.6** $\pm$1.2 | 61.3 $\pm$0.9 | 56.2 $\pm$0.6 | **57.5** $\pm$1.1 | 50.8 $\pm$1.1 | **61.9** $\pm$0.4 |
| **w/o Resp.** | 64.1 $\pm$1.9 | 63.1 $\pm$1.8 | 57.5 $\pm$1.0 | 69.1 $\pm$3.9 | 64.7 $\pm$1.2 | 63.4 $\pm$2.0 | 59.7 $\pm$4.6 | 56.6 $\pm$1.8 | 54.8 $\pm$1.9 | 53.7 $\pm$1.7 | 52.4 $\pm$1.0 | 59.7 $\pm$1.4 |
| **w/o Cal.** | 60.9 $\pm$2.8 | 62.3 $\pm$2.6 | 56.6 $\pm$2.5 | 63.1 $\pm$1.5 | 62.9 $\pm$2.6 | 61.2 $\pm$1.6 | 58.3 $\pm$1.8 | **62.3** $\pm$0.8 | **58.4** $\pm$1.8 | 56.8 $\pm$2.4 | **54.9** $\pm$2.0 | 59.8 $\pm$0.4 |
| **w/o CAUD** | 55.9 $\pm$1.3 | 54.8 $\pm$0.6 | 49.7 $\pm$0.8 | 56.3 $\pm$2.6 | 55.0 $\pm$1.9 | 58.8 $\pm$0.9 | 57.1 $\pm$0.6 | 52.6 $\pm$1.1 | 50.4 $\pm$1.8 | 51.3 $\pm$2.6 | 48.4 $\pm$2.6 | 53.7 $\pm$0.4 |
| **Cal-Only** | 51.0 $\pm$1.4 | 49.1 $\pm$3.8 | 44.2 $\pm$2.0 | 52.8 $\pm$1.4 | 48.4 $\pm$2.7 | 51.2 $\pm$1.6 | 51.4 $\pm$3.0 | 30.3 $\pm$0.6 | 43.8 $\pm$0.9 | 50.9 $\pm$1.6 | 51.3 $\pm$1.5 | 47.6 $\pm$0.4 |
| **FrzInf-Only** | 41.6 $\pm$3.1 | 41.6 $\pm$3.1 | 37.7 $\pm$0.6 | 43.2 $\pm$0.9 | 41.2 $\pm$2.0 | 45.9 $\pm$2.4 | 41.8 $\pm$2.6 | 30.8 $\pm$2.2 | 37.4 $\pm$2.2 | 43.8 $\pm$0.9 | 41.5 $\pm$2.1 | 40.6 $\pm$0.2 |

**Note.** Gap (%) $\pm$ standard deviation over three audit seeds; higher is better. 0% and 100% denote uniform replay and the cross-fitted held-out oracle, respectively. Red and blue shades denote the top main methods and structural ablations, respectively.




(a) Measured CAUD components.

(b) First-order CAUD correspondence.

Figure 4: **CAUD mechanisms and first-order correspondence.** (a) Measured complementarity and response effects across five benchmark families. (b) At $H = 2$, first-order predictions correlate with matched-audit measurements for complementarity ($\rho = 0.69$) and response ($\rho = 0.58$).

### 4.1 RQ1: CoRVM changes replay decisions.

**Figure 3(a)** shows that frozen and co-adaptive utilities can rank the same replay batches differently. This disagreement is decision-consequential: **Figure 3(b)** shows a positive held-out co-adaptive selection gap across all five benchmark families. Thus, valuing replay against frozen teammates can change which updates should receive the limited replay budget.

### 4.2 RQ2: CAUD captures the mechanisms behind CoRVM.

**Figure 4(a)** shows substantial and heterogeneous complementarity and response effects across all benchmark families; neither component collapses to zero. Moreover, the corresponding first-order channels predict their matched-audit variation with Spearman correlations of $\rho = 0.69$ and $\rho = 0.58$, respectively (**Figure 4(b)**). These results connect the CAUD decomposition to measurable cross-agent learning effects rather than treating residual co-adaptation gap as an undifferentiated quantity.

Table 2: Normalized area under the learning curve (nAUC, %) under matched total resources.

| Method | SMACv2 | | | LBF | | RWARE | | Overcooked | | GRF | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Terran° | Protoss | Zerg | LBF-2 | LBF-3 | tiny-4 | small-4 | CRing | FCoord | 3v1 | CAeasy |
| ***Basic controls*** | | | | | | | | | | | |
| **QMIX** | 42.9±2.3 | 41.4±4.0 | 27.1±2.6 | 56.6±2.5 | 36.9±5.4 | 45.2±8.5 | 33.1±6.2 | 32.8±5.9 | 25.9±3.2 | 41.2±5.4 | 32.9±3.9 |
| **Focal-U** | 42.8±3.5 | 40.5±3.7 | 27.8±3.6 | 58.2±1.1 | 37.9±3.4 | 45.8±7.0 | 33.0±6.6 | 34.1±4.6 | 25.8±4.2 | 41.1±4.7 | 32.8±4.7 |
| **TG-Replay** | 43.7±3.7 | 44.4±3.2 | 29.9±3.5 | 57.3±3.2 | 37.5±4.3 | 46.1±5.1 | 34.3±5.6 | 33.1±3.9 | 28.1±5.0 | 43.3±8.0 | 33.3±5.9 |
| ***Replay-priority methods*** | | | | | | | | | | | |
| **PER** | 43.9±3.3 | 45.1±5.0 | 27.8±2.7 | 59.4±2.7 | 40.0±3.9 | 46.0±7.2 | 31.9±6.4 | 36.2±4.0 | 27.3±5.6 | 40.2±5.2 | 31.9±4.6 |
| **ReaPER** | 43.7±4.4 | 42.9±4.2 | 28.7±3.5 | 59.3±1.8 | 39.8±5.6 | 47.6±6.7 | 34.2±7.0 | 36.4±3.7 | 29.1±5.6 | 45.4±6.5 | 35.3±4.3 |
| **ReLo** | 45.7±3.0 | 44.9±3.2 | 30.3±4.4 | 58.7±1.8 | 40.7±5.1 | 47.9±8.0 | 32.1±7.0 | 36.8±3.5 | 29.3±4.7 | 46.5±7.0 | 34.1±6.8 |
| **ReMERT** | 44.7±4.6 | 42.2±2.6 | 29.8±3.5 | 60.5±1.9 | 41.5±5.6 | 46.8±6.8 | 35.8±6.2 | 32.8±4.8 | 27.0±4.7 | 41.9±4.7 | 35.6±7.0 |
| **DIFFER** | 46.3±4.0 | 45.4±3.6 | 29.4±2.5 | 59.1±2.8 | 41.3±4.5 | 46.6±5.5 | 34.8±6.9 | 37.3±3.3 | 28.3±3.2 | 41.3±4.2 | 35.2±4.6 |
| **MAC-PO** | 47.5±2.9 | 44.3±3.0 | 29.6±3.8 | 60.2±3.3 | 42.4±3.9 | 49.9±7.4 | 36.7±5.4 | 37.7±4.0 | 28.8±2.5 | 45.2±4.9 | 36.3±5.6 |
| ***Influence-based methods*** | | | | | | | | | | | |
| **LocalInf+Cal** | 47.1±3.2 | 45.8±4.0 | 32.8±3.1 | 59.9±3.2 | 40.8±4.3 | 46.7±6.4 | 35.9±6.7 | 37.0±3.4 | 28.0±3.7 | 45.5±5.9 | 38.1±3.8 |
| **JointInf+Cal** | 48.2±3.7 | 45.1±4.1 | 33.5±3.6 | 60.7±2.3 | 42.5±5.8 | 47.9±6.1 | 36.6±6.0 | 38.4±5.2 | 32.2±5.2 | **47.7**±4.2 | 37.8±5.5 |
| ***Resource-matched methods*** | | | | | | | | | | | |
| **PER+XR** | 46.3±2.4 | 43.2±2.5 | 30.0±3.7 | 59.4±2.8 | 43.4±4.3 | 47.8±5.0 | 33.4±5.5 | 36.0±4.6 | 29.2±5.2 | 43.6±8.5 | 36.6±6.2 |
| **MAC-PO+XR** | 48.7±4.1 | 47.5±2.5 | 32.2±4.0 | 59.5±2.4 | 44.6±3.9 | 47.9±7.0 | 35.9±6.0 | 38.9±3.9 | 32.6±2.8 | 45.3±4.6 | 39.4±4.3 |
| **MARR** | 46.9±4.2 | 45.7±3.1 | **35.5**±4.8 | 59.3±2.0 | 41.4±4.1 | **52.0**±9.3 | **40.8**±4.8 | 35.5±5.5 | 30.2±4.1 | 46.6±5.1 | 34.6±4.9 |
| **ExSpectra** | **50.9**±2.9 | **49.2**±3.1 | 32.8±3.9 | **60.8**±3.7 | **47.1**±3.6 | 51.5±6.1 | 35.5±7.3 | **41.5**±3.2 | **33.9**±4.4 | 46.2±5.8 | **40.0**±3.8 |
| **w/o Comp.** | **48.5**±4.1 | **48.1**±2.8 | **31.9**±3.7 | 60.0±3.0 | **44.7**±3.9 | 50.2±5.2 | 36.1±7.1 | 39.0±3.2 | 32.5±3.4 | 45.2±5.9 | 38.5±3.9 |
| **w/o Resp.** | 48.1±2.5 | 47.0±3.5 | 31.5±3.9 | **60.5**±3.5 | 44.1±4.0 | 49.8±5.6 | 35.4±7.1 | 39.4±3.1 | 32.6±3.8 | 46.3±5.1 | 38.0±4.2 |
| **w/o Cal.** | 48.0±2.8 | 47.1±4.7 | 30.9±3.7 | 56.9±3.0 | 44.2±5.3 | **54.5**±5.8 | **38.1**±7.5 | **39.7**±4.1 | **33.0**±3.9 | 45.4±5.9 | **38.9**±4.2 |
| **w/o CAUD** | 45.0±4.1 | 45.7±2.8 | 31.0±5.5 | 60.1±2.3 | 42.2±4.2 | 49.6±6.2 | 37.0±5.4 | 36.9±3.2 | 31.6±3.5 | **46.8**±6.7 | 36.8±5.1 |
| **Cal-Only** | 46.9±2.4 | 44.6±3.6 | 29.5±4.1 | 50.0±4.0 | 39.0±4.3 | 48.3±7.2 | 33.3±8.0 | 35.0±2.7 | 28.4±6.3 | 43.1±6.8 | 37.5±3.7 |
| **FrzInf-Only** | 44.7±2.7 | 43.3±3.7 | 28.4±4.4 | 48.7±3.2 | 37.9±3.9 | 46.2±9.6 | 31.5±7.6 | 36.5±3.1 | 29.5±3.9 | 44.3±7.2 | 38.9±3.9 |

**Note.** Entries report nAUC (%) as mean ± SD over eight matched seeds; higher is better. Rankings use unrounded means. Red and blue shades mark the top three main methods and structural ablations separately; only the best is boldfaced. Terran° is the development task; XR denotes matched extra resources. Cal-Only and FrzInf-Only retain only calibration and frozen influence.

### 4.3 RQ3: ExSpectra better recovers co-adaptive replay value.

**Table 1** shows that ExSpectra closes 68.7% of the oracle gap overall, compared with 65.6% for JointInf+Cal and 45.0% for the strongest replay-priority baseline, MAC-PO, and ranks first on 9/11 tasks. Removing complementarity, response, or calibration reduces the overall Gap to 61.9%, 59.7%, and 59.8%, respectively, while removing both CAUD corrections lowers it further to 53.7%. Hence both co-adaptive structure and residual calibration contribute to replay allocation quality.

### 4.4 RQ4: Better selection improves matched-resource learning.

Under the matched-resource protocol, ExSpectra achieves the highest mean nAUC on 7/11 tasks in **Table 2**, with its largest non-development gains on LBF-3 and Coordination Ring. The advantage is not universal: MARR is stronger on Zerg and both RWARE tasks, while JointInf+Cal is strongest on GRF 3v1. Removing either complementarity or response lowers mean nAUC on 10/11 tasks, although many task-level differences are small relative to run variation. Removing calibration instead raises nAUC on both RWARE tasks despite lowering Gap, indicating that the relationship between short-horizon selection quality and learning is task dependent.

### 4.5 Sensitivity analysis.

**Table 3** shows that $H = 4$ and $|\mathcal{W}| = 256$ provide stable operating points, whereas $H = 1$ and a small calibration window degrade selection on several tasks. $K_p$ and $K_v$ instead control information freshness: more frequent probing or derivative refresh generally improves Gap, while sparse probing ($K_p = 2000$) and stale derivatives ($K_v = 500$) degrade it. The defaults $(H, K_p, K_v, |\mathcal{W}|) = (4, 500, 200, 256)$ are shared settings fixed for the main experiments rather than per-task optima.

Table 3: Sensitivity of ExSpectra nAUC and replay-selection Gap (%) to core parameters.

| Parameter | Value | SMACv2 Zerg | | LBF LBF-3 | | RWARE small-4 | | Overcooked FCoord | | GRF CAeasy | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | nAUC | Gap | nAUC | Gap | nAUC | Gap | nAUC | Gap | nAUC | Gap |
| $H$ | 1 | $25.3_{\pm 3.0}$ | $42.7_{\pm 7.8}$ | $45.6_{\pm 5.8}$ | $62.8_{\pm 4.8}$ | $29.2_{\pm 10.2}$ | $48.5_{\pm 2.4}$ | $27.1_{\pm 3.9}$ | $50.2_{\pm 7.2}$ | $35.6_{\pm 4.3}$ | $41.8_{\pm 4.7}$ |
| | 2 | $32.1_{\pm 3.9}$ | $62.6_{\pm 3.8}$ | $\mathbf{48.2_{\pm 3.4}}$ | $71.9_{\pm 3.5}$ | $33.0_{\pm 4.0}$ | $65.3_{\pm 14.4}$ | $33.3_{\pm 8.7}$ | $60.8_{\pm 6.1}$ | $38.6_{\pm 2.8}$ | $53.7_{\pm 6.7}$ |
| | 4 | $\mathbf{33.1_{\pm 3.9}}$ | $\mathbf{66.1_{\pm 5.9}}$ | $47.5_{\pm 4.0}$ | $75.1_{\pm 6.6}$ | $\mathbf{34.1_{\pm 9.2}}$ | $\mathbf{69.9_{\pm 3.5}}$ | $\mathbf{33.5_{\pm 5.3}}$ | $\mathbf{64.3_{\pm 4.7}}$ | $40.5_{\pm 4.5}$ | $\mathbf{59.9_{\pm 2.5}}$ |
| | 8 | $31.6_{\pm 4.0}$ | $65.9_{\pm 3.3}$ | $47.2_{\pm 5.4}$ | $\mathbf{75.6_{\pm 6.0}}$ | $31.6_{\pm 7.5}$ | $69.0_{\pm 2.2}$ | $\mathbf{33.5_{\pm 5.2}}$ | $64.1_{\pm 9.4}$ | $\mathbf{42.9_{\pm 5.5}}$ | $57.5_{\pm 2.4}$ |
| $K_p$ | 250 | $\mathbf{34.2_{\pm 5.2}}$ | $\mathbf{66.2_{\pm 5.4}}$ | $48.0_{\pm 4.4}$ | $\mathbf{76.5_{\pm 9.1}}$ | $\mathbf{36.1_{\pm 7.5}}$ | $\mathbf{71.0_{\pm 8.5}}$ | $\mathbf{36.0_{\pm 7.5}}$ | $\mathbf{66.4_{\pm 4.5}}$ | $39.9_{\pm 4.0}$ | $\mathbf{61.9_{\pm 5.3}}$ |
| | 500 | $33.1_{\pm 3.9}$ | $66.1_{\pm 5.9}$ | $47.5_{\pm 4.0}$ | $75.1_{\pm 6.6}$ | $34.1_{\pm 9.2}$ | $69.9_{\pm 3.5}$ | $33.5_{\pm 5.3}$ | $64.3_{\pm 4.7}$ | $\mathbf{40.5_{\pm 4.5}}$ | $59.9_{\pm 2.5}$ |
| | 1000 | $32.4_{\pm 3.0}$ | $62.7_{\pm 4.8}$ | $\mathbf{48.3_{\pm 3.7}}$ | $73.4_{\pm 12.1}$ | $36.0_{\pm 7.4}$ | $65.1_{\pm 6.4}$ | $32.2_{\pm 5.0}$ | $61.2_{\pm 1.2}$ | $38.1_{\pm 4.9}$ | $57.4_{\pm 14.9}$ |
| | 2000 | $29.9_{\pm 4.1}$ | $59.7_{\pm 10.3}$ | $46.6_{\pm 3.0}$ | $68.3_{\pm 8.2}$ | $30.2_{\pm 12.1}$ | $61.8_{\pm 4.7}$ | $27.7_{\pm 6.2}$ | $53.3_{\pm 1.6}$ | $36.1_{\pm 4.4}$ | $48.3_{\pm 4.5}$ |
| $K_v$ | 50 | $31.9_{\pm 3.0}$ | $\mathbf{67.0_{\pm 7.2}}$ | $\mathbf{48.6_{\pm 2.3}}$ | $\mathbf{77.2_{\pm 15.2}}$ | $32.6_{\pm 8.7}$ | $69.6_{\pm 4.9}$ | $\mathbf{34.5_{\pm 5.0}}$ | $\mathbf{68.1_{\pm 7.3}}$ | $40.5_{\pm 4.5}$ | $\mathbf{63.2_{\pm 3.5}}$ |
| | 100 | $\mathbf{34.8_{\pm 3.8}}$ | $66.3_{\pm 6.2}$ | $47.9_{\pm 4.1}$ | $75.8_{\pm 5.3}$ | $33.8_{\pm 8.0}$ | $\mathbf{72.0_{\pm 8.9}}$ | $33.4_{\pm 4.3}$ | $64.3_{\pm 4.2}$ | $\mathbf{41.8_{\pm 3.8}}$ | $62.6_{\pm 3.2}$ |
| | 200 | $33.1_{\pm 3.9}$ | $66.1_{\pm 5.9}$ | $47.5_{\pm 4.0}$ | $75.1_{\pm 6.6}$ | $\mathbf{34.1_{\pm 9.2}}$ | $69.9_{\pm 3.5}$ | $33.5_{\pm 5.3}$ | $64.3_{\pm 4.7}$ | $40.5_{\pm 4.5}$ | $59.9_{\pm 2.5}$ |
| | 500 | $31.0_{\pm 1.5}$ | $59.4_{\pm 3.8}$ | $47.5_{\pm 4.0}$ | $70.1_{\pm 9.3}$ | $30.3_{\pm 6.7}$ | $63.3_{\pm 4.6}$ | $29.1_{\pm 7.3}$ | $58.1_{\pm 8.4}$ | $35.5_{\pm 5.5}$ | $52.3_{\pm 7.5}$ |
| $\lvert\mathcal{W}\rvert$ | 64 | $28.1_{\pm 2.4}$ | $59.0_{\pm 2.8}$ | $45.7_{\pm 4.0}$ | $69.3_{\pm 1.4}$ | $29.3_{\pm 8.5}$ | $59.3_{\pm 1.8}$ | $29.0_{\pm 7.3}$ | $55.8_{\pm 9.6}$ | $33.2_{\pm 3.7}$ | $47.7_{\pm 2.7}$ |
| | 128 | $32.0_{\pm 3.5}$ | $61.4_{\pm 13.7}$ | $\mathbf{47.6_{\pm 2.9}}$ | $74.1_{\pm 5.6}$ | $33.1_{\pm 8.0}$ | $66.7_{\pm 6.4}$ | $30.8_{\pm 7.1}$ | $62.1_{\pm 6.8}$ | $39.7_{\pm 4.5}$ | $59.1_{\pm 4.7}$ |
| | 256 | $\mathbf{33.1_{\pm 3.9}}$ | $\mathbf{66.1_{\pm 5.9}}$ | $47.5_{\pm 4.0}$ | $75.1_{\pm 6.6}$ | $\mathbf{34.1_{\pm 9.2}}$ | $\mathbf{69.9_{\pm 3.5}}$ | $\mathbf{33.5_{\pm 5.3}}$ | $\mathbf{64.3_{\pm 4.7}}$ | $\mathbf{40.5_{\pm 4.5}}$ | $\mathbf{59.9_{\pm 2.5}}$ |
| | 512 | $32.7_{\pm 2.4}$ | $64.8_{\pm 7.2}$ | $47.0_{\pm 2.5}$ | $\mathbf{75.7_{\pm 4.4}}$ | $32.4_{\pm 10.1}$ | $69.4_{\pm 9.3}$ | $32.2_{\pm 5.6}$ | $63.3_{\pm 7.4}$ | $40.3_{\pm 6.9}$ | $59.7_{\pm 3.3}$ |

**Note.** All nAUC and Gap entries use common five-seed sensitivity and three-seed audit panels, respectively. The shared default $(H, K_p, K_v, |\mathcal{W}|) = (4, 500, 200, 256)$ is repeated across sweeps and may differ from the eight-seed headline results in Table 2. Gray denotes defaults; blue bold marks the best mean within each block.

## 5 Related Work

**Replay valuation and data attribution.** Replay methods connect sampling to learning through TD error, reducible loss, and bootstrap reliability (Schaul et al., 2015; Sujit et al., 2023; Pleiss et al., 2026). Learning-based replay instead optimizes observed return improvement (Zha et al., 2019), while cooperative methods use decomposed individual rewards or joint-policy regret (Hu et al., 2023; Mei et al., 2023; Zheng et al., 2026). These objectives guide sampling through learning proxies or policy performance, rather than a candidate-level comparison under matched teammate-learning interventions. Data attribution makes the training intervention explicit through occurrence-level removal along a fixed optimization sequence (Wang et al., 2025), attribution within a PPO training round (Hu et al., 2026), and non-local on-policy attribution (Liu et al.). Adaptive-attribution theory further distinguishes fixed-log replay from interventions that let future data collection respond (Rege, 2026). Tracing a perturbation under a specified learning process, however, does not by itself distinguish how its utility changes across teammate-learning interventions.

**Adaptation and learning-aware optimization.** Research on teammate adaptation addresses both historical-data compatibility and anticipated learning responses. Importance weighting (Du et al., 2026) and fingerprints mitigate replay mismatch caused by changing teammate policies (Foerster et al., 2017a), while communication correction relabels stored messages under current communication policies (Ahilan & Dayan, 2020). These corrections address whether historical experience remains suitable for training, rather than its comparative team benefit after subsequent joint updates. LOLA incorporates anticipated co-player learning into policy optimization (Foerster et al., 2017b), while POLA develops a proximal formulation of this learning-aware objective (Zhao et al., 2022). Optimization variable is the policy update, rather than experience receives an existing update slot.

## 6 Conclusion

We showed that replay value in cooperative MARL depends on how teammates continue to learn, proving the existence of CoRVM. Therefore, we introduced CAUD to decompose this effect into complementarity and induced response, and ExSpectra to estimate co-adaptive replay utility for replay selection. Across the benchmarks, ExSpectra selects replay closer to the co-adaptive oracle, and generally improves learning. These results indicate that cooperative replay should be valued with respect to future team adaptation, rather than only the current team state.

## 7 AI USE STATEMENT

We used generative AI tools solely for language polishing and proofreading of the manuscript. We did not use generative AI for experiments and data analysis. The authors take full responsibility for the final content, including all text, analyses, and conclusions.

## 8 REPRODUCIBILITY STATEMENT

All benchmarks used in our experiments are publicly available. We provide implementation details, hyperparameters, source code, and configurations in an anonymous repository to support independent replication.

## REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems*, 34:29304–29320, 2021.

Sanjeevan Ahilan and Peter Dayan. Correcting experience replay for multi-agent communication. *arXiv preprint arXiv:2010.01192*, 2020.

Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32, 2019.

Filippos Christianos, Lukas Schäfer, and Stefano Albrecht. Shared experience actor-critic for multi-agent reinforcement learning. *Advances in neural information processing systems*, 33:10707–10717, 2020.

Xuefeng Du, Jiajun Wu, Yuduo Zheng, and Fengqi Li. IEC: When information-driven exploration meets spectral consensus via primal–dual reward regularization in decentralized MARL. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=FsKoSE1xq8.

Benjamin Ellis, Jonathan Cook, Skander Moalla, Mikayel Samvelyan, Mingfei Sun, Anuj Mahajan, Jakob Foerster, and Shimon Whiteson. Smacv2: An improved benchmark for cooperative multi-agent reinforcement learning. *Advances in neural information processing systems*, 36:37567–37593, 2023.

Jakob Foerster, Nantas Nardelli, Gregory Farquhar, Triantafyllos Afouras, Philip HS Torr, Pushmeet Kohli, and Shimon Whiteson. Stabilising experience replay for deep multi-agent reinforcement learning. In *International conference on machine learning*, pp. 1146–1155. PMLR, 2017a.

Jakob Foerster, Gregory Farquhar, Maruan Al-Shedivat, Tim Rocktäschel, Eric Xing, and Shimon Whiteson. Dice: The infinitely differentiable monte carlo estimator. In *International Conference on Machine Learning*, pp. 1529–1538. PMLR, 2018.

Jakob N Foerster, Richard Y Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. *arXiv preprint arXiv:1709.04326*, 2017b.

Xunhan Hu, Jian Zhao, Wengang Zhou, Ruili Feng, and Houqiang Li. Differ: Decomposing individual reward for fair experience replay in multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 36:75048–75066, 2023.

Yuzheng Hu, Fan Wu, Haotian Ye, David Forsyth, James Zou, Nan Jiang, Jiaqi Ma, and Han Zhao. A snapshot of influence: A local data attribution framework for online reinforcement learning. *Advances in Neural Information Processing Systems*, 38:39044–39077, 2026.

Michal Korniak, Mikołaj Czarnecki, Yarden As, Piotr Miłoś, Pieter Abbeel, and Michal Nauman. When does non-uniform replay matter in reinforcement learning? *arXiv preprint arXiv:2605.10236*, 2026.

Karol Kurach, Anton Raichuk, Piotr Stanczyk, Michal Zajkac, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, et al. Google research football: A novel reinforcement learning environment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 4501–4510, 2020.

Shixuan Liu, Yuzheng Hu, Han Zhao, and Jiaqi W Ma. Non-local data attribution for on-policy reinforcement learning.

Xu-Hui Liu, Zhenghai Xue, Jingcheng Pang, Shengyi Jiang, Feng Xu, and Yang Yu. Regret minimization experience replay in off-policy reinforcement learning. *Advances in neural information processing systems*, 34:17604–17615, 2021.

Yongsheng Mei, Hanhan Zhou, Tian Lan, Guru Venkataramani, and Peng Wei. Mac-po: Multi-agent experience replay via collective priority optimization. *arXiv preprint arXiv:2302.10418*, 2023.

Leonard S Pleiss, Tobias Sutter, and Maximilian Schiffer. Reliability-adjusted prioritized experience replay. In *International Conference on Learning Representations*, volume 2026, pp. 59206–59232, 2026.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51, 2020.

Amit Kiran Rege. Data attribution in adaptive learning. *arXiv preprint arXiv:2604.04892*, 2026.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

Shivakanth Sujit, Somjit Nath, Pedro Braga, and Samira Ebrahimi Kahou. Prioritizing samples in reinforcement learning with reducible loss. *Advances in Neural Information Processing Systems*, 36:23237–23258, 2023.

Dhruva Tirumala, Thomas Lampe, Jose Enrique Chen, Tuomas Haarnoja, Sandy Huang, Guy Lever, Ben Moran, Tim Hertweck, Leonard Hasenclever, Martin Riedmiller, et al. Replay across experiments: A natural extension of off-policy rl. In *International Conference on Learning Representations*, volume 2024, pp. 31533–31549, 2024.

Jiachen Tianhao Wang, Dawn Song, James Y Zou, Prateek Mittal, and Ruoxi Jia. Capturing the temporal dependence of training data influence. In *International conference on learning representations*, volume 2025, pp. 11236–11273, 2025.

Yaodong Yang, Guangyong Chen, Jianye Hao, and Pheng-Ann Heng. Sample-efficient multiagent reinforcement learning with reset replay. In *ICML*, pp. 55961–55975, 2024.

Chao Yu, Akash Velu, Eugene Vinitsky, Yu Wang, Alexandre Bayen, and Yi Wu. Benchmarking multi-agent deep reinforcement learning algorithms. 2020.

Daochen Zha, Kwei-Herng Lai, Kaixiong Zhou, and Xia Hu. Experience replay optimization. *arXiv preprint arXiv:1906.08387*, 2019.

Stephen Zhao, Chris Lu, Roger B Grosse, and Jakob Foerster. Proximal learning with opponent-learning awareness. *Advances in Neural Information Processing Systems*, 35:26324–26336, 2022.

Yuduo Zheng, XueFeng Du, Yanqi Cheng, Li Yin, and Fengqi Li. Eva-gen: When perception learns from value via generative models in decentralized multi-agent systems. In Diego Calvanese (ed.), *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence, IJCAI-26*, pp. 472–480. International Joint Conferences on Artificial Intelligence Organization, 8 2026. doi: 10.24963/ijcai.2026/54. URL https://doi.org/10.24963/ijcai.2026/54. Main Track.

# Appendix

## CONTENTS


**A Formal Setup and Matched Interventions . . . 13**

A.1 Summary of Notation . . . 13

A.2 Learner State and Block Ownership . . . 13

A.3 Pre-Allocation Information and Temporal Holdout . . . 13

A.4 Matched Replay Interventions and Continuation Branches . . . 14

**B Theoretical Analysis and Proofs . . . 14**

B.1 Exact Decomposition and Audit Validity . . . 14

B.2 Proof of Proposition 1 . . . 14

B.3 Two-Step Expansion of the Co-Adaptive Mechanisms . . . 15

B.4 Multi-Step Propagation and Shared Influence Estimation . . . 16

B.5 Finite-Budget Replay Allocation Guarantee . . . 18

**C Algorithmic Details and Implementation . . . 19**

C.1 Replay Scoring and Budgeted Updates . . . 19

C.2 Matched Auditing and Residual Calibration . . . 20

C.3 Caching and Resource Accounting . . . 21

**D Experimental Methodology . . . 21**

D.1 Benchmarks and Unified Configuration . . . 21

D.2 Baselines and Structural Controls . . . 21

D.3 Audit Design and Information Separation . . . 22

D.4 Metrics and Statistical Analysis . . . 22

D.5 Resource Matching . . . 23

**E Additional Experimental Results . . . 23**

E.1 Learning Trajectories under Matched Resources . . . 23

E.2 Audit Reliability and Cross-Fit Stability . . . 24

E.3 Mechanism Fidelity and Allocation Diagnostics . . . 24

E.4 Sensitivity to Core Parameters . . . 24

E.5 Scaling with Team Size . . . 25

# A Formal Setup and Matched Interventions

This section fixes the notation, learner-state ownership, information boundary, and matched continuation rules used throughout the analysis. It specifies learner states copied, frozen, or scripted across intervention branches, and information available before replay allocation. Proof-local symbols and implementation-specific parameters are introduced only where they are needed.

## A.1 Summary of Notation

Table A.1: Main notation.

| Symbol | Definition |
|---|---|
| **Learner and State** | |
| $t,\ k,\ h$ | Environment step, learner-update index, and local continuation index |
| $\boldsymbol{\Theta},\ X_k$ | Online learner parameters and complete learner checkpoint at update $k$ |
| $\mathcal{F}_k$ | Pre-allocation information available before replay selection |
| $\mathcal{R},\ \mathcal{T}$ | Active learner block and teammate learner block |
| **Replay and Intervention** | |
| $\mathcal{D}_k,\ \mathcal{E}_k,\ M_k$ | Replay buffer, candidate set, and candidate count $M_k = \lvert\mathcal{E}_k\rvert$ |
| $e,\ e_\iota^0,\ S,\ B$ | Candidate replay unit, reference unit at slot $\iota$, selected set, and replay budget |
| $\mathcal{B}_k^0,\ q_{\mathrm{ref}}$ | Candidate-independent reference batch and its sampling distribution |
| $\iota,\ \mathcal{B}_{k,\iota\to e}^0$ | Uniform replacement slot and reference batch with $e_\iota^0$ replaced by $e$ |
| $H,\ \mathcal{P}_k$ | Local learner-update horizon and matched continuation protocol |
| $p \in \{\mathrm{frz}, \mathrm{scr}, \mathrm{co}\}$ | Frozen, scripted, and co-adaptive teammate continuation rules |
| $\Phi_k^p(\mathcal{B})$ | Terminal online parameters after the initial update and continuation under rule $p$ |
| **Utility and CAUD** | |
| $J(\boldsymbol{\Theta})$ | Smooth team-return objective used for local replay valuation |
| $u_k^p(e)$ | Matched singleton replay utility under continuation rule $p$ |
| $\Gamma_{\mathrm{comp}}(e),\ \Gamma_{\mathrm{resp}}(e)$ | Complementarity and induced-response components of co-adaptive utility |
| $U_k^p(S)$ | Measured whole-batch replay utility under continuation rule $p$ |
| $\mathcal{S}_k$ | Feasible replay sets, $\mathcal{S}_k = \{S \subseteq \mathcal{E}_k : \lvert S\rvert = B\}$ |
| **Estimation and Calibration** | |
| $\widetilde{u}_k(e),\ s_k(e)$ | Structural utility estimate and calibrated replay score |
| $\widehat{u}_k^{\mathrm{frz}}(e),\ \widehat{\Gamma}_{\mathrm{comp},k}(e)$ | Frozen-utility estimate and estimated complementarity correction |
| $\widehat{\Gamma}_{\mathrm{resp},k}(e)$ | Estimated induced-response correction |
| $\boldsymbol{\varphi}_k(e),\ \boldsymbol{\beta}_k,\ \mathcal{W}_k$ | Calibration features, fitted coefficients, and historical calibration window |
| $\mu_k(e),\ \nu_k(e)$ | Probe propensity and target calibration distribution |
| **Replay Allocation** | |
| $\widehat{V}_k(S),\ \widehat{S}_k$ | Additive score $\widehat{V}_k(S) = \sum_{e\in S} s_k(e)$ and replay set sampled by ExSpectra |

## A.2 Learner State and Block Ownership

Under the candidate-independent round-robin schedule, $\mathcal{R}$ contains the focal agent and mixer, while $\mathcal{T}$ contains the remaining agents. Candidates select replay units, not focal blocks.

Block ownership includes all state governing subsequent updates, such as target networks, optimizer state, and counters. Each branch therefore forks the complete checkpoint $X_k$, including RNG state and shared bookkeeping. Freezing or scripting $\mathcal{T}$ acts on this complete update-relevant state rather than only on its online parameters.

## A.3 Pre-Allocation Information and Temporal Holdout

At update $k$, all quantities used for scoring—including the focal schedule, structural estimates, calibration coefficients, and derivative cache—are determined from $\mathcal{F}_k$ before allocation. The calibration window $\mathcal{W}_k$ uses only completed probe labels from source checkpoints $j < k$. Thus, an audit forked from $X_k$ cannot affect any score evaluated at the same checkpoint.

### A.4 Matched Replay Interventions and Continuation Branches

For replacement slot $\iota$, the treatment batch $\mathcal{B}^0_{k,\iota\to e}$ differs from $\mathcal{B}^0_k$ only by replacing $e^0_\iota$ with $e$. The resulting candidate-dependent perturbation is confined initially to $\mathcal{R}$.

The remaining $H-1$ minibatches, update schedule, target-copy schedule, and exogenous random keys are fixed by $\mathcal{P}_k$ before allocation and shared across branches. Hence the intervention changes the initial replay allocation while holding the subsequent data stream fixed. Under $\mathrm{frz}$, the complete $\mathcal{T}$ state is fixed; under $\mathrm{co}$, teammates follow their prescribed updates; and under $\mathrm{scr}$, their state is replaced at each continuation boundary by the matched reference–$\mathrm{co}$ state. The active block follows its own branch updates throughout. Consequently, reference–$\mathrm{scr}$ coincides with reference–$\mathrm{co}$ by construction.Consequently, reference–$\mathrm{scr}$ coincides with reference–$\mathrm{co}$ by construction.

## B Theoretical Analysis and Proofs

Throughout, $\mathcal{F}_k$ fixes the reference batch and continuation protocol; slot sampling, evaluation sampling, and replay selection are kept distinct.

### B.1 Exact Decomposition and Audit Validity

For $p \in \{\mathrm{frz}, \mathrm{scr}, \mathrm{co}\}$, write $J^p_{e,\iota} = J(\Phi^p_k(\mathcal{B}^0_{k,\iota\to e}))$ and $J^p_0 = J(\Phi^p_k(\mathcal{B}^0_k))$, so that $u^p_k(e) = \mathbb{E}_\iota[J^p_{e,\iota} - J^p_0 \mid \mathcal{F}_k]$ by equation 2. Scripted reference updates copy the teammate states and give the active update the same inputs at each boundary; induction yields $J^{\mathrm{scr}}_0 = J^{\mathrm{co}}_0$. Inserting the scripted treatment gives the exact decomposition

$$\begin{aligned} u^{\mathrm{co}}_k(e) - u^{\mathrm{frz}}_k(e) &= \mathbb{E}_\iota\big[(J^{\mathrm{scr}}_{e,\iota} - J^{\mathrm{scr}}_0) - (J^{\mathrm{frz}}_{e,\iota} - J^{\mathrm{frz}}_0) \mid \mathcal{F}_k\big] \\ &\quad + \mathbb{E}_\iota\big[J^{\mathrm{co}}_{e,\iota} - J^{\mathrm{scr}}_{e,\iota} \mid \mathcal{F}_k\big] \\ &= \Gamma_{\mathrm{comp},k}(e) + \Gamma_{\mathrm{resp},k}(e). \end{aligned} \tag{B.1}$$

The first difference contains the change in reference teammate learning; the second isolates candidate-induced response against a common reference. No smoothness or batch-additivity assumption is used.

For a paired audit, let $Y^p_k(e) = \widehat{J}^p_{e,\iota} - \widehat{J}^p_0$. If each evaluation estimate has the correct conditional marginal mean given $(\mathcal{F}_k, e, \iota)$, uniform slot sampling and the tower property give

$$\begin{aligned} \mathbb{E}[Y^p_k(e) \mid \mathcal{F}_k, e] &= \frac{1}{B}\sum_{\iota=1}^{B} \mathbb{E}[\widehat{J}^p_{e,\iota} - \widehat{J}^p_0 \mid \mathcal{F}_k, e, \iota] \\ &= \frac{1}{B}\sum_{\iota=1}^{B} (J^p_{e,\iota} - J^p_0) = u^p_k(e). \end{aligned} \tag{B.2}$$

Common random numbers preserve this mean when marginals are correct and reduce variance only when the induced covariance is positive. Probe positivity concerns population coverage, not the conditional validity of an already selected candidate.

### B.2 Proof of Proposition 1

We use $B = 1$, $H = 2$, and unclipped plain SGD without optimizer memory, i.e., the specialization of Theorem 1. The two scenarios share the checkpoint and differ only in the continuation minibatch prescribed by $\mathcal{P}_k$.

*Proof of Proposition 1.* **Step 1: One checkpoint, two continuation protocols.** Two agents take $a_i \in \{-1, +1\}$. All transitions are terminal, so TD targets are rewards. Let $s_\star, s_T, s_0$ be equiprobable, with $r(s_\star, \mathbf{a}) = a_1 a_2$, $r(s_T, \mathbf{a}) = a_1$, and $r(s_0, \mathbf{a}) = 0$. Set $(f_1, f_2, b)$ to $(1, 1, 1)$ at $s_\star$, $(0, 1, 0)$ at $s_T$, and $(0, 0, 0)$ at $s_0$, and use

$$\begin{cases} Q_1(s, a_1; x) = f_1(s) a_1 x, \\ Q_2(s, a_2; y) = f_2(s) a_2 y, \\ Q_{\mathrm{tot}}(s, \mathbf{a}; x, y, \psi) = Q_1(s, a_1; x) + Q_2(s, a_2; y) + b(s)\psi. \end{cases} \tag{B.3}$$

The mixer is monotone because $\partial Q_{\text{tot}}/\partial Q_i = 1$, and $b(s)\psi$ is a state-conditioned bias, as in QMIX hypernetworks. The active block is $(x, \psi)$ and the teammate block is $y$. Under the evaluation policy of equation 1, agent 1 is uniform at $s_T$, so $s_T$ and $s_0$ contribute zero expected return and $J(x, y, \psi) = \frac{1}{3}\tanh(x/\beta_{\text{eval}})\tanh(y/\beta_{\text{eval}})$. The checkpoint is $(x, y, \psi) = (0, \vartheta, 0)$ with $0 < \vartheta < \min\{\beta_{\text{eval}}, 1/2\}$. The reference is $e^0 = (s_0, (+1, +1), 0)$; the candidates are $e_\pm = (s_\star, (\pm 1, \pm 1), +1)$; and the buffer also stores $e_T^+ = (s_T, (+1, +1), +1)$ and $e_T^- = (s_T, (-1, +1), -1)$, all genuine transitions of the same environment. For equiprobable $\sigma \in \{-1, +1\}$, protocol $\mathcal{P}^\sigma$ prescribes the continuation minibatch $\{e_T^\sigma\}$, masks active updates, and uses step size $\eta_1 = 2\vartheta$. Checkpoint, candidates, and reference are common to both protocols, so all Q-values, policies, and TD errors coincide.

**Step 2: Frozen and co-adaptive utilities.** All features vanish at $s_0$, so the reference gradient is zero. For the half-squared TD loss, each candidate's active gradient is $(Q_{\text{tot}} - 1)(a_1, 1)$, and an initial step of size $\eta_0$ gives

$$\begin{cases} (x_+, \psi_+) = (\eta_0(1-\vartheta), \eta_0(1-\vartheta)) = (d_+, d_+), \\ (x_-, \psi_-) = (-\eta_0(1+\vartheta), \eta_0(1+\vartheta)) = (-d_-, d_-), \end{cases} \tag{B.4}$$

with $d_\pm > 0$. The teammate gradient on $e_T^\sigma$ is $y - \sigma$, independent of $(x, \psi)$ because $f_1(s_T) = b(s_T) = 0$. Hence, in the reference and in every treatment, $y_\sigma^+ = \vartheta - \eta_1(\vartheta - \sigma)$, i.e., $y_+^+ = \vartheta(3 - 2\vartheta) > 0$ and $y_-^+ = -\vartheta(1 + 2\vartheta) < 0$. Let $A_\pm = \tanh(d_\pm/\beta_{\text{eval}})$, $q_0 = \tanh(\vartheta/\beta_{\text{eval}})$, and $q_\sigma = \tanh(|y_\sigma^+|/\beta_{\text{eval}})$, all positive. The reference has $x = 0$ and zero return, so

$$\begin{cases} u^{\text{frz}}(e_+) = \frac{1}{3} q_0 A_+, \\ u^{\text{frz}}(e_-) = -\frac{1}{3} q_0 A_-, \\ u_\sigma^{\text{co}}(e_+) = \frac{1}{3}\sigma q_\sigma A_+, \\ u_\sigma^{\text{co}}(e_-) = -\frac{1}{3}\sigma q_\sigma A_-. \end{cases} \tag{B.5}$$

Thus $e_+$ uniquely maximizes frozen utility under both protocols, whereas the co-adaptive maximizer switches with $\sigma$. Every teammate update is independent of active-owned state, so $\Gamma_{\text{resp},\sigma} = 0$ (Appendix B.4); the reversal is entirely complementarity.

**Step 3: Positive expected regret.** A selector restricted to the common summaries chooses $e_+$ with the same probability $\alpha$ under both protocols. Its expected regret is

$$\begin{aligned} \mathbb{E}_{\sigma,\text{sel}}[\text{Regret}] &= \tfrac{1}{6}(1-\alpha) q_+ (A_+ + A_-) + \tfrac{1}{6}\alpha q_- (A_+ + A_-) \\ &\geq \tfrac{1}{6}\min\{q_+, q_-\}(A_+ + A_-) > 0, \end{aligned} \tag{B.6}$$

which includes the frozen oracle ($\alpha = 1$).

**Step 4: The structural score recovers both rankings.** At the post-reference anchor $\mathbf{z} = (0, \vartheta, 0)$, masked active continuation gives $\mathbf{v}_\mathcal{R} = 0$, while $\mathbf{v}_\mathcal{T} = \eta_1(\vartheta - \sigma)$ is independent of $(x, \psi)$; hence $\mathbf{h}_{\text{own}} = \mathbf{h}_{\text{resp}} = 0$. Since $\nabla_\psi J = 0$, with exact return derivatives and $t = \vartheta/\beta_{\text{eval}} \in (0, 1)$, equation 6 reduces to $\widetilde{u}_\sigma(e) = x_e\, c_\sigma$, where

$$c_\sigma = \partial_x J - \partial_{xy}^2 J\, \mathbf{v}_\mathcal{T} = \frac{1}{3\beta_{\text{eval}}}\left[\tanh t - \operatorname{sech}^2(t)\, \frac{\eta_1(\vartheta - \sigma)}{\beta_{\text{eval}}}\right]. \tag{B.7}$$

Because $x_{e_+} > 0 > x_{e_-}$, the score selects $e_+$ iff $c_\sigma > 0$. For $\sigma = +1$, $\vartheta - 1 < 0$ gives $c_+ > 0$. For $\sigma = -1$ and $\eta_1 = 2\vartheta$, $c_- < 0$ is equivalent to $\sinh(2t)/(2t) < 2(1 + \vartheta)$, which holds because $\sinh(2t)/(2t) < \sinh(2)/2 < 1.82$ on $(0, 1)$. Summary-restricted selectors cannot separate the two scenarios, whereas $\widetilde{u}_k$ reads $\mathcal{P}_k$ through $\mathbf{v}_\mathcal{T}$, the complementarity channel of Theorem 1, and recovers the co-adaptive maximizer under both protocols. □

### B.3 Two-Step Expansion of the Co-Adaptive Mechanisms

For Theorem 1, specialize to fixed-target, unclipped plain SGD without optimizer memory, with post-reference parameters $\mathbf{z} = (\mathbf{r}, \mathbf{t})$ and initial active displacement $\mathbf{d}_{e,\iota}^\mathcal{R}$ given by equation 4. The

continuation step is synchronous, $\mathbf{z}^+ = \mathbf{z} - \eta\mathbf{G}(\mathbf{z})$, and derivatives are evaluated at $\mathbf{z}$ unless specified. Write $\mathbf{A}^{\mathcal{TR}} = D_{\mathcal{R}}\mathbf{G}_{\mathcal{T}}$, $\mathbf{C}_{\mathcal{RT}} = \nabla^2_{\mathcal{RT}}J$, $\mathbf{g}_{\mathcal{T}} = \nabla_{\mathcal{T}}J$, $\bar{\mathbf{d}}_e = \mathbb{E}_\iota[\mathbf{d}^{\mathcal{R}}_{e,\iota}]$, and $D_e = \max_\iota \|\mathbf{d}^{\mathcal{R}}_{e,\iota}\|$.

We prove equation 5 assuming $J \in C^3$ and $\mathbf{G} \in C^2$ have bounded derivatives on a convex neighborhood containing the matched states and interpolation segments; the constant $C$ in $|r_c(e)| + |r_r(e)| \le C(\eta D_e^2 + \eta^2 D_e)$ depends only on these bounds and fixed neighborhood sizes.

*Proof of Theorem 1.* **Step 1: Branch displacements.** Fix a slot and abbreviate $\mathbf{d} = \mathbf{d}^{\mathcal{R}}_{e,\iota}$, $D = \|\mathbf{d}\|$, $\mathbf{r}_f = \mathbf{r} - \eta\mathbf{G}_{\mathcal{R}}(\mathbf{z})$, and $\mathbf{v} = -\eta\mathbf{G}_{\mathcal{T}}(\mathbf{z})$. The treatment-minus-reference displacements after continuation are

$$\begin{cases} \mathbf{a} = \mathbf{d} - \eta\big[\mathbf{G}_{\mathcal{R}}(\mathbf{z} + (\mathbf{d}, 0)) - \mathbf{G}_{\mathcal{R}}(\mathbf{z})\big] = (I - \eta D_{\mathcal{R}}\mathbf{G}_{\mathcal{R}})\mathbf{d} + \boldsymbol{\rho}_{\mathcal{R}}, \\ \mathbf{b} = -\eta\big[\mathbf{G}_{\mathcal{T}}(\mathbf{z} + (\mathbf{d}, 0)) - \mathbf{G}_{\mathcal{T}}(\mathbf{z})\big] = -\eta\mathbf{A}^{\mathcal{TR}}\mathbf{d} + \boldsymbol{\rho}_{\mathcal{T}}, \\ \boldsymbol{\rho}_{\mathcal{A}} = -\eta\displaystyle\int_0^1 (1-s)D^2_{\mathcal{RR}}\mathbf{G}_{\mathcal{A}}(\mathbf{z} + s(\mathbf{d}, 0))[\mathbf{d}, \mathbf{d}]\, ds, \end{cases} \tag{B.8}$$

for $\mathcal{A} \in \{\mathcal{R}, \mathcal{T}\}$, with $\|\boldsymbol{\rho}_{\mathcal{A}}\| \le C\eta D^2$. Local derivative bounds also give $\|\mathbf{a} - \mathbf{d}\| \le C\eta D$, $\|\mathbf{a}\| \le CD$, $\|\mathbf{v}\| \le C\eta$, and $\|\mathbf{b}\| \le C\eta D$, where $C$ may grow between inequalities.

**Step 2: Complementarity as a mixed difference.** The frozen and scripted treatments share the active coordinate $\mathbf{r}_f + \mathbf{a}$ but have teammate coordinates $\mathbf{t}$ and $\mathbf{t} + \mathbf{v}$. Their slot-specific utility difference is Specifically, $c_\iota(e) = J(\mathbf{r}_f + \mathbf{a}, \mathbf{t} + \mathbf{v}) - J(\mathbf{r}_f, \mathbf{t} + \mathbf{v}) - J(\mathbf{r}_f + \mathbf{a}, \mathbf{t}) + J(\mathbf{r}_f, \mathbf{t})$. By the mixed integral identity, $c_\iota(e) = \int_0^1\int_0^1 \mathbf{a}^\top\nabla^2_{\mathcal{RT}}J(\mathbf{r}_f + s\mathbf{a}, \mathbf{t} + q\mathbf{v})\mathbf{v}\, dq\, ds$. Therefore, $c_\iota(e) = -\eta\mathbf{d}^\top\mathbf{C}_{\mathcal{RT}}\mathbf{G}_{\mathcal{T}} + \rho_{c,\iota}$. The mixed Hessian is locally Lipschitz and is evaluated within $C\eta + \|\mathbf{a}\|$ of $\mathbf{z}$, so $|\rho_{c,\iota}| \le C\left[\|\mathbf{a} - \mathbf{d}\|\|\mathbf{v}\| + \|\mathbf{a}\|\|\mathbf{v}\|(\eta + \|\mathbf{a}\|)\right] \le C(\eta^2 D + \eta D^2)$. The double integral retains the factor $\|\mathbf{v}\| = O(\eta)$ that four independent Taylor remainders would lose.

**Step 3: Induced teammate response.** The co-adaptive and scripted references coincide, while their treatment teammate coordinates differ by $\mathbf{b}$. Thus

$$\begin{aligned} r_\iota(e) &= \int_0^1 \nabla_{\mathcal{T}}J(\mathbf{r}_f + \mathbf{a}, \mathbf{t} + \mathbf{v} + s\mathbf{b})^\top\mathbf{b}\, ds \\ &= -\eta\mathbf{d}^\top(\mathbf{A}^{\mathcal{TR}})^\top\mathbf{g}_{\mathcal{T}} + \mathbf{g}_{\mathcal{T}}^\top\boldsymbol{\rho}_{\mathcal{T}} \\ &\quad + \int_0^1 \left[\nabla_{\mathcal{T}}J(\mathbf{r}_f + \mathbf{a}, \mathbf{t} + \mathbf{v} + s\mathbf{b}) - \mathbf{g}_{\mathcal{T}}\right]^\top\mathbf{b}\, ds. \end{aligned} \tag{B.9}$$

Calling the final two terms $\rho_{r,\iota}$, bounded gradients and Hessians imply $|\rho_{r,\iota}| \le C[\eta D^2 + (\eta + \|\mathbf{a}\| + \|\mathbf{b}\|)\|\mathbf{b}\|] \le C(\eta D^2 + \eta^2 D)$.

**Step 4: Slot averaging.** Reference derivatives are common to all slots, so averaging the linear terms replaces $\mathbf{d}$ by $\bar{\mathbf{d}}_e$, and

$$\begin{aligned} |r_c(e)| + |r_r(e)| &\le \mathbb{E}_\iota[|\rho_{c,\iota}| + |\rho_{r,\iota}| \mid \mathcal{F}_k] \\ &\le C\left(\eta\mathbb{E}_\iota\|\mathbf{d}^{\mathcal{R}}_{e,\iota}\|^2 + \eta^2\mathbb{E}_\iota\|\mathbf{d}^{\mathcal{R}}_{e,\iota}\|\right) \le C(\eta D_e^2 + \eta^2 D_e). \end{aligned} \tag{B.10}$$

For $D_e = O(\eta)$, the leading terms are $O(\eta^2)$ and the remainders $O(\eta^3)$, as stated in Theorem 1. □

### B.4 Multi-Step Propagation and Shared Influence Estimation

Let $\mathbf{z}_h^{0,p}$ and $\mathbf{z}_h^{e,p}$ be reference and treatment states in continuous coordinates, with schedules and random keys fixed. Define $\mathcal{J}(\mathbf{z}) = J(\mathrm{pr}_\Theta\,\mathbf{z})$ and $\boldsymbol{\delta}_h = \mathbf{z}_h^{e,p} - \mathbf{z}_h^{0,p}$. The initial displacement $\boldsymbol{\delta}_1 = (\boldsymbol{\delta}_1^{\mathcal{R}}, \mathbf{0})$ includes changed active optimizer or target coordinates and is parameter-only under SGD. Let $\mathbf{K}_h^p = DF_h^p(\mathbf{z}_h^{0,p})$ and $\mathbf{K}_{b:a}^p = \mathbf{K}_b^p \cdots \mathbf{K}_a^p$ (empty products are the identity); scripted maps hold the reference teammate path fixed when differentiated.

**Lemma B.1.** *Suppose $F_h^p$ and $\mathcal{J}$ are $C^2$ on convex neighborhoods containing the branch paths and interpolation segments, with $\|DF_h^p\| \le a$, $\|D^2F_h^p\| \le b$, $\|\nabla\mathcal{J}\| \le g$, and $\|\nabla^2\mathcal{J}\| \le L_J$, where*

$a \geq 1$. *For each fixed slot,*

$$\begin{cases} \mathcal{J}(\mathbf{z}_H^{e,p}) - \mathcal{J}(\mathbf{z}_H^{0,p}) = \nabla \mathcal{J}(\mathbf{z}_H^{0,p})^\top \mathbf{K}_{H-1:1}^p \boldsymbol{\delta}_1 + \mathcal{R}_{e,\iota}^p, \\ |\mathcal{R}_{e,\iota}^p| \leq C_H \|\boldsymbol{\delta}_1\|^2, \\ C_H = \dfrac{gb}{2} \displaystyle\sum_{j=0}^{H-2} a^{H-2-j} a^{2j} + \dfrac{L_J}{2} a^{2(H-1)}, \end{cases} \tag{B.11}$$

*where the sum is empty for $H = 1$.*

*Proof.* Mean-value formula gives $\|\boldsymbol{\delta}_h\| \leq a^{h-1}\|\boldsymbol{\delta}_1\|$, and a Taylor expansion of one update yields $\boldsymbol{\delta}_{h+1} = \mathbf{K}_h^p \boldsymbol{\delta}_h + \boldsymbol{\rho}_h$, with $\boldsymbol{\rho}_h = \int_0^1 (1-s) D^2 F_h^p(\mathbf{z}_h^{0,p} + s\boldsymbol{\delta}_h)[\boldsymbol{\delta}_h, \boldsymbol{\delta}_h] ds$ and $\|\boldsymbol{\rho}_h\| \leq (b/2) a^{2(h-1)} \|\boldsymbol{\delta}_1\|^2$. Set $\mathbf{v}_1^p = \boldsymbol{\delta}_1$, $\mathbf{v}_{h+1}^p = \mathbf{K}_h^p \mathbf{v}_h^p$, and $\mathbf{e}_h = \boldsymbol{\delta}_h - \mathbf{v}_h^p$. Since $\mathbf{e}_1 = 0$, $\mathbf{e}_H = \sum_{\ell=1}^{H-1} \mathbf{K}_{H-1:\ell+1}^p \boldsymbol{\rho}_\ell$, $\|\mathbf{e}_H\| \leq \frac{b}{2}\|\boldsymbol{\delta}_1\|^2 \sum_{\ell=1}^{H-1} a^{H-1-\ell} a^{2(\ell-1)} = \frac{b}{2}\|\boldsymbol{\delta}_1\|^2 \sum_{j=0}^{H-2} a^{H-2-j} a^{2j}$. Expanding $\mathcal{J}$ at $\mathbf{z}_H^{0,p}$ bounds the residual by $g\|\mathbf{e}_H\| + (L_J/2)\|\boldsymbol{\delta}_H\|^2$, proving equation B.11; averaging slots gives the utility bound. The constant depends on $H$, so the statement is finite-horizon. □

**Response propagation and feedback.** Partition $\mathbf{K}_h^{\mathrm{co}}$ into blocks $\left(\begin{smallmatrix} \mathbf{P}_h & \mathbf{Q}_h \\ \mathbf{R}_h & \mathbf{S}_h \end{smallmatrix}\right)$ and write $\mathbf{v}_h^{\mathrm{co}} = (\mathbf{a}_h^{\mathrm{co}}, \mathbf{b}_h^{\mathrm{co}})$, so that $\mathbf{a}_{h+1}^{\mathrm{co}} = \mathbf{P}_h \mathbf{a}_h^{\mathrm{co}} + \mathbf{Q}_h \mathbf{b}_h^{\mathrm{co}}$ and $\mathbf{b}_{h+1}^{\mathrm{co}} = \mathbf{R}_h \mathbf{a}_h^{\mathrm{co}} + \mathbf{S}_h \mathbf{b}_h^{\mathrm{co}}$. The scripted tangent obeys $\mathbf{b}_h^{\mathrm{scr}} = 0$ and $\mathbf{a}_{h+1}^{\mathrm{scr}} = \mathbf{P}_h \mathbf{a}_h^{\mathrm{scr}}$. Repeated substitution gives

$$\begin{cases} \mathbf{b}_H^{\mathrm{co}} = \displaystyle\sum_{\ell=1}^{H-1} \mathbf{S}_{H-1:\ell+1} \mathbf{R}_\ell \mathbf{a}_\ell^{\mathrm{co}}, \\ \mathbf{a}_H^{\mathrm{co}} - \mathbf{a}_H^{\mathrm{scr}} = \displaystyle\sum_{h=2}^{H-1} \sum_{\ell=1}^{h-1} \mathbf{P}_{H-1:h+1} \mathbf{Q}_h \mathbf{S}_{h-1:\ell+1} \mathbf{R}_\ell \mathbf{a}_\ell^{\mathrm{co}}, \end{cases} \tag{B.12}$$

where the second line retains ordered active-to-teammate-to-active feedback. With $\mathbf{g}_{\mathcal{A}}^p = \nabla_{\mathcal{A}} \mathcal{J}(\mathbf{z}_H^{0,p})$, slot averaging yields

$$\begin{cases} \Gamma_{\mathrm{comp},k}(e) = \mathbb{E}_\iota\left[(\mathbf{g}_{\mathcal{R}}^{\mathrm{co}})^\top \mathbf{a}_H^{\mathrm{scr}} - (\mathbf{g}_{\mathcal{R}}^{\mathrm{frz}})^\top \mathbf{a}_H^{\mathrm{frz}} \mid \mathcal{F}_k\right] + R_{c,H}(e), \\ \Gamma_{\mathrm{resp},k}(e) = \mathbb{E}_\iota\left[(\mathbf{g}_{\mathcal{R}}^{\mathrm{co}})^\top (\mathbf{a}_H^{\mathrm{co}} - \mathbf{a}_H^{\mathrm{scr}}) + (\mathbf{g}_{\mathcal{T}}^{\mathrm{co}})^\top \mathbf{b}_H^{\mathrm{co}} \mid \mathcal{F}_k\right] + R_{r,H}(e), \end{cases} \tag{B.13}$$

with each remainder bounded by $2C_H \mathbb{E}_\iota \|\boldsymbol{\delta}_1\|^2$. The frozen tangent uses its own reference path, not necessarily the same $\mathbf{P}_h$.

**Shared-anchor approximation.** For fixed $H$ under fixed-target SGD with nonnegative step sizes, let $\zeta_H = \sum_{h=1}^{H-1} \eta_h$ and $\mathbf{v}_{\mathcal{A}} = \sum_{h=1}^{H-1} \eta_h \mathbf{G}_{\mathcal{A},h}(\mathbf{z})$, where scheduled masks are included. These accumulated directions differ from the state tangents $\mathbf{v}_h^p$. Let $\mathbf{g}_{\mathcal{A}} = \nabla_{\mathcal{A}} J(\mathbf{z})$ and $\mathbf{C}_{\mathcal{A}\mathcal{B}} = \nabla^2_{\mathcal{A}\mathcal{B}} J(\mathbf{z})$.

*Derivation of the structural score equation 6.* **Step 1: Expand the continuation maps.** Under the two-step regularity conditions, reference states drift by $O(\zeta_H)$ uniformly over scheduled fields, and for small $\zeta_H$

$$\begin{aligned} \overleftarrow{\prod}_{h=1}^{H-1} [I - \eta_h D\mathbf{G}_h(\mathbf{z}_h)] &= I - \sum_{h=1}^{H-1} \eta_h D\mathbf{G}_h(\mathbf{z}) + O(\zeta_H^2), \\ \Psi_k^{\mathrm{co}}(\mathbf{z}) = \Psi_k^{\mathrm{scr}}(\mathbf{z}) &= \mathbf{z} - (\mathbf{v}_{\mathcal{R}}, \mathbf{v}_{\mathcal{T}}) + O(\zeta_H^2), \\ \Psi_k^{\mathrm{frz}}(\mathbf{z}) &= \mathbf{z} - (\mathbf{v}_{\mathcal{R}}, 0) + O(\zeta_H^2), \end{aligned} \tag{B.14}$$

where later factors act on the left and the error collects two-step-size terms and derivative drift. To first order, the co-adaptive $\mathcal{R}$-Jacobian has blocks $I - D_{\mathcal{R}} \mathbf{v}_{\mathcal{R}}$ and $-D_{\mathcal{R}} \mathbf{v}_{\mathcal{T}}$, while the scripted and frozen teammate-output derivatives are zero: equal reference endpoints do not imply equal derivatives, because the scripted path is fixed.

**Step 2: Separate the shared coefficients.** For $\mathbf{a}_p = [D_{\mathcal{R}}\Psi_k^p(\mathbf{z})]^\top \nabla J(\Psi_k^p(\mathbf{z}))$, expanding the terminal gradient and keeping terms linear in $\zeta_H$ gives

$$\begin{cases} \mathbf{a}_{\mathrm{frz}} = \mathbf{g}_{\mathcal{R}} - (D_{\mathcal{R}}\mathbf{v}_{\mathcal{R}})^\top \mathbf{g}_{\mathcal{R}} - \mathbf{C}_{\mathcal{R}\mathcal{R}}\mathbf{v}_{\mathcal{R}} + O(\zeta_H^2), \\ \mathbf{a}_{\mathrm{scr}} - \mathbf{a}_{\mathrm{frz}} = -\mathbf{C}_{\mathcal{R}\mathcal{T}}\mathbf{v}_{\mathcal{T}} + O(\zeta_H^2), \\ \mathbf{a}_{\mathrm{co}} - \mathbf{a}_{\mathrm{scr}} = -(D_{\mathcal{R}}\mathbf{v}_{\mathcal{T}})^\top \mathbf{g}_{\mathcal{T}} + O(\zeta_H^2), \end{cases} \tag{B.15}$$

whose leading terms are the population counterparts of $\mathbf{b}_{\mathrm{frz}}$, $\mathbf{b}_{\mathrm{comp}}$, and $\mathbf{b}_{\mathrm{resp}}$ in equation 6.

**Step 3: Recover the candidate score.** Linearizing $J \circ \Psi_k^p$ in the initial displacement and averaging slots gives $u_k^p(e) = \bar{\mathbf{d}}_e^\top \mathbf{a}_p + O(\mathbb{E}_\iota \|\mathbf{d}_{e,\iota}^{\mathcal{R}}\|^2)$, hence $u_k^{\mathrm{co}}(e) = \bar{\mathbf{d}}_e^\top \left[\mathbf{g}_{\mathcal{R}} - (D_{\mathcal{R}}\mathbf{v}_{\mathcal{R}})^\top \mathbf{g}_{\mathcal{R}} - \mathbf{C}_{\mathcal{R}\mathcal{R}}\mathbf{v}_{\mathcal{R}} - \mathbf{C}_{\mathcal{R}\mathcal{T}}\mathbf{v}_{\mathcal{T}} - (D_{\mathcal{R}}\mathbf{v}_{\mathcal{T}})^\top \mathbf{g}_{\mathcal{T}}\right] + O\big(\zeta_H^2 D_e + \mathbb{E}_\iota \|\mathbf{d}_{e,\iota}^{\mathcal{R}}\|^2\big)$. Replacing return derivatives by estimates gives equation 6. Optimizer memory, clipping, target-copy discrepancies, derivative sampling, and cache staleness enter the total score error of Proposition 2, not this SGD Taylor remainder. □

**Return derivatives and fixed-vector products.** For a trajectory $\xi$, let $R(\xi)$ be discounted return, $\mathbf{s}^{\mathcal{A}} = \nabla_{\mathcal{A}} \log p_\Theta(\xi)$, and $\mathbf{H}_{\mathcal{A}\mathcal{B}} = \nabla^2_{\mathcal{A}\mathcal{B}} \log p_\Theta(\xi)$. If differentiation and expectation can be interchanged, $\mathbf{g}_{\mathcal{A}} = \mathbb{E}[R\mathbf{s}^{\mathcal{A}}]$ and $\mathbf{C}_{\mathcal{A}\mathcal{B}} = \mathbb{E}[R(\mathbf{s}^{\mathcal{A}}(\mathbf{s}^{\mathcal{B}})^\top + \mathbf{H}_{\mathcal{A}\mathcal{B}})]$. Non-shared policies and parameter-independent environment laws imply $\mathbf{H}_{\mathcal{R}\mathcal{T}} = 0$ on fixed trajectories; hence, with evaluation directions held fixed,

$$\begin{cases} \mathbf{C}_{\mathcal{R}\mathcal{T}}\mathbf{v}_{\mathcal{T}} = \mathbb{E}\big[R\mathbf{s}^{\mathcal{R}}\langle \mathbf{s}^{\mathcal{T}}, \mathbf{v}_{\mathcal{T}}\rangle\big], \\ \mathbf{C}_{\mathcal{R}\mathcal{R}}\mathbf{v}_{\mathcal{R}} = \mathbb{E}\big[R\left(\mathbf{s}^{\mathcal{R}}\langle \mathbf{s}^{\mathcal{R}}, \mathbf{v}_{\mathcal{R}}\rangle + \mathbf{H}_{\mathcal{R}\mathcal{R}}\mathbf{v}_{\mathcal{R}}\right)\big]. \end{cases} \tag{B.16}$$

The within-block log-likelihood Hessian must be retained, as in the DiCE construction. Update sensitivity instead fixes the return gradient: for $\mathcal{A} \in \{\mathcal{R}, \mathcal{T}\}$, $\nabla_{\mathcal{R}}\big[\mathbf{v}_{\mathcal{A}}^\top \mathrm{sg}(\widehat{\mathbf{g}}_{\mathcal{A}})\big] = (D_{\mathcal{R}}\mathbf{v}_{\mathcal{A}})^\top \widehat{\mathbf{g}}_{\mathcal{A}}$; otherwise the product rule adds evaluation curvature a second time. Although $\nabla_\psi J = 0$, mixer perturbations still act through update Jacobians.

**Vanishing cases.** For $H = 1$, the branches coincide and both corrections vanish. If every complete teammate update is independent of active-owned state, identical initial teammate states remain identical by induction, so $u_k^{\mathrm{co}} = u_k^{\mathrm{scr}}$ and $\Gamma_{\mathrm{resp},k} = 0$ without forcing complementarity to vanish (Appendix B.2 is such a case). Conversely, $\mathbf{G}_{\mathcal{T}}(\mathbf{z}) = 0$ does not imply $D_{\mathcal{R}}\mathbf{G}_{\mathcal{T}}(\mathbf{z}) = 0$, so zero reference drift alone does not eliminate induced response.

### B.5 Finite-Budget Replay Allocation Guarantee

Condition on the checkpoint and realized scores, with $\mathcal{S}_k = \{S \subseteq \mathcal{E}_k : |S| = B\}$ and $\widehat{V}_k(S) = \sum_{e\in S} s_k(e)$. Let $I_k(S) = U_k^{\mathrm{co}}(S) - \sum_{e\in S} u_k^{\mathrm{co}}(e)$. The score maximizer $S_k^{s,*}$ is distinct from the randomized selection $\widehat{S}_k$, and their score shortfall is $\Lambda_k = \widehat{V}_k(S_k^{s,*}) - \mathbb{E}[\widehat{V}_k(\widehat{S}_k)] \geq 0$.

In equation 9, $\kappa_k$ bounds the entire batch residual $\max_S |I_k(S)|$, not one pairwise interaction, and the expectation is over replay sampling only.

*Proof of Proposition 2.* Fixed cardinality and the triangle inequality give, for every feasible batch, including negative utilities,

$$\begin{aligned} |U_k^{\mathrm{co}}(S) - \widehat{V}_k(S)| &= \left| I_k(S) + \sum_{e\in S} [u_k^{\mathrm{co}}(e) - s_k(e)] \right| \\ &\leq |I_k(S)| + \sum_{e\in S} |u_k^{\mathrm{co}}(e) - s_k(e)| \leq \kappa_k + B\varepsilon_k. \end{aligned} \tag{B.17}$$

Score optimality and the definition of $\Lambda_k$ then give

$$\begin{aligned} U_k^{\mathrm{co}}(S_k^{\mathrm{co},*}) &\leq \widehat{V}_k(S_k^{\mathrm{co},*}) + B\varepsilon_k + \kappa_k \\ &\leq \widehat{V}_k(S_k^{s,*}) + B\varepsilon_k + \kappa_k \\ &= \mathbb{E}[\widehat{V}_k(\widehat{S}_k)] + \Lambda_k + B\varepsilon_k + \kappa_k. \end{aligned} \tag{B.18}$$

Applying equation B.17 to each realization of $\widehat{S}_k$ before taking expectation, $\mathbb{E}[\widehat{V}_k(\widehat{S}_k)] \leq \mathbb{E}[U_k^{\mathrm{co}}(\widehat{S}_k)] + B\varepsilon_k + \kappa_k$, and substitution proves equation 9 without identifying the stochastic allocator with deterministic top-$B$ selection. □

**Advantage over the frozen oracle.** Let $\mathcal{M}_k = \Delta U_{\mathrm{sel},k} - (2B\varepsilon_k + 2\kappa_k + \Lambda_k)$. Subtracting the allocation loss from the CoRVM gap gives $\mathbb{E}[U_k^{\mathrm{co}}(\widehat{S}_k)] - U_k^{\mathrm{co}}(S_k^{\mathrm{frz},*}) = \Delta U_{\mathrm{sel},k} - \left[U_k^{\mathrm{co}}(S_k^{\mathrm{co},*}) - \mathbb{E}[U_k^{\mathrm{co}}(\widehat{S}_k)]\right] \geq \mathcal{M}_k$, so $\mathcal{M}_k > 0$ ensures a positive expected local advantage. Exact additivity with deterministic score maximization recovers the $2B\varepsilon_k$ bound. With scores fixed, sorting and independent sampler runs estimate $\Lambda_k$ without learner rollouts. Paired-audit validity and finite audit accuracy do not, by themselves, establish the two uniform error bounds.

# C Algorithmic Details and Implementation

Scoring uses the fixed-target SGD surrogate of Appendix B.4, while main updates and audits use the configured learner. Decisions use only pre-allocation information, and completed audits supervise later decisions. Rollout collection and buffer insertion are unchanged and counted toward interaction cost.

## C.1 Replay Scoring and Budgeted Updates

Candidates are joint sequences scored on learning positions $\mathcal{I}(e)$, excluding burn-in and padding, with the same loss normalization and focal mask as Section 2.1. A reference-batch gradient gives the slot-averaged displacement

$$\begin{cases} \bar{\mathbf{g}}_k^0 = \dfrac{1}{B}\displaystyle\sum_{\iota=1}^{B} \nabla_{\mathcal{R}} \ell(e_\iota^0; X_k), \\ \bar{\mathbf{d}}_e = -\dfrac{\eta_0}{B}\left[\nabla_{\mathcal{R}} \ell(e; X_k) - \bar{\mathbf{g}}_k^0\right]. \end{cases} \tag{C.1}$$

The scoring anchor is $\mathbf{z} = \boldsymbol{\Theta}_k - \eta_0(\bar{\mathbf{g}}_k^0, \mathbf{0})$, whereas audits copy the complete configured learner. At refresh, define $\mathbf{v}_{\mathcal{A}} = \sum_{h=1}^{H-1} \eta_h \mathbf{G}_{\mathcal{A},h}(\mathbf{z})$ from the scheduled, masked, fixed-target fields of $\mathcal{P}_k$ for $\mathcal{A} \in \{\mathcal{R}, \mathcal{T}\}$. Using $n_v$ evaluation trajectories with return $R_\ell$, score $\mathbf{s}_\ell^{\mathcal{A}}$, and log-likelihood Hessian $\mathbf{H}_\ell^{\mathcal{R}\mathcal{R}}$, compute

$$\begin{cases} \widehat{\mathbf{g}}_{\mathcal{A}} = \dfrac{1}{n_v}\displaystyle\sum_{\ell=1}^{n_v} R_\ell \mathbf{s}_\ell^{\mathcal{A}}, \\ \widehat{\mathbf{c}}_{\mathcal{R}\mathcal{T}} = \dfrac{1}{n_v}\displaystyle\sum_{\ell=1}^{n_v} R_\ell \mathbf{s}_\ell^{\mathcal{R}} \langle \mathbf{s}_\ell^{\mathcal{T}}, \mathbf{v}_{\mathcal{T}} \rangle, \\ \widehat{\mathbf{c}}_{\mathcal{R}\mathcal{R}} = \dfrac{1}{n_v}\displaystyle\sum_{\ell=1}^{n_v} R_\ell \left[\mathbf{s}_\ell^{\mathcal{R}} \langle \mathbf{s}_\ell^{\mathcal{R}}, \mathbf{v}_{\mathcal{R}} \rangle + \mathbf{H}_\ell^{\mathcal{R}\mathcal{R}} \mathbf{v}_{\mathcal{R}}\right]. \end{cases} \tag{C.2}$$

Together with $\mathbf{h}_{\mathrm{own}}$ and $\mathbf{h}_{\mathrm{resp}}$ from equation 7, these yield

$$\begin{cases} \mathbf{b}_{\mathrm{frz}} = \widehat{\mathbf{g}}_{\mathcal{R}} - \mathbf{h}_{\mathrm{own}} - \widehat{\mathbf{c}}_{\mathcal{R}\mathcal{R}}, \\ \mathbf{b}_{\mathrm{comp}} = -\widehat{\mathbf{c}}_{\mathcal{R}\mathcal{T}}, \\ \mathbf{b}_{\mathrm{resp}} = -\mathbf{h}_{\mathrm{resp}}, \\ \widetilde{u}_k(e) = \bar{\mathbf{d}}_e^{\top}\left(\mathbf{b}_{\mathrm{frz}} + \mathbf{b}_{\mathrm{comp}} + \mathbf{b}_{\mathrm{resp}}\right), \\ s_k(e) = \widetilde{u}_k(e) + \boldsymbol{\beta}_k^{\top} \boldsymbol{\varphi}_k(e). \end{cases} \tag{C.3}$$

DiCE computes the within-block curvature product without a dense Hessian. Between refreshes, focal-block coefficients are reused with current candidate and reference gradients. The $B$ samples are drawn without replacement by equation 8 and applied in the fixed intervention order $\mathrm{ord}(\widehat{S}_k)$.

**Algorithm C.1:** ExSpectra: shared replay valuation, budgeted update, and sparse CAUD audit

**Define:** $M > B$: pool and batch sizes; $H$: horizon; $K_v, K_p$: refresh and audit periods; $m$: probe draws; $R_p$: paired replicates; $\mathcal{C}$: focal caches; $\mathcal{W}$: completed probe records.

**Function** `ReplayStep(`$X_k$`)`**:**

  $f_k \leftarrow$ round-robin focal identity; $X_k^{\mathrm{save}} \leftarrow$ complete copy of $X_k$; draw $\mathcal{E}_k, \mathcal{B}_k^0 \sim q_{\mathrm{ref}}$, and fix $\mathcal{P}_k$

  ▷ `Stage I: shared estimation and calibrated scores`

  $\boldsymbol{\beta}_k \leftarrow$ ridge fit equation C.5 on records in $\mathcal{W}$ with $k_a < k$; compute $\bar{\mathbf{g}}_k^0$ and the SGD anchor $\mathbf{z}$

  **if** $\mathcal{C}[f_k]$ *is missing or* $k - k_v(f_k) \geq K_v$ **then**

    Form $\mathbf{v}_{\mathcal{R}}, \mathbf{v}_{\mathcal{T}}$ from $\mathcal{P}_k$; compute equation C.2 and equation 7 with $n_v$ trajectories

    Cache $\mathbf{b}_{\mathrm{frz}}, \mathbf{b}_{\mathrm{comp}}, \mathbf{b}_{\mathrm{resp}}$ with anchor, protocol, and target versions; $k_v(f_k) \leftarrow k$

  **foreach** $e \in \mathcal{E}_k$ **do** compute $\bar{\mathbf{d}}_e, \boldsymbol{\varphi}_k(e), \widetilde{u}_k(e), s_k(e)$ by equation C.1 and equation C.3

  ▷ `Stage II: sequential selection and main update`

  $\widehat{S}_k \leftarrow B$ draws without replacement by equation 8; $X_{k+1} \leftarrow$ configured update on $\mathrm{ord}(\widehat{S}_k)$ with focal block $f_k$

  ▷ `Stage III: shadow audit and delayed supervision`

  **if** $k \bmod K_p = 0$ *and both budgets permit a complete audit* **then**

    Fix $\mu_k, \nu_k$ from saved pre-allocation data; $\mathcal{Y}_k \leftarrow$ `MatchedAudit(`$X_k^{\mathrm{save}}, \mathcal{E}_k, \mathcal{B}_k^0, \mathcal{P}_k, \mu_k$`)`; store with frozen $\widetilde{u}_k, \boldsymbol{\varphi}_k, \mu_k, \nu_k$

  Charge realized costs to the simulator and gradient ledgers; **return** $X_{k+1}$

**Function** `MatchedAudit(`$X_k^{\mathrm{save}}, \mathcal{E}_k, \mathcal{B}_k^0, \mathcal{P}_k, \mu_k$`)`**:**

  Draw $e_1, \ldots, e_m \sim \mu_k$ independently; record per-draw probabilities

  **for** $\rho = 1$ **to** $R_p$ **do**

    Draw evaluation keys $\zeta_\rho$ disjoint from derivative trajectories; run reference–co (storing $\mathcal{T}_{1:H}^{0,\mathrm{co}}$) and reference–frz; evaluate $\widehat{J}_{0,\rho}^p$

    **for** $j = 1$ **to** $m$ **do**

      $\iota \sim \mathrm{Unif}\{1, \ldots, B\}$; for $p \in \{\mathrm{co}, \mathrm{frz}, \mathrm{scr}\}$ run `RunBranch(`$X_k^{\mathrm{save}}, \mathcal{B}_{k,\iota \to e_j}^0, \mathcal{P}_k, p; \mathcal{T}_{1:H}^{0,\mathrm{co}}$`)` and evaluate $\widehat{J}_{j,\rho}^p$ with $\zeta_\rho$

      Form equation C.4; store slot and identifier $(k, \rho)$

  Average replicates into $Y_j^p$; store $Y_j^{\mathrm{co}}$, $Y_j^{\mathrm{scr}} - Y_j^{\mathrm{frz}}$, $Y_j^{\mathrm{co}} - Y_j^{\mathrm{scr}}$; discard shadow learners; **return** labels, metadata, and costs

## C.2 Matched Auditing and Residual Calibration

Probe candidates are sampled with replacement from $\mu_k(e) = \epsilon_p / M_k + (1 - \epsilon_p)\mu_k^{\mathrm{audit}}(e)$, where $\mu_k^{\mathrm{audit}}$ and target law $\nu_k$ are fixed before audit outcomes; duplicates remain distinct samples. Each branch clones $X_k^{\mathrm{save}}$, applies the active-only initial update, and executes the common $\mathcal{P}_k$ with the configured learner. Replicates keep the source checkpoint, reference batch, and protocol fixed while resampling slots and evaluation keys. The matched labels are

$$\begin{cases} Y_{j,\rho}^{\mathrm{co}} = \widehat{J}_{j,\rho}^{\mathrm{co}} - \widehat{J}_{0,\rho}^{\mathrm{co}}, \\ Y_{j,\rho}^{\mathrm{frz}} = \widehat{J}_{j,\rho}^{\mathrm{frz}} - \widehat{J}_{0,\rho}^{\mathrm{frz}}, \\ Y_{j,\rho}^{\mathrm{scr}} = \widehat{J}_{j,\rho}^{\mathrm{scr}} - \widehat{J}_{0,\rho}^{\mathrm{co}}, \end{cases} \tag{C.4}$$

and are averaged over $R_p$ replicates. Only completed records with source checkpoint $k_a < k$ enter $\mathcal{W}_k$. For $\lambda_{\mathrm{ridge}} > 0$, calibration solves

$$\boldsymbol{\beta}_k = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{a \in \mathcal{W}_k} \frac{\nu_a(e_a)}{\mu_a(e_a)} \left[ Y_a^{\mathrm{co}} - \widetilde{u}_a(e_a) - \boldsymbol{\beta}^\top \boldsymbol{\varphi}_a(e_a) \right]^2 + \lambda_{\mathrm{ridge}} \|\boldsymbol{\beta}\|_2^2 \right\}. \tag{C.5}$$

Empty windows use $\boldsymbol{\beta}_k = \mathbf{0}$, and unchanged windows reuse the previous fit. Current audits only append records for later decisions. The feature vector $\boldsymbol{\varphi}_k(e) \in \mathbb{R}^{d_\varphi}$ uses only pre-allocation candidate information and is fixed before held-out auditing.

### C.3 Caching and Resource Accounting

Caches are focal-specific, refreshed every $K_v$ updates, and retain the anchor, protocol, target versions, and derivative-trajectory identifiers; stale-cache effects enter the total score error. Scoring, calibration, refresh, main-update, and audit costs are charged to simulator and gradient ledgers, and an audit runs only when both budgets cover the complete paired computation. Algorithm C.1 schedules $R_p(2+3m)H$ branch-update positions per audit. Shared derivative-product counts are independent of $M_k$, whereas candidate-gradient cost scales with $M_k$. Materializing candidate gradients costs $O(M_k d_{\mathcal{R}})$ memory, which streaming avoids, while calibration stores $O(|\mathcal{W}_k| d_\varphi)$ values.

## D Experimental Methodology

Controlled audits test whether co-adaptation changes finite-budget replay decisions (RQ1), whether CAUD explains the change (RQ2), and how closely scored selections match held-out co-adaptive utility (RQ3). End-to-end runs test whether these selection gains accompany improved learning under matched resources (RQ4).

### D.1 Benchmarks and Unified Configuration

All comparisons use non-shared QMIX with a common architecture, optimizer, observation interface, and replay representation. Method-specific replay, reset, augmentation, or auxiliary computation remains part of the corresponding method. Terran$^\circ$ is used only for development and is marked wherever reported. Pressure-Plate is a diagnostic two-agent $7 \times 7$ task with horizon 40; the doorway depends on the pressure plate with probability $p \in \{0, 0.5, 1\}$, with $p = 1$ as the primary coupled setting. All benchmark learning results use eight matched training seeds and greedy evaluation over 32 episodes per checkpoint. Local replay utilities instead use the smooth evaluation objective defined in Section 2.1.

### D.2 Baselines and Structural Controls

**QMIX / Focal-U.** QMIX uses native uniform replay; Focal-U uses ExSpectra's round-robin focal schedule and candidate pool but samples uniformly, isolating the effect of replay valuation.

**Replay-priority baselines.** TG-Replay (Korniak et al., 2026), PER (Schaul et al., 2015), ReaPER (Pleiss et al., 2026), ReLo (Sujit et al., 2023), ReMERT (Liu et al., 2021), DIFFER (Hu et al., 2023), and MAC-PO (Mei et al., 2023) use their native replay signals. Transition-level scores are aggregated over valid learning positions of the same joint sequence without adding unavailable candidate information. PER uses mean absolute TD error with priority exponent $0.6$ and an importance-weight exponent annealed from $0.4$ to $1.0$.

**Influence controls.** LocalInf+Cal uses the immediate local influence $\bar{\mathbf{d}}_e^\top \widehat{\mathbf{g}}_{\mathcal{R}}$. JointInf+Cal propagates the same focal perturbation through the matched joint continuation, $\bar{\mathbf{d}}_e^\top [D_{\mathcal{R}} \Psi_k^{\mathrm{co}}]^\top \widehat{\nabla J}$. Both use the same historical calibration-label budget as ExSpectra. JointInf directly propagates joint-continuation influence, whereas ExSpectra uses the shared-anchor first-order structure in Section 3.2, exposing complementarity and response as separate auditable channels.

**Resource controls.** PER+XR and MAC-PO+XR receive the same additional resource allowance otherwise used by ExSpectra for derivative estimation, shadow auditing, and calibration, spent on additional interaction or learner computation under the same task-level resource caps. MARR (Yang et al., 2024) retains its method-defining high replay ratio, reset, and augmentation mechanisms; all associated additional computation is charged explicitly to MARR.

**Structural ablations.** Let $f_k(e) = \bar{\mathbf{d}}_e^\top \mathbf{b}_{\mathrm{frz}}$, $c_k(e) = \bar{\mathbf{d}}_e^\top \mathbf{b}_{\mathrm{comp}}$, and $r_k(e) = \bar{\mathbf{d}}_e^\top \mathbf{b}_{\mathrm{resp}}$. w/o Comp uses $f_k + r_k$; w/o Resp uses $f_k + c_k$; w/o Cal uses $f_k + c_k + r_k$ without calibration; w/o CAUD uses calibrated $f_k$; Cal-Only removes the structural score; and FrzInf-Only uses uncalibrated $f_k$. Each calibrated ablation refits its own residual model with the same historical-label budget, excluding removed components and equivalent transformations from its features.

Table D.1: **Unified benchmark, learner, ExSpectra, and audit configurations.** Training-step limits denote backbone interaction budgets; auxiliary method costs are accounted for separately.

| Setting | Symbol / scope | Configuration |
|---|---|---|
| **Benchmark configurations** | | |
| SMACv2 | Terran°/Protoss/Zerg | 5v5; 10M steps; eval. every 100K |
| LBF-2 | 2 agents / 2 foods | $8\times8$; sight 2; forced coop.; $H_{\text{env}} = 50$; 2M; eval. every 20K |
| LBF-3 | 3 agents / 3 foods | $10\times10$; sight 2; forced coop.; horizon 50; 2M; eval. every 20K |
| RWARE | tiny-4 / small-4 | 4 agents; $H_{\text{env}} = 500$; 4M; eval. every 40K |
| Overcooked | CRing / FCoord | Coordination Ring / Forced Coordination |
| GRF | 3v1 / CAeasy | 3-vs-1 Keeper / Counterattack Easy |
| Pressure-Plate | Diagnostic | 2 agents; $7\times7$; $H_{\text{env}} = 40$; $p \in \{0, .5, 1\}$; diagnostics only |
| **Common learner and replay** | | |
| Network | Agent / mixer | 64 / 32 hidden units; non-shared agent parameters |
| Optimization | RMSProp | $\eta = 5\times10^{-4}$; decay 0.99; stabilizer $10^{-5}$; $\gamma = 0.99$ |
| Stabilization | Clip / target | Gradient norm 10; target copy every 200 learner updates |
| Exploration | $\epsilon$-greedy | $1.0 \to 0.05$ over the first 10% of training |
| Replay | $\lvert\mathcal{D}\rvert$ / $B$ | 5,000 episode-equivalents / 32 joint sequences |
| Sequence loss | Burn-in / learning | 8 / 8; half-squared mean TD loss over valid positions |
| **ExSpectra configuration** | | |
| Allocation | $M_k$ / $B$ / $H$ | 128 / 32 / 4 |
| Focal schedule | $f_k$ | Candidate-independent round robin |
| Sparse auditing | $m$ / $K_p$ / $R_p$ | 4 / 500 updates / 10 benchmark repeats; 20 on Pressure-Plate |
| Derivative refresh | $K_v$ | 200 learner updates per focal block |
| Sampling mixtures | $\epsilon_p$ / $\epsilon_s$ | 0.10 / 0.05 |
| Calibration | $\lvert\mathcal{W}_k\rvert$ / $\lambda_{\text{ridge}}$ | 256 historical labels / $10^{-4}$ |
| Local evaluation | $\beta_{\text{eval}}$ | 0.2 |
| **Held-out audits and evaluation** | | |
| Independent seeds | Training / audit | 8 matched training seeds / 3 independent audit seeds |
| Audit checkpoints | Training progress | 20%, 40%, 60%, 80% |
| RQ1 batch audit | $M_a$ / $B_a$ | 8 / 2; all $\binom{8}{2} = 28$ candidate batches |
| RQ2 mechanism audit | Candidates / horizon | 32 singletons; $H = 2$ and $H = 4$ |
| RQ3 ranking audit | $M_a$ / $B_a$ | 64 / 16 singletons |
| Cross-fitting | A/B | Disjoint half-splits; select on one half, evaluate on the other |
| Leakage control | Calibration / audit | Only $k_a < k$ labels; derivative and audit trajectories disjoint |
| Greedy reporting | Episodes / checkpoint | 32 episodes per seed |
| Resource matching | $X_q$ / $C_q$ | Per-task simulator / gradient-computation caps |

All benchmark comparisons use non-shared QMIX. $B = 32$ is the training replay budget, whereas $B_a$ is an audit selection budget. Terran° is the development task. Local replay utility uses the smooth objective with $\beta_{\text{eval}} = 0.2$; benchmark learning curves use greedy evaluation.

### D.3 Audit Design and Information Separation

Each audit fixes the task, seed, checkpoint, focal block, reference batch, and continuation protocol, with matched continuation and evaluation randomness across branches. RQ1 exhaustively evaluates all $\binom{8}{2} = 28$ batches for $M_a = 8, B_a = 2$ under $U^{\text{co}}$ and $U^{\text{frz}}$. RQ2 audits 32 singleton candidates using $\Gamma_{\text{comp}}$ and $\Gamma_{\text{resp}}$ at $H = 2$ and $H = 4$. RQ3 evaluates held-out $u^{\text{co}}$ on 64 singleton candidates with $B_a = 16$ for additive ranking. Audits use three independent seeds at $20\%, 40\%, 60\%, 80\%$ of training, with 20 paired repeats on Pressure-Plate and 10 on benchmark tasks.

**Cross-fitting and leakage control.** Paired repeats are split into disjoint halves A and B, keeping shared-reference observations together. Selections formed on one half are evaluated on the other and then swapped. Derivative trajectories are disjoint from audit evaluations, and current-checkpoint labels never enter the evaluated score or calibrator. Audits record focal identity and use balanced candidate-independent coverage whenever possible.

### D.4 Metrics and Statistical Analysis

**Normalized area under the learning curve.** For task $q$ and seed $r$, normalize greedy performance as $\widetilde{P}_{q,r}(x) = (P_{q,r}(x) - L_q)/(U_q - L_q)$ and compute

$$\text{nAUC}_{q,r} = \frac{100}{X_q} \sum_{\ell=1}^{K} (x_\ell - x_{\ell-1}) \frac{\widetilde{P}_{q,r}(x_\ell) + \widetilde{P}_{q,r}(x_{\ell-1})}{2}. \tag{D.1}$$

Here $X_q$ is the common simulator cap. SMACv2 and LBF use $(L_q, U_q) = (0, 1)$; RWARE, Overcooked, and GRF use the random-policy reference and final Native-QMIX performance from an independent seed set. Anchors are fixed independently of the compared methods, and early-ended runs carry forward their final observation without extrapolation.

**CoRVM selection gap.** For batch-audit block $b$, let $S_b^{p,*,\mathsf{A}}$ maximize half-A whole-batch utility for $p \in \{\mathrm{co}, \mathrm{frz}\}$. The held-out decision gap is $\Delta U_{\mathrm{sel},b}^{\mathsf{A}\to\mathsf{B}} = U_b^{\mathrm{co},\mathsf{B}}(S_b^{\mathrm{co},*,\mathsf{A}}) - U_b^{\mathrm{co},\mathsf{B}}(S_b^{\mathrm{frz},*,\mathsf{A}})$, averaged over both cross-fit directions. Finite-sample estimates may be negative.

**Oracle Gap Closed.** For singleton block $b$, define $V_b^{\mathsf{B}}(S) = \sum_{e\in S} \bar{u}_b^{\mathrm{co},\mathsf{B}}(e)$ and $S_b^{*,\mathsf{A}} = \mathrm{Top}_{B_a}(\bar{u}_b^{\mathrm{co},\mathsf{A}})$. Uniform selection has expected utility $V_{0,b}^{\mathsf{B}} = (B_a/M_a)\sum_e \bar{u}_b^{\mathrm{co},\mathsf{B}}(e)$. For a score-selected set $S_b^s$, let

$$\begin{cases} N_b(s) = \dfrac{1}{2}\left[V_b^{\mathsf{B}}(S_b^s) - V_{0,b}^{\mathsf{B}} + V_b^{\mathsf{A}}(S_b^s) - V_{0,b}^{\mathsf{A}}\right], \\ D_b = \dfrac{1}{2}\left[V_b^{\mathsf{B}}(S_b^{*,\mathsf{A}}) - V_{0,b}^{\mathsf{B}} + V_b^{\mathsf{A}}(S_b^{*,\mathsf{B}}) - V_{0,b}^{\mathsf{A}}\right], \\ \widehat{\mathrm{Gap}}_q(s) = 100\,\dfrac{\sum_b N_b(s)}{\sum_b D_b}. \end{cases} \tag{D.2}$$

Thus uniform replay has expected Gap zero and the cross-fitted oracle defines $100\%$. Values are not clipped; if the pooled denominator is nonpositive, the ratio is left undefined and its numerator and denominator are reported separately. Overall Gap is the equal-weight average of the five family means, with the same aggregation applied to aligned audit-seed estimates.

**Statistics.** Training seeds are the replication unit for end-to-end comparisons. Audit bootstraps preserve checkpoint sequences and shared-reference candidate groups, recomputing oracle selection, cross-fitting, and ratio aggregation within each replicate. We report stratified-bootstrap 95% confidence intervals and probability of improvement (Agarwal et al., 2021); prespecified pairwise tests use Holm correction. Spearman and Kendall correlations are secondary ranking diagnostics.

### D.5 Resource Matching

RQ4 matches simulator interaction and gradient computation. We track $N_{\mathrm{sim}} = N_{\mathrm{train}} + N_{\text{probe-eval}} + N_{\text{deriv-eval}} + N_{\text{other-learning-eval}}$ and $C_{\mathrm{grad}} = C_{\mathrm{main}} + C_{\mathrm{candidate}} + C_{\text{shared-derivatives}} + C_{\mathrm{shadow}} + C_{\mathrm{calibration}}$. All method-specific evaluation, derivative, audit, and calibration costs are charged to each method; reporting-only evaluations are excluded. PER+XR and MAC-PO+XR spend ExSpectra's additional allowance on extra interaction or learner updates under the same caps, while MARR's additional replay, reset, and augmentation costs are charged to MARR. Wall-clock time and peak memory are reported separately from the two matched resource axes.

## E Additional Experimental Results

All comparisons follow the matched-resource protocol of Appendix D, with method-specific auxiliary costs charged to the corresponding method.

### E.1 Learning Trajectories under Matched Resources

Figure E.1 complements the nAUC results with training dynamics. Protoss exhibits a delayed crossover, with ExSpectra improving more strongly after the early phase. Zerg instead favors MARR through much of training, consistent with its higher nAUC. On GRF, JointInf+Cal obtains the stronger trajectory on 3v1, whereas the CAeasy gain of ExSpectra is primarily trajectory-level rather than uniform terminal dominance. Thus the nAUC differences reflect task-dependent changes over the full finite-budget learning trajectory.

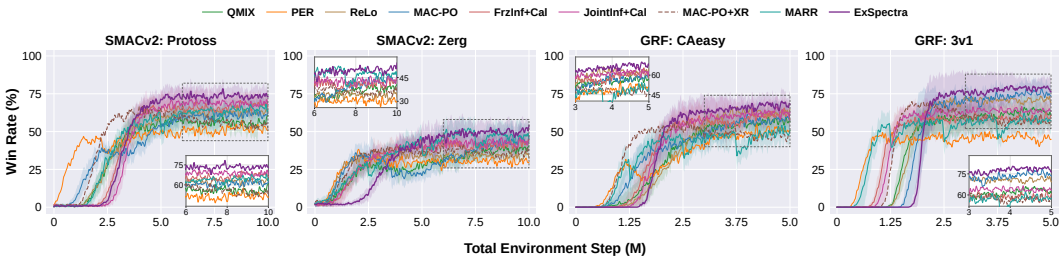



Figure E.1: **Learning trajectories under matched total resources.** Mean test performance over eight matched seeds with one-standard-deviation shading; dashed regions are enlarged in the insets.

### E.2 Audit Reliability and Cross-Fit Stability

For audit block $b$, split paired replicates into disjoint halves $\mathsf{A}, \mathsf{B}$ and measure $\rho_b^{\mathrm{SH}} = \mathrm{Spearman}_{e\in\mathcal{E}_b}\left(\bar{u}_b^{\mathrm{co},\mathsf{A}}(e), \bar{u}_b^{\mathrm{co},\mathsf{B}}(e)\right)$ and $O_b = |S_b^{*,\mathsf{A}} \cap S_b^{*,\mathsf{B}}|/B_a$.

Table E.1 shows a family-balanced split-half correlation of $0.666 \pm 0.012$ and top-$B$ overlap of $0.603 \pm 0.003$, compared with random expected overlap $B_a/M_a = 0.25$. The Gap denominator is positive in 467 of 468 benchmark blocks (99.8%). Agreement is weaker on GRF and RWARE but retains substantial rank and top-$B$ consistency. Three blocks with constant half-split utilities have undefined Spearman correlation and are excluded only from the correlation aggregate.

Table E.1: **Cross-fit audit reliability.**

| Family | Split-half $\rho \uparrow$ | Top-$B$ overlap $\uparrow$ | $D_b > 0$ (%) |
|---|---|---|---|
| SMACv2 | $0.704_{\pm 0.008}$ | $0.616_{\pm 0.020}$ | 100.0 |
| LBF | $0.743_{\pm 0.040}$ | $0.671_{\pm 0.020}$ | 100.0 |
| RWARE | $0.606_{\pm 0.011}$ | $0.568_{\pm 0.009}$ | 99.0 |
| Overcooked | $0.705_{\pm 0.006}$ | $0.633_{\pm 0.018}$ | 100.0 |
| GRF | $0.573_{\pm 0.040}$ | $0.526_{\pm 0.032}$ | 100.0 |
| **Overall** | $\mathbf{0.666}_{\pm 0.012}$ | $\mathbf{0.603}_{\pm 0.003}$ | **99.8** |
| Pressure-Plate | $0.673_{\pm 0.036}$ | $0.591_{\pm 0.020}$ | 100.0 |

Split-half/overlap: mean ± SD over three audit runs; Overall is family-balanced and $D_b > 0$ is pooled over blocks.

### E.3 Mechanism Fidelity and Allocation Diagnostics

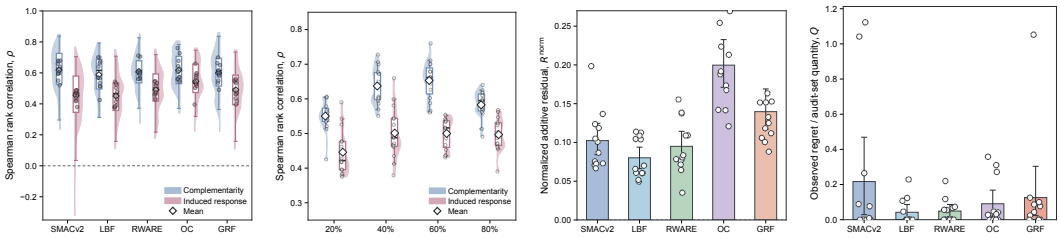
**Mechanism fidelity.** Figure E.2(a–b) shows that at the deployed horizon $H = 4$, family-balanced rank correlations between structural predictions and matched component audits are $0.606$ for complementarity and $0.486$ for induced response. Across the $20\%, 40\%, 60\%$, and $80\%$ checkpoints, complementarity correlations are $0.550, 0.637, 0.653, 0.583$, while response correlations are $0.446, 0.501, 0.500, 0.497$. Relative to $H = 2$, where the corresponding values are $0.686$ and $0.576$, longer propagation reduces but does not eliminate the ranking signal.

**Batch interaction.** For replay batch $S$, define $V_k(S) = \sum_{e\in S} u_k^{\mathrm{co}}(e)$ and $I_k(S) = U_k^{\mathrm{co}}(S) - V_k(S)$. Figure E.2(c) summarizes the normalized interaction residuals from exhaustive $M = 8, B = 2$ audits. The mean correlations between $V_k(S)$ and measured batch utility are $0.979, 0.989, 0.981, 0.941$, and $0.961$ on SMACv2, LBF, RWARE, Overcooked, and GRF, respectively. The median normalized absolute interaction residual is $0.0102$, its 90th percentile is $0.0317$, and the maximum is $0.1852$. Thus interactions are nonzero and occasionally substantial, but the singleton surrogate explains most observed batch-utility variation.

**Allocation-bound diagnostic.** On each audit pool, define $\widehat{\varepsilon}_k = \max_{e\in\mathcal{E}_k} |s_k(e) - u_k^{\mathrm{co}}(e)|$ and $\widehat{\kappa}_k = \max_{|S|=B} |U_k^{\mathrm{co}}(S) - V_k(S)|$. For audit optimum $S_k^\star$, $R_k^{\mathrm{obs}} = U_k^{\mathrm{co}}(S_k^\star) - U_k^{\mathrm{co}}(\widehat{S}_k)$. Figure E.2(d) compares this regret with $B_k^{\mathrm{audit}} = 2B\widehat{\varepsilon}_k + 2\widehat{\kappa}_k + \widehat{\Lambda}_k$. The diagnostic is conservative on audited blocks but is not a uniform certificate of Proposition 2.

### E.4 Sensitivity to Core Parameters

We vary the continuation horizon $H$, probe interval $K_p$, shared-derivative refresh interval $K_v$, and calibration-window size $|\mathcal{W}|$ independently while keeping all other settings fixed. The default is




(a) Mechanism fidelity at $H = 4$. (b) Fidelity across stages. (c) Batch additivity residual. (d) Allocation-bound diagnostic.

Figure E.2: **Mechanism and allocation diagnostics.** (a) Rank correspondence between CAUD estimates and matched audits at $H = 4$. (b) Correspondence across training stages. (c) Normalized discrepancy between measured batch utility and the additive singleton surrogate. (d) Observed allocation regret relative to $2B\widehat{\varepsilon}_k + 2\widehat{\kappa}_k + \widehat{\Lambda}_k$. Dashed lines denote zero correlation in (a–b) and equality in (d); panel (d) is a finite-audit diagnostic rather than a uniform certificate.

Table E.2: **Sensitivity of ExSpectra nAUC (%) to core parameters.**

| Parameter | Value | SMACv2 | | | LBF | | RWARE | | Overcooked | | GRF | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Terran° | Protoss | Zerg | LBF-2 | LBF-3 | T-4 | S-4 | CRing | FCoord | 3v1 | CAeasy |
| $H$ | 1 | 45.6 $\pm$3.2 | 40.7 $\pm$2.8 | 25.3 $\pm$3.0 | 57.9 $\pm$4.1 | 45.6 $\pm$5.8 | 43.1 $\pm$1.8 | 29.2 $\pm$10.2 | 32.8 $\pm$2.4 | 27.1 $\pm$3.9 | 41.5 $\pm$2.8 | 35.6 $\pm$4.3 |
| | 2 | 50.7 $\pm$2.9 | 46.2 $\pm$1.5 | 32.1 $\pm$3.9 | 61.3 $\pm$5.8 | 48.2 $\pm$3.4 | 48.4 $\pm$3.3 | 33.0 $\pm$4.0 | 39.7 $\pm$2.5 | 33.3 $\pm$8.7 | 43.7 $\pm$4.0 | 38.6 $\pm$2.8 |
| | **4** | 52.3 $\pm$2.7 | 48.3 $\pm$3.5 | 33.1 $\pm$3.9 | 61.9 $\pm$4.4 | 47.5 $\pm$4.0 | 50.2 $\pm$2.8 | 34.1 $\pm$9.2 | 40.5 $\pm$3.4 | 33.5 $\pm$5.3 | 45.7 $\pm$4.1 | 40.5 $\pm$4.5 |
| | 8 | 50.4 $\pm$3.3 | 49.2 $\pm$2.3 | 31.6 $\pm$4.0 | 58.8 $\pm$3.9 | 47.2 $\pm$5.4 | 50.2 $\pm$2.3 | 31.6 $\pm$7.5 | 40.9 $\pm$3.5 | 33.5 $\pm$5.2 | 46.6 $\pm$1.9 | 42.9 $\pm$5.5 |
| $K_p$ | 250 | 53.0 $\pm$2.9 | 50.5 $\pm$2.7 | 34.2 $\pm$5.2 | 61.2 $\pm$2.0 | 48.0 $\pm$4.4 | 52.3 $\pm$3.0 | 36.1 $\pm$7.5 | 41.1 $\pm$3.7 | 36.0 $\pm$7.5 | 44.5 $\pm$3.0 | 39.9 $\pm$4.0 |
| | **500** | 52.3 $\pm$2.7 | 48.3 $\pm$3.5 | 33.1 $\pm$3.9 | 61.9 $\pm$4.4 | 47.5 $\pm$4.0 | 50.2 $\pm$2.8 | 34.1 $\pm$9.2 | 40.5 $\pm$3.4 | 33.5 $\pm$5.3 | 45.7 $\pm$4.1 | 40.5 $\pm$4.5 |
| | 1000 | 51.5 $\pm$2.4 | 45.4 $\pm$1.7 | 32.4 $\pm$3.0 | 60.9 $\pm$3.3 | 48.3 $\pm$3.7 | 49.5 $\pm$2.6 | 36.0 $\pm$7.4 | 39.1 $\pm$4.0 | 32.2 $\pm$5.0 | 44.1 $\pm$3.6 | 38.1 $\pm$4.9 |
| | 2000 | 49.1 $\pm$2.6 | 45.6 $\pm$3.6 | 29.9 $\pm$4.1 | 59.4 $\pm$4.7 | 46.6 $\pm$3.0 | 47.1 $\pm$1.0 | 30.2 $\pm$12.1 | 38.0 $\pm$2.7 | 27.7 $\pm$6.2 | 43.5 $\pm$6.2 | 36.1 $\pm$4.4 |
| $K_v$ | 50 | 52.9 $\pm$2.5 | 48.5 $\pm$4.0 | 31.9 $\pm$3.0 | 60.6 $\pm$4.0 | 48.6 $\pm$2.3 | 50.3 $\pm$2.1 | 32.6 $\pm$8.7 | 41.3 $\pm$4.2 | 34.5 $\pm$5.0 | 45.4 $\pm$2.5 | 40.5 $\pm$4.5 |
| | 100 | 54.8 $\pm$3.4 | 48.4 $\pm$4.7 | 34.8 $\pm$3.8 | 60.5 $\pm$5.5 | 47.9 $\pm$4.1 | 49.8 $\pm$3.7 | 33.8 $\pm$8.0 | 40.1 $\pm$2.8 | 33.4 $\pm$4.3 | 46.0 $\pm$3.2 | 41.8 $\pm$3.8 |
| | **200** | 52.3 $\pm$2.7 | 48.3 $\pm$3.5 | 33.1 $\pm$3.9 | 61.9 $\pm$4.4 | 47.5 $\pm$4.0 | 50.2 $\pm$2.8 | 34.1 $\pm$9.2 | 40.5 $\pm$3.4 | 33.5 $\pm$5.3 | 45.7 $\pm$4.1 | 40.5 $\pm$4.5 |
| | 500 | 49.1 $\pm$2.9 | 46.7 $\pm$2.0 | 31.0 $\pm$1.5 | 61.2 $\pm$4.1 | 47.5 $\pm$4.0 | 49.8 $\pm$3.5 | 30.3 $\pm$6.7 | 39.2 $\pm$2.8 | 29.1 $\pm$7.3 | 43.1 $\pm$3.7 | 35.5 $\pm$5.5 |
| $\lvert\mathcal{W}\rvert$ | 64 | 51.9 $\pm$1.8 | 47.4 $\pm$3.5 | 28.1 $\pm$2.4 | 61.0 $\pm$2.1 | 45.7 $\pm$4.0 | 47.9 $\pm$1.2 | 29.3 $\pm$8.5 | 36.4 $\pm$2.7 | 29.0 $\pm$7.3 | 44.0 $\pm$6.0 | 33.2 $\pm$3.7 |
| | 128 | 51.9 $\pm$2.1 | 46.1 $\pm$2.8 | 32.0 $\pm$3.5 | 59.4 $\pm$2.7 | 47.6 $\pm$2.9 | 50.2 $\pm$2.0 | 33.1 $\pm$8.0 | 38.7 $\pm$3.4 | 30.8 $\pm$7.1 | 44.6 $\pm$4.2 | 39.7 $\pm$4.5 |
| | **256** | 52.3 $\pm$2.7 | 48.3 $\pm$3.5 | 33.1 $\pm$3.9 | 61.9 $\pm$4.4 | 47.5 $\pm$4.0 | 50.2 $\pm$2.8 | 34.1 $\pm$9.2 | 40.5 $\pm$3.4 | 33.5 $\pm$5.3 | 45.7 $\pm$4.1 | 40.5 $\pm$4.5 |
| | 512 | 52.4 $\pm$2.5 | 47.9 $\pm$4.4 | 32.7 $\pm$2.4 | 61.9 $\pm$4.4 | 47.0 $\pm$2.5 | 50.0 $\pm$3.0 | 32.4 $\pm$10.1 | 38.6 $\pm$2.8 | 32.2 $\pm$5.6 | 43.5 $\pm$4.5 | 40.3 $\pm$6.9 |

**Note.** nAUC reports mean $\pm$ SD over five matched sensitivity seeds; Gap reports pooled estimates $\pm$ audit-seed SD over three seeds. The shared default $(H, K_p, K_v, |\mathcal{W}|) = (4, 500, 200, 256)$ is repeated across sweeps and may differ from the eight-seed headline results in Table 2. Gray denotes defaults.

$(H, K_p, K_v, |\mathcal{W}|) = (4, 500, 200, 256)$. Tables E.2 and E.3 report end-to-end nAUC and replay-selection Gap, respectively.

**Sensitivity results.** Tables E.2 and E.3 show the largest degradation at $H = 1$, sparse probing ($K_p = 2000$), stale derivatives ($K_v = 500$), and a small calibration history ($|\mathcal{W}| = 64$). Increasing $H$ from 4 to 8 yields no consistent improvement. More frequent probing or refresh improves some tasks but does not dominate uniformly. Overall, the default $(4, 500, 200, 256)$ lies in a stable operating region rather than at a narrow per-task optimum.

### E.5 Scaling with Team Size

We vary system-level team scale in LBF and SMACv2 while preserving the common replay protocol within each family, testing whether CoRVM, CAUD mechanisms, selection quality, and computational cost remain meaningful as the team grows.

Table E.4 shows that CoRVM remains measurable as team scale increases, although its normalized selection gap decreases from 14.2 to 11.6 in LBF and from 10.1 to 7.6 in SMACv2. The

Table E.3: **Sensitivity of ExSpectra replay-selection Gap (%) to core parameters.**

| Parameter | Value | SMACv2 | | | LBF | | RWARE | | Overcooked | | GRF | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Terran° | Protoss | Zerg | LBF-2 | LBF-3 | T-4 | S-4 | CRing | FCoord | 3v1 | CAeasy |
| $H$ | 1 | 51.4 $\pm$3.3 | 47.5 $\pm$11.0 | 42.7 $\pm$7.8 | 71.5 $\pm$1.6 | 62.8 $\pm$4.8 | 57.5 $\pm$5.6 | 48.5 $\pm$2.4 | 49.2 $\pm$9.9 | 50.2 $\pm$7.2 | 47.0 $\pm$8.3 | 41.8 $\pm$4.7 |
| | 2 | 66.2 $\pm$0.7 | 66.7 $\pm$8.6 | 62.6 $\pm$3.8 | 76.2 $\pm$7.4 | 71.9 $\pm$3.5 | 64.3 $\pm$7.1 | 65.3 $\pm$14.4 | 63.2 $\pm$9.7 | 60.8 $\pm$6.1 | 56.5 $\pm$6.3 | 53.7 $\pm$6.7 |
| | **4** | 72.4 $\pm$8.3 | 71.8 $\pm$3.5 | 66.1 $\pm$5.9 | 79.2 $\pm$5.2 | 75.1 $\pm$6.6 | 71.1 $\pm$7.4 | 69.9 $\pm$3.5 | 67.5 $\pm$2.5 | 64.3 $\pm$4.7 | 61.2 $\pm$4.4 | 59.9 $\pm$2.5 |
| | 8 | 71.4 $\pm$4.4 | 72.9 $\pm$10.6 | 65.9 $\pm$3.3 | 79.8 $\pm$1.6 | 75.6 $\pm$6.0 | 70.5 $\pm$5.8 | 69.0 $\pm$2.2 | 66.0 $\pm$6.4 | 64.1 $\pm$9.4 | 61.2 $\pm$6.3 | 57.5 $\pm$2.4 |
| $K_p$ | 250 | 73.4 $\pm$7.7 | 74.6 $\pm$5.3 | 66.2 $\pm$5.4 | 78.8 $\pm$4.2 | 76.5 $\pm$9.1 | 72.6 $\pm$3.3 | 71.0 $\pm$8.5 | 68.2 $\pm$1.9 | 66.4 $\pm$4.5 | 65.0 $\pm$12.1 | 61.9 $\pm$5.3 |
| | **500** | 72.4 $\pm$8.3 | 71.8 $\pm$3.5 | 66.1 $\pm$5.9 | 79.2 $\pm$5.2 | 75.1 $\pm$6.6 | 71.1 $\pm$7.4 | 69.9 $\pm$3.5 | 67.5 $\pm$2.5 | 64.3 $\pm$4.7 | 61.2 $\pm$4.4 | 59.9 $\pm$2.5 |
| | 1000 | 69.6 $\pm$4.0 | 70.0 $\pm$4.8 | 62.7 $\pm$4.8 | 76.1 $\pm$6.0 | 73.4 $\pm$12.1 | 67.6 $\pm$11.3 | 65.1 $\pm$6.4 | 63.0 $\pm$7.5 | 61.2 $\pm$1.2 | 57.8 $\pm$10.2 | 57.4 $\pm$14.9 |
| | 2000 | 63.6 $\pm$8.7 | 65.3 $\pm$4.7 | 59.7 $\pm$10.3 | 73.9 $\pm$4.5 | 68.3 $\pm$8.2 | 62.8 $\pm$5.4 | 61.8 $\pm$4.7 | 53.9 $\pm$12.0 | 53.3 $\pm$1.6 | 55.3 $\pm$3.0 | 48.3 $\pm$4.5 |
| $K_v$ | 50 | 72.8 $\pm$0.6 | 74.4 $\pm$0.6 | 67.0 $\pm$7.2 | 79.3 $\pm$4.3 | 77.2 $\pm$15.2 | 74.0 $\pm$3.5 | 69.6 $\pm$4.9 | 70.5 $\pm$5.8 | 68.1 $\pm$7.3 | 64.9 $\pm$3.5 | 63.2 $\pm$3.5 |
| | 100 | 71.6 $\pm$1.3 | 73.7 $\pm$12.4 | 66.3 $\pm$6.2 | 79.2 $\pm$8.5 | 75.8 $\pm$5.3 | 69.8 $\pm$4.2 | 72.0 $\pm$8.9 | 68.6 $\pm$7.7 | 64.3 $\pm$4.2 | 61.0 $\pm$1.9 | 62.6 $\pm$3.2 |
| | **200** | 72.4 $\pm$8.3 | 71.8 $\pm$3.5 | 66.1 $\pm$5.9 | 79.2 $\pm$5.2 | 75.1 $\pm$6.6 | 71.1 $\pm$7.4 | 69.9 $\pm$3.5 | 67.5 $\pm$2.5 | 64.3 $\pm$4.7 | 61.2 $\pm$4.4 | 59.9 $\pm$2.5 |
| | 500 | 64.1 $\pm$11.9 | 62.7 $\pm$4.2 | 59.4 $\pm$3.8 | 74.4 $\pm$6.2 | 70.1 $\pm$9.3 | 65.7 $\pm$7.9 | 63.3 $\pm$4.6 | 61.9 $\pm$7.1 | 58.1 $\pm$8.4 | 54.5 $\pm$2.8 | 52.3 $\pm$7.5 |
| $\lvert\mathcal{W}\rvert$ | 64 | 61.9 $\pm$4.0 | 63.1 $\pm$4.0 | 59.0 $\pm$2.8 | 72.1 $\pm$1.7 | 69.3 $\pm$1.4 | 60.0 $\pm$0.7 | 59.3 $\pm$1.8 | 55.1 $\pm$6.7 | 55.8 $\pm$9.6 | 52.5 $\pm$7.5 | 47.7 $\pm$2.7 |
| | 128 | 69.6 $\pm$7.4 | 71.0 $\pm$8.1 | 61.4 $\pm$13.7 | 77.3 $\pm$1.9 | 74.1 $\pm$5.6 | 67.7 $\pm$7.5 | 66.7 $\pm$6.4 | 64.1 $\pm$0.9 | 62.1 $\pm$6.8 | 59.3 $\pm$2.4 | 59.1 $\pm$4.7 |
| | **256** | 72.4 $\pm$8.3 | 71.8 $\pm$3.5 | 66.1 $\pm$5.9 | 79.2 $\pm$5.2 | 75.1 $\pm$6.6 | 71.1 $\pm$7.4 | 69.9 $\pm$3.5 | 67.5 $\pm$2.5 | 64.3 $\pm$4.7 | 61.2 $\pm$4.4 | 59.9 $\pm$2.5 |
| | 512 | 70.6 $\pm$16.3 | 70.4 $\pm$11.9 | 64.8 $\pm$7.2 | 78.7 $\pm$5.7 | 75.7 $\pm$4.4 | 69.4 $\pm$6.3 | 69.4 $\pm$9.3 | 66.9 $\pm$0.3 | 63.3 $\pm$7.4 | 60.4 $\pm$2.7 | 59.7 $\pm$3.3 |

Gap is the pooled estimate $\pm$ audit-seed SD over three seeds. Gray bands denote frozen defaults. T-4 and S-4 denote RWARE tiny-4 and small-4.

Table E.4: **Co-adaptive replay under increasing team scales.** Comp. and Resp. are relative mechanism magnitudes. Replay-selection deltas report ExSpectra Gap differences from w/o CAUD and JointInf+Cal; the learning delta is the nAUC difference from JointInf+Cal.

| Family | Scale | Co-adaptation | | | Replay selection | | | Learning | | Cost | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Sel. Gap | Comp. | Resp. | Ex. Gap | $\Delta\mathbf{Gap}_{\mathrm{Frz}}$ | $\Delta\mathbf{Gap}_{\mathrm{Joint}}$ | Ex. nAUC | $\Delta\mathbf{nAUC}_{\mathrm{Joint}}$ | ms/cand. | Joint/Ex |
| LBF | 2 | 14.2 $\pm$0.4 | 26.0 $\pm$0.3 | 44.0 $\pm$3.3 | 79.0 $\pm$1.1 | +22.8 | +5.6 | 60.8 $\pm$2.9 | +0.1 | 0.93 | 1.72× |
| | 3 | 14.0 $\pm$0.6 | 28.0 $\pm$2.4 | 42.0 $\pm$2.2 | 75.0 $\pm$1.9 | +20.0 | +6.4 | 47.1 $\pm$2.2 | +4.6 | 1.25 | 1.66× |
| | 4 | 13.1 $\pm$0.5 | 31.0 $\pm$2.4 | 38.0 $\pm$4.3 | 74.0 $\pm$2.6 | +19.8 | +5.1 | 43.8 $\pm$3.3 | +2.9 | 1.36 | 1.96× |
| | 6 | 11.6 $\pm$0.6 | 35.0 $\pm$3.3 | 33.0 $\pm$2.7 | 69.8 $\pm$0.9 | +17.2 | +3.4 | 37.9 $\pm$1.9 | +1.2 | 1.90 | 2.15× |
| SMACv2 | 5v5 | 10.1 $\pm$0.6 | 32.0 $\pm$3.1 | 36.0 $\pm$2.9 | 72.3 $\pm$3.2 | +17.4 | +2.7 | 49.2 $\pm$3.5 | +4.1 | 1.42 | 1.73× |
| | 10v10 | 9.2 $\pm$0.1 | 34.0 $\pm$2.8 | 33.0 $\pm$4.3 | 68.1 $\pm$3.0 | +15.0 | +2.8 | 40.2 $\pm$1.1 | +1.1 | 2.91 | 2.27× |
| | 20v20 | 7.6 $\pm$0.6 | 37.0 $\pm$2.9 | 29.0 $\pm$1.4 | 63.2 $\pm$3.8 | +11.5 | −0.1 | 28.6 $\pm$2.9 | −0.8 | 7.02 | 2.53× |

Sel. Gap, Comp., Resp., and Gap use three audit seeds; nAUC uses five matched training seeds. Comp. and Resp. are $100\,\mathbb{E}|\Gamma_{\mathrm{comp}}|/\mathbb{E}|u^{\mathrm{co}}|$ and $100\,\mathbb{E}|\Gamma_{\mathrm{resp}}|/\mathbb{E}|u^{\mathrm{co}}|$. Team scale changes with the benchmark configuration and is interpreted as system-level scaling rather than an isolated intervention on agent count.

mechanism composition shifts toward complementarity as induced response decreases. ExSpectra retains a substantial advantage over frozen valuation at every scale, while its Gap advantage over JointInf+Cal narrows and essentially vanishes at SMACv2 20v20. Meanwhile, the JointInf/ExSpectra scoring-cost ratio grows from 1.73× at 5v5 to 2.53× at 20v20, exposing an increasing accuracy–computation tradeoff at larger team scales.