RECOVERABLE COORDINATION DISTILLATION FOR DECENTRALIZED MULTI-AGENT REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Centralized training with decentralized execution (CTDE) allows multi-agent learners to exploit privileged information during training, but deployed agents must act from local inputs. Existing approaches may attempt to reconstruct missing teammate information, although much of this information can be irrelevant to the current decision. We propose Recoverable Coordination Distillation (RCD), which instead predicts the locally recoverable *effect* of privileged information on a teacher's action preferences. A frozen teacher defines this coordination effect as the centered difference between its action logits with and without privileged input. A local predictor estimates the conditional mean and uncertainty of this effect, while a reliability-aware fusion mechanism regulates its influence on the decentralized policy. At deployment, all computations use only execution-available local information, without the teacher, privileged inputs, or communication. We characterize the locally recoverable component of the teacher effect and derive a return bound relating prediction errors to policy performance. Across seven tasks from MPE, LBF, and RWARE, RCD achieves higher mean final return than our MAPPO-based MA^{2}E baseline in all 14 task–input comparisons. With local history, it also achieves higher mean learning-curve AUC than the Local baseline on all seven tasks. These results support recovering decision-relevant coordination effects rather than reconstructing missing information for decentralized multi-agent control. Our experimental code is available at https://anonymous.4open.science/r/RCD-30C2/.

1 INTRODUCTION

Centralized training with decentralized execution (CTDE) allows multi-agent learners to exploit joint information during training while requiring each deployed agent to act from locally available information (Lowe et al., 2017; Foerster et al., 2018; Rashid et al., 2018; Yu et al., 2022). This creates a fundamental information asymmetry: privileged information, such as other agents' observations or the global state, can reveal how agents should coordinate during training, but is unavailable to an individual agent at execution. A central question is therefore how to retain the *decision-relevant benefit* of such privileged information under the decentralized execution constraint.

One approach is to infer the missing information itself. For example, MA^{2}E reconstructs unavailable teammate observations and actions from an agent's local history and retains this reconstruction module at deployment (Kang et al., 2025). The reconstructed teammate information is then used for action selection. This is useful when the missing trajectories can be predicted accurately, but reconstruction can be unnecessarily demanding: much of a teammate's state or trajectory may be irrelevant to the current decision.

Consider an agent deciding whether to move left or right. To avoid a collision, it may not need to recover the precise positions, velocities, and action histories of all teammates. It may only need to know that the information unavailable to it would make moving right less desirable. More generally, two hidden situations can be difficult to distinguish from local information while inducing essentially the same change in the preferred action. This suggests a different objective: rather than predicting *what privileged information is missing*, predict *how that information would change the decision*.

Motivated by this principle, we introduce Recoverable Coordination Distillation (RCD). RCD transfers the locally predictable effect of privileged information on a centralized teacher's action preferences. A frozen teacher is evaluated on the same transition with and without access to privileged information. Their centered logit difference defines a *teacher coordination effect*: a decision-level representation of how the privileged information changes the teacher's relative action preferences. The deployed student does not observe this effect directly. Instead, it predicts the part of the effect that can be inferred from its legal local information.

This distinction is important because decentralized inference cannot create information that is absent from the local input. If several privileged situations are indistinguishable to an agent, no local predictor can identify which one actually occurred. Their effects on action selection, however, may still share a predictable component. RCD therefore models both the conditional mean of the teacher coordination effect and its conditional uncertainty. The predicted mean provides coordination guidance, while the predicted uncertainty is converted into a reliability score that regulates how strongly this guidance

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.

influences the student policy. When the hidden situations imply consistent action corrections, the guidance can be used strongly; when they imply conflicting corrections, its influence can be reduced.

This perspective also differs from conventional privileged teacher–student approaches. Methods such as CTDS, ADVISOR, and MAGPO use privileged teachers or guiders to improve decentralized policies during training (Zhao et al., 2022; Weihs et al., 2021; Li et al., 2026). Their privileged guidance is removed once training is complete. In contrast, RCD learns an explicit *local prediction of the privileged coordination effect* that remains part of the deployed policy. Like MA^{2}E, it therefore retains an inference mechanism at execution, but changes its prediction target from missing teammate information to the decision-relevant effect of that information.

Concretely, a local predictor estimates the mean and uncertainty of the teacher's coordination effect. The predicted effect is projected into the student representation, while its uncertainty controls a reliability-aware fusion gate. The predictor learns from teacher targets, and reinforcement learning trains the policy to use this guidance. At deployment, all components rely only on execution-available local information; the teacher, privileged information, and centralized critic are removed.

Our contributions are threefold:

- We formulate decentralized privileged-information transfer as prediction of *recoverable decision effects*, separating the locally predictable teacher coordination effect from irreducible uncertainty.
- We introduce RCD, which combines coordination-effect prediction, uncertainty estimation, and reliability-aware fusion, with a return bound relating prediction errors to policy performance.
- We evaluate RCD on seven MPE, LBF, and RWARE tasks. It outperforms our MAPPO-based MA^{2}E baseline in all 14 task–input comparisons, and with local history achieves higher mean AUC than Local on all seven tasks.

2 Related Work

Centralized training and decentralized execution. Centralized training with decentralized execution (CTDE) uses joint information during training while requiring each deployed agent to act from locally available inputs. Centralized critics, counterfactual objectives, and value decomposition are widely used within this framework (Lowe et al., 2017; Foerster et al., 2018; Rashid et al., 2018; Yu et al., 2022). These methods exploit information unavailable to individual agents at execution to improve decentralized policies. RCD operates under the same execution constraint, but uses privileged information to supervise a local predictor whose output remains part of the deployed policy.

Deployment-time reconstruction. One way to cope with missing execution-time information is to predict the unavailable information itself. MA^{2}E uses a masked autoencoder to reconstruct missing teammate observations and actions from an agent's local history (Kang et al., 2025). The reconstructed information remains available to the policy at deployment and supports decentralized action selection. RCD similarly retains an inference mechanism during execution, but changes the prediction target. Rather than reconstructing hidden teammate trajectories, it predicts how privileged information would change a teacher's action preferences. This focuses prediction on a decision-relevant effect and additionally models its uncertainty to regulate how strongly the predicted guidance influences the policy.

Privileged teacher guidance and distillation. Teacher-guided methods transfer information from a policy or value function with richer training-time information to a locally executable student. CTDS trains decentralized students to approximate centralized teacher values (Zhao et al., 2022). ADVISOR adapts the balance between imitation and reinforcement learning according to how well a student can imitate a privileged teacher (Weihs et al., 2021). MAGPO uses a centralized guider to supervise decentralized learners during policy optimization (Li et al., 2026). In these approaches, privileged guidance is used during training, while the deployed student acts independently of the teacher.

RCD also uses privileged teacher supervision, but retains an explicit *local prediction* of the teacher's coordination effect within the deployed policy. A frozen teacher provides a fixed decision-effect target, while reinforcement learning trains the student to use the predicted guidance through reliability-aware fusion. Thus, RCD separates prediction of privileged guidance from its use for action selection. We do not claim that learning locally predictable teacher outputs is itself new; standard distillation can also transfer predictable aspects of a teacher. The distinction lies in predicting the *change in action preference induced by privileged information*, modeling the ambiguity of this effect, and explicitly using both at deployment.

Local history and predictive uncertainty. Local history is commonly used to mitigate partial observability in reinforcement learning (Hausknecht & Stone, 2015; Omidshafiei et al., 2017). RCD can likewise condition on local history, but additional processing cannot recover distinctions that are absent from the execution-available information. We therefore model the conditional mean and variance of the teacher coordination effect using input-dependent Gaussian prediction, following the general approach of heteroscedastic uncertainty estimation (Kendall & Gal, 2017). The predicted variance is converted into a

reliability signal that regulates the use of coordination guidance. This uncertainty estimate characterizes ambiguity in the predicted teacher effect; by itself, it does not imply calibration or improved policy return.

3 Problem Formulation

We consider a cooperative decentralized partially observable Markov decision process (Dec-POMDP) with N agents, $\mathcal{M} = \langle \mathcal{S}, \{\mathcal{A}_i\}_{i=1}^N, P, r, \{\mathcal{O}_i\}_{i=1}^N, O, \gamma \rangle$, where $\mathcal{S}$ is the global state space, $\mathcal{A}_i$ and $\mathcal{O}_i$ are the action and observation spaces of agent i, P and O are the transition and observation functions, r is the shared reward, and $\gamma \in [0, 1)$ is the discount factor. At time t, agent i acts using only its execution-available information $X_{i,t}$, which may include its current observation, local history, identity, role, and permitted messages.

The joint decentralized policy factorizes as $\pi^S(\mathbf{a}_t \mid \mathbf{X}_t) = \prod_{i=1}^N \pi_i^S(a_{i,t} \mid X_{i,t})$, where $\mathbf{a}_t$ and $\mathbf{X}_t$ denote the joint action and execution-available information. During centralized training, a teacher additionally observes privileged information $U_{i,t}$, such as the global state or other agents' observations, and follows $\pi_i^T(a_{i,t} \mid X_{i,t}, U_{i,t})$. Because $U_{i,t}$ is unavailable at execution, a decentralized student cannot in general reproduce teacher decisions that depend on privileged distinctions not inferable from $X_{i,t}$.

Our goal is therefore to learn a decentralized student that maximizes the expected discounted team return $J(\pi^S) = \mathbb{E}_{\pi^S}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$, while using the teacher during training to identify coordination guidance that is predictable from $X_{i,t}$. The remaining ambiguity arises from privileged distinctions that cannot be resolved from local information. Section 4 formalizes this notion of recoverability, and Section 5 develops the proposed method.

4 Coordination Recoverability

Unless stated otherwise, expectations are taken under a fixed reference distribution $\mathcal{D}$ over training transitions. The teacher is frozen, so its target function is fixed; however, the conditional moments of that target depend on the distribution of local information under $\mathcal{D}$.

4.1 Local Information Limit

Proposition 1 (Local prediction adds no execution information)**.** *For any deterministic local prediction $R_i = h(X_i)$, an actor $\pi_i(a_i \mid X_i, R_i)$ has an equivalent local-only actor $\widetilde{\pi}_i(a_i \mid x_i) = \pi_i(a_i \mid x_i, h(x_i))$.*

The result follows directly by substitution; Appendix A.1 provides the proof. A local predictor can reorganize or expose information already contained in X_i, but cannot distinguish privileged situations that produce the same X_i. Importantly, such situations may still induce similar changes in action preference. Predicting this common decision effect can therefore be easier than reconstructing the hidden situation itself.

4.2 Teacher Coordination Effect

Let $\ell_i^T(X_i, U_i) \in \mathbb{R}^{d_i}$ denote the action logits of a frozen teacher with access to privileged information U_i. On the same transition, we also evaluate the teacher with the privileged path removed, yielding $\ell_i^T(X_i)$. We define the teacher coordination effect as

$$C_i^T = \mathrm{ctr}\big(\ell_i^T(X_i, U_i) - \ell_i^T(X_i)\big), \qquad \mathrm{ctr}(v) = v - \frac{\mathbf{1}^\top v}{d_i}\mathbf{1}. \tag{1}$$

Centering removes action-independent logit shifts, which do not affect softmax probabilities. Thus, each component of C_i^T represents the change in the teacher's relative preference for an action caused by privileged information.

Both evaluations use the same frozen teacher parameters. The restricted evaluation is therefore a reference intervention, not a separately trained decentralized teacher, and must remove every privileged information path. Appendix B discusses the target construction in detail.

4.3 Recoverable Effect and Remaining Ambiguity

Assuming $\mathbb{E}\|C_i^T\|_2^2 < \infty$, we decompose the teacher effect as

$$C_i^T = \underbrace{\mathbb{E}[C_i^T \mid X_i]}_{\mu_i^*(X_i) = C_{i,\mathrm{rec}}^T} + \underbrace{C_i^T - \mathbb{E}[C_i^T \mid X_i]}_{\varepsilon_i = C_{i,\mathrm{irr}}^T}, \qquad \mathbb{E}[\varepsilon_i \mid X_i] = 0. \tag{2}$$

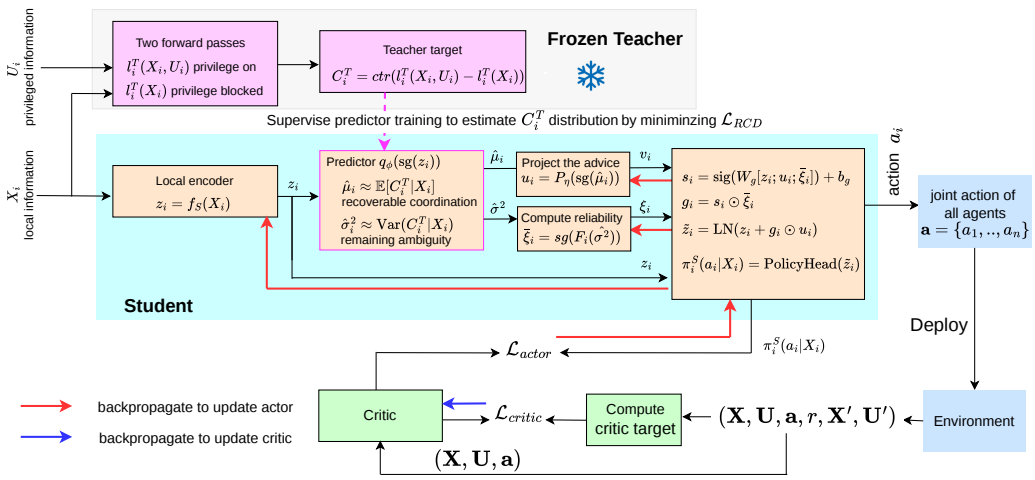

Figure 1: RCD first trains and freezes a privileged teacher, then trains a decentralized student from teacher coordination-effect targets. At deployment, only the local student remains. Prediction and policy learning use separate gradient paths.

The recoverable effect $\mu_i^*(X_i)$ is the part of the teacher's coordination effect that is statistically predictable from execution-available information. The residual ε_i captures variation that remains unresolved given X_i. For example, two hidden teammate configurations may be locally indistinguishable yet both make moving left more desirable. Their shared preference change remains predictable. If they instead favor opposite actions, the realized teacher effects cannot both be recovered from the same local input.

For any square-integrable local predictor h, conditional orthogonality gives

$$\mathbb{E}\|C_i^T - h(X_i)\|_2^2 = \mathbb{E}\|\varepsilon_i\|_2^2 + \mathbb{E}\|\mu_i^*(X_i) - h(X_i)\|_2^2. \tag{3}$$

Hence, μ_i^* minimizes the local squared prediction error, while $\mathbb{E}\|\varepsilon_i\|_2^2$ is the irreducible prediction error given X_i. The conditional covariance

$$\Sigma_i^*(X_i) = \mathrm{Var}(C_i^T \mid X_i) = \mathbb{E}[\varepsilon_i \varepsilon_i^\top \mid X_i] \tag{4}$$

quantifies the remaining ambiguity. Similar effects across hidden situations produce low uncertainty, whereas conflicting effects produce higher uncertainty. Appendix A.2 proves Equation (3). Appendix A.3 further shows that, for a fixed teacher target and data distribution, discarding local information cannot reduce the minimum achievable squared prediction error.

5 Recoverable Coordination Distillation

RCD consists of two training stages followed by decentralized deployment (Figure 1). The teacher and student are distinct networks. Parameters may be shared across agents within each network, but each deployed student processes only its own execution-available information.

Stage 1: train and freeze the teacher. We first train a privileged MARL teacher using both local information X_i and privileged information U_i. After training, the teacher is frozen and used only to construct the coordination-effect target C_i^T; it does not select actions during student rollouts.

Stage 2: train the local student. The student interacts with the environment using only X_i. During training, the system retains U_i to evaluate the frozen teacher with and without privileged information and construct C_i^T. The student predicts the recoverable effect and its uncertainty from local information, then uses a reliability-aware gate to incorporate this guidance into action selection. Prediction learning fits the teacher target, while reinforcement learning determines how the predicted guidance should influence team return. Appendix D details the training sequence and gradient routing.

5.1 Local Coordination-Effect Predictor

The student encodes its local information as $z_i = f_S(X_i) \in \mathbb{R}^{d_z}$. A local predictor estimates the action-wise mean and variance of the teacher coordination effect:

$$(\widehat{\mu}_i, \log \widehat{\sigma}_i^2) = q_\phi(\mathrm{sg}(z_i)), \qquad \widehat{\mu}_i, \widehat{\sigma}_i^2 \in \mathbb{R}^{d_i}, \tag{5}$$

where sg denotes stop-gradient.

We fit a diagonal Gaussian surrogate by minimizing

$$\mathcal{L}_{\mathrm{RCD}} = \frac{1}{2N} \sum_{i=1}^{N} \sum_{k=1}^{d_i} \left[\frac{(C_{i,k}^{T} - \widehat{\mu}_{i,k})^2}{\widehat{\sigma}_{i,k}^2} + \log \widehat{\sigma}_{i,k}^2 \right]. \tag{6}$$

The predicted mean approximates the locally recoverable effect, while the variance captures residual prediction ambiguity. The variance is clipped to $[\sigma_{\min}^2, \sigma_{\max}^2]$ for numerical stability. Because the centered action components are dependent, the diagonal Gaussian models their marginal conditional moments rather than the full joint distribution.

Since the predictor receives z_i rather than X_i directly, its population-optimal moments are conditioned on z_i. They coincide with the moments conditioned on X_i when the encoder preserves the relevant information; otherwise, representation loss introduces additional ambiguity. Appendix C gives the corresponding analysis.

We summarize the predicted uncertainty by the reliability score

$$\widehat{\xi}_i = F_i(\widehat{\sigma}_i^2), \qquad F_i(v) = \exp\left[-\frac{1}{d_i} \sum_{k=1}^{d_i} \log(1 + v_k) \right] \in (0, 1]. \tag{7}$$

Lower predicted variance gives higher reliability, while larger uncertainty reduces the influence available to the predicted guidance.

5.2 Reliability-Aware Fusion

The predicted coordination effect is projected into the student's feature space, while both predictor-derived quantities are detached from the reinforcement-learning gradient:

$$u_i = P_\eta(\mathrm{sg}(\widehat{\mu}_i)) \in \mathbb{R}^{d_z}, \qquad \bar{\xi}_i = \mathrm{sg}(\widehat{\xi}_i). \tag{8}$$

We then define a feature-wise fusion gate

$$g_i = \bar{\xi}_i \, \sigma\big(W_g[z_i; u_i; \bar{\xi}_i] + b_g\big), \qquad g_i \in (0, 1)^{d_z}, \tag{9}$$

$$\widetilde{z}_i = \mathrm{LN}(z_i + g_i \odot u_i), \qquad \pi_i^S(\cdot \mid X_i) = \mathrm{PolicyHead}_S(\widetilde{z}_i). \tag{10}$$

Here $[\cdot;\cdot]$ denotes concatenation, σ is the sigmoid, $\odot$ is element-wise multiplication, and LN denotes layer normalization. The learned sigmoid determines which feature corrections are useful, while ξ_i limits their magnitude according to predicted reliability. When the predicted effect is uncertain, the gate is suppressed and the policy remains closer to its local pathway. A negative initial gate bias further keeps early training close to this pathway while the predictor is still inaccurate.

5.3 Learning and Deployment

Learning. The student uses reinforcement-learning and prediction objectives,

$$\mathcal{L} = \mathcal{L}_{\mathrm{RL}} + \lambda_C \mathcal{L}_{\mathrm{RCD}}, \qquad \mathcal{L}_{\mathrm{RL}} = \mathcal{L}_{\mathrm{actor}} + c_V \mathcal{L}_{\mathrm{critic}} - c_H \mathcal{H}(\pi^S). \tag{11}$$

In our MAPPO implementation, $\mathcal{L}_{\mathrm{actor}}$ is the PPO clipped objective, the centralized critic estimates returns, and the entropy term encourages exploration. Let θ_S denote the encoder, projection, gate, and policy-head parameters. We separate prediction learning from policy learning through

$$\nabla_{\theta_S} \mathcal{L}_{\mathrm{RCD}} = 0, \qquad \nabla_{\phi} \mathcal{L}_{\mathrm{RL}} = 0. \tag{12}$$

Thus, $\mathcal{L}_{\mathrm{RCD}}$ trains only the local effect predictor, whereas reinforcement learning trains the student to use the predicted guidance when it improves team return. Detaching both the predicted mean and reliability prevents policy gradients from modifying the predictor. The teacher remains frozen throughout student training.

Deployment. At execution, agent i computes $X_i \to z_i \to (\widehat{\mu}_i, \widehat{\sigma}_i^2) \to \widehat{\xi}_i \to \widetilde{z}_i \to \pi_i^S(\cdot \mid X_i)$. All quantities are functions only of execution-available information X_i. The teacher, privileged information, teacher-effect target, and centralized critic are discarded. Thus, RCD introduces no additional execution information; it uses local information to predict and selectively exploit the recoverable decision effect learned from privileged supervision.

6 Effect of Prediction Error on Return

We analyze how errors in predicting the recoverable coordination effect and its reliability affect return. To isolate prediction error from policy optimization, we compare the learned student with a decentralized reference that uses the exact conditional moments in the *same fixed* fusion and policy network.

Recall $\mu_i^*(X_i) = \mathbb{E}[C_i^T \mid X_i]$. Define the corresponding conditional variance and reliability as

$$\sigma_{i,k}^{*2}(X_i) = \mathrm{Var}(C_{i,k}^T \mid X_i), \qquad \xi_i^*(X_i) = F_i(\sigma_i^{*2}(X_i)), \tag{13}$$

where F_i is defined in Equation (7). These are population quantities used only for analysis; they are not additional information available to the deployed agent.

Let $\Pi_\theta(\cdot \mid X, \boldsymbol{\mu}, \boldsymbol{\xi})$ denote the joint action distribution induced by the student's fixed fusion and policy network. We compare

$$\pi_C(\cdot \mid X) = \Pi_\theta(\cdot \mid X, \boldsymbol{\mu}^*, \boldsymbol{\xi}^*), \qquad \widehat{\pi}(\cdot \mid X) = \Pi_\theta(\cdot \mid X, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\xi}}). \tag{14}$$

Both policies remain decentralized because each agent's supplied quantities depend only on its own X_i. The reference π_C is neither the privileged teacher nor an optimal decentralized policy; it differs from $\widehat{\pi}$ only by replacing learned predictions with their exact conditional counterparts.

Theorem 1 (Return sensitivity to coordination-prediction error)**.** *Assume $|r_t| \le R_{\max}$ and a common initial-state distribution. Suppose that, uniformly over X, the fixed policy network satisfies*

$$D_{\mathrm{TV}}(\Pi_\theta(\cdot \mid X, \boldsymbol{\mu}, \boldsymbol{\xi}), \Pi_\theta(\cdot \mid X, \boldsymbol{\mu}', \boldsymbol{\xi}')) \le L_\mu \sum_{i=1}^N \|\mu_i - \mu_i'\|_2 + L_\xi \sum_{i=1}^N |\xi_i - \xi_i'|. \tag{15}$$

Let $d_{\widehat{\pi}}$ denote the normalized discounted occupancy of X under $\widehat{\pi}$, and define

$$e_\mu = \sum_{i=1}^N \mathbb{E}_{d_{\widehat{\pi}}}\left[\|\widehat{\mu}_i(X_i) - \mu_i^*(X_i)\|_2\right], \qquad e_\xi = \sum_{i=1}^N \mathbb{E}_{d_{\widehat{\pi}}}\left[|\widehat{\xi}_i(X_i) - \xi_i^*(X_i)|\right]. \tag{16}$$

Then

$$|J(\pi_C) - J(\widehat{\pi})| \le \frac{2R_{\max}}{(1-\gamma)^2}\left(L_\mu e_\mu + L_\xi e_\xi\right). \tag{17}$$

Appendix E gives the proof on full interaction histories, so X_i may represent either the current local input or a local history. The bound shows that, for a fixed policy network, small errors in the predicted coordination effect and reliability imply a small return deviation from the exact-conditional-moment reference. The constants L_μ and L_ξ measure the policy's sensitivity to these two sources of error. Representation loss, finite-sample estimation, variance clipping, and distribution shift may all contribute to e_μ and e_ξ.

Importantly, the result does not require the conditional uncertainty to vanish. When locally indistinguishable privileged situations induce different teacher effects, even π_C uses a reduced reliability score. The theorem therefore quantifies *estimation error around recoverable guidance*, rather than the full gap to the privileged teacher. It does not establish training convergence, superiority over a local baseline, or optimality of the coordination-effect target. Even with exact conditional moments, a gap to the teacher may remain because the teacher observes the realized privileged information U_i.

7 Experiments

We evaluate seven tasks: SpreadXY-2, SpreadXY-4, and SpreadBlindfold from MPE (Santos et al., 2025); LBF-8×8 and LBF-10×10 with sight range two; and RWARE-Tiny and RWARE-Small (Papoudakis et al., 2021; Christianos et al., 2020). They cover controlled information removal, spatial partial observability, and distributed warehouse sensing. SpreadXY partitions motion information along the two spatial axes, whereas SpreadBlindfold hides teammate motion states. In LBF and RWARE, missing information changes with the agents' locations. This suite therefore covers both fixed information restrictions and information distributed dynamically across agents. Appendix F.1 discusses the rationale for the environment selection, and Appendix F.4 provides detailed descriptions of the corresponding environments.

All methods use MAPPO (Yu et al., 2022), with actor and critic parameters shared across agents within each method and a training-only centralized critic. We compare four actor designs: **Joint**, a privileged reference and teacher using all agents' observations; **Local**, a decentralized policy; **Local+MA^{2}E**, our MAPPO adaptation of teammate-information reconstruction (Kang et al., 2025); and **Local+RCD**, which predicts the coordination effect of the corresponding frozen Joint teacher. Each design uses either current observations or a five-step window of observations and previous actions, giving eight methods. Decentralized methods receive the same raw local inputs within each setting. The primary history encoders are feed-forward. The details of these methods are provided in Appendix F.5 with information of MAPPO implementation and padding masks. Appendix H provides the hyperparameter settings used in the experiments.

For each method, we conduct five independent training runs using the same set of seeds across methods. Each run is trained for 10 million environment steps. During training, the policy is evaluated periodically; at each evaluation checkpoint, we execute 100 evaluation episodes and use their mean episodic return as the checkpoint-level performance estimate. For each run, we then average these checkpoint-level returns over the final 20% of training, corresponding to the last 2 million environment steps. This produces one final-phase return for each seed. We report the mean ± sample standard deviation of these five run-level values. Higher returns indicate better performance, and all comparisons below refer to these reported means.

7.1 Main Results

Table 1: Final return (mean ± sample SD, five seeds). MA^{2}E and RCD denote Local+MA^{2}E and Local+RCD, respectively. Bold: best decentralized mean per row; higher is better.

Input	Task	Joint	Local	MA^{2}E	RCD
Current observation	SpreadXY-2	-76.16 ± 0.38	-92.39 ± 9.20	-109.73 ± 0.68	$\mathbf{-87.74 \pm 0.50}$
	SpreadXY-4	-241.44 ± 9.38	$\mathbf{-277.90 \pm 11.71}$	-301.34 ± 6.82	-286.97 ± 7.73
	SpreadBlindfold	-137.89 ± 4.83	-151.14 ± 4.71	-177.22 ± 1.49	$\mathbf{-134.85 \pm 0.68}$
	LBF-8×8	0.87 ± 0.02	0.73 ± 0.05	0.50 ± 0.04	$\mathbf{0.76 \pm 0.09}$
	LBF-10×10	0.59 ± 0.03	0.38 ± 0.03	0.24 ± 0.01	$\mathbf{0.50 \pm 0.01}$
	RWARE-Tiny	35.83 ± 1.01	$\mathbf{21.90 \pm 1.53}$	13.31 ± 2.44	21.81 ± 2.30
	RWARE-Small	13.12 ± 1.86	$\mathbf{8.60 \pm 1.09}$	2.94 ± 0.44	8.14 ± 1.20
Five-step history	SpreadXY-2	-76.05 ± 0.45	-87.68 ± 4.25	-98.12 ± 10.44	$\mathbf{-85.86 \pm 3.85}$
	SpreadXY-4	-235.50 ± 19.23	$\mathbf{-283.92 \pm 11.97}$	-286.82 ± 5.71	-286.80 ± 6.34
	SpreadBlindfold	-135.13 ± 0.42	-147.79 ± 2.09	-155.08 ± 6.97	$\mathbf{-136.80 \pm 4.76}$
	LBF-8×8	0.96 ± 0.00	0.95 ± 0.02	0.75 ± 0.03	$\mathbf{0.96 \pm 0.00}$
	LBF-10×10	0.84 ± 0.02	0.71 ± 0.03	0.45 ± 0.02	$\mathbf{0.83 \pm 0.02}$
	RWARE-Tiny	38.37 ± 1.12	$\mathbf{26.39 \pm 4.59}$	24.18 ± 1.34	26.26 ± 1.66
	RWARE-Small	13.62 ± 2.94	12.30 ± 1.17	3.93 ± 0.74	$\mathbf{12.38 \pm 1.81}$

Final return. RCD has higher mean final return than MA^{2}E in all 14 task–input comparisons (Table 1). Against Local, it improves in $4/7$ tasks with current observations and $5/7$ with history. The clearest gains occur on SpreadBlindfold and LBF-10×10. On SpreadBlindfold, the improvements over Local are 16.29 return points without history and 10.99 with history. The gains are task dependent: RCD remains below Local on SpreadXY-4 in both settings, while final returns on RWARE are similar to Local.

Contribution of local history. History improves RCD's final return in $6/7$ tasks, with particularly clear gains on LBF and RWARE. For example, RCD increases from 0.50 to 0.83 on LBF-10×10 and from 8.14 to 12.38 on RWARE-Small. Local also benefits from history in six tasks, showing that temporal context is useful independently of RCD. With history, RCD nearly matches Joint on both LBF tasks, while the remaining RWARE-Tiny gap (26.26 versus 38.37) shows that a substantial benefit from privileged information can persist.

7.2 Learning Efficiency

Final return reported in Section 7.1 summarizes performance near the end of training. To assess performance during learning, we also report the area under the evaluation learning curve (AUC). This curve plots the average episodic return at each evaluation checkpoint against the number of training steps. For each seed, we connect consecutive evaluation points with straight lines and calculate the resulting area using trapezoidal integration. We then divide this area by the 10M-step training horizon to obtain a measure of average performance during training (Table 2). Higher AUC favors methods that achieve high returns earlier and sustain them. We compare AUC only within the same task because reward scales differ across tasks. Appendix F.2 provides the exact calculation.

Current observations. RCD has higher mean AUC than Local in five of seven tasks and higher mean AUC than MA^{2}E in all seven. On SpreadBlindfold, it improves over Local from -162.717 to -145.383, consistent with the final-return advantage. It also improves on SpreadXY-2 and both LBF tasks, while RWARE-Tiny is nearly tied (13.945 versus 13.764). The exceptions are SpreadXY-4 (-311.866 versus -297.424) and RWARE-Small (3.487 versus 3.694). Thus, useful transfer from instantaneous observations remains task dependent.

Five-step history. With history, RCD has higher mean AUC than both Local and MA^{2}E on all seven tasks. Its gains on the warehouse tasks are more apparent in AUC than in final return: RWARE-Tiny increases from 17.581 to 21.190, and RWARE-Small from 5.790 to 7.517, despite similar final returns. On LBF-10×10, AUC increases from 0.580 to 0.759. RCD also exceeds the trained Joint-History reference on LBF-10×10 and RWARE-Small; these references are empirical comparisons, not performance upper bounds. SpreadXY-4 illustrates the distinction between metrics: RCD has slightly higher AUC but lower final return than Local.

What history adds. History increases RCD's mean AUC in all seven tasks, although the SpreadBlindfold difference is only 0.086. For LBF-10×10, AUC rises from 0.379 to 0.759; for RWARE-Tiny, from 13.945 to 21.190. History also improves

Table 2: Return AUC (mean ± sample SD, five seeds), divided by the 10M-step horizon. Compare values within each task. Method abbreviations follow Table 1; bold marks the best decentralized mean per row.

Input	Task	Joint	Local	MA^{2}E	RCD
Current observation	SpreadXY-2	-79.068 ± 0.762	-98.860 ± 6.113	-112.686 ± 1.174	$\mathbf{-94.382 \pm 2.971}$
	SpreadXY-4	-273.117 ± 7.406	$\mathbf{-297.424 \pm 6.848}$	-334.895 ± 5.364	-311.866 ± 9.994
	SpreadBlindfold	-156.759 ± 6.123	-162.717 ± 3.313	-186.644 ± 1.581	$\mathbf{-145.383 \pm 2.960}$
	LBF-8×8	0.755 ± 0.015	0.580 ± 0.055	0.321 ± 0.013	$\mathbf{0.637 \pm 0.073}$
	LBF-10×10	0.448 ± 0.017	0.294 ± 0.014	0.166 ± 0.012	$\mathbf{0.379 \pm 0.014}$
	RWARE-Tiny	23.918 ± 1.147	13.764 ± 1.296	5.754 ± 0.910	$\mathbf{13.945 \pm 1.339}$
	RWARE-Small	5.881 ± 0.945	$\mathbf{3.694 \pm 0.548}$	1.075 ± 0.169	3.487 ± 0.501
Five-step history	SpreadXY-2	-79.682 ± 0.413	-99.459 ± 3.924	-105.287 ± 5.333	$\mathbf{-93.358 \pm 4.076}$
	SpreadXY-4	-274.084 ± 8.848	-303.557 ± 4.196	-306.312 ± 5.633	$\mathbf{-300.466 \pm 4.031}$
	SpreadBlindfold	-155.428 ± 5.195	-162.704 ± 2.020	-170.984 ± 2.362	$\mathbf{-145.297 \pm 5.624}$
	LBF-8×8	0.906 ± 0.007	0.866 ± 0.020	0.506 ± 0.021	$\mathbf{0.891 \pm 0.012}$
	LBF-10×10	0.708 ± 0.022	0.580 ± 0.023	0.298 ± 0.008	$\mathbf{0.759 \pm 0.021}$
	RWARE-Tiny	26.236 ± 1.090	17.581 ± 2.177	10.593 ± 0.623	$\mathbf{21.190 \pm 1.319}$
	RWARE-Small	5.590 ± 1.293	5.790 ± 0.522	1.313 ± 0.296	$\mathbf{7.517 \pm 1.023}$

MA^{2}E in all seven tasks, Local in five, and Joint in four. These results distinguish the value of temporal information from the additional benefit of RCD within the same input setting. They also show why both metrics matter: a method can improve return during learning without substantially changing the final endpoint.

7.3 Uncertainty and Local Information

Uncertainty-aware fusion. We compare full RCD with **RCD-Mean**, which uses the same teacher-effect target but predicts only the mean effect, removing variance prediction and the reliability-aware gate. Full RCD additionally estimates uncertainty and uses it to control how strongly the predicted guidance influences the policy. Full RCD achieves higher mean return in five of six comparisons (Table 3). Without history, it improves over RCD-Mean by 6.18 on SpreadXY-2 and 6.10 on SpreadBlindfold. With history, the gains are 4.26 on SpreadXY-2, 2.31 on SpreadXY-4, and 2.00 on SpreadBlindfold. The only exception is SpreadXY-4 without history, where RCD-Mean achieves a slightly higher mean return, by 0.96. These results support using uncertainty prediction and gating together. Because RCD-Mean removes both components, this comparison does not show how much each component contributes individually. It also does not test whether the predicted uncertainty accurately reflects the observed prediction errors.

Table 3: Reliability ablation: final return (mean ± SD, five seeds). RCD-Mean removes variance prediction and reliability-aware gating. C: current observation; H: five-step history. Bold: higher mean in each pair.

Task	Full RCD (C)	RCD-Mean (C)	Full RCD (H)	RCD-Mean (H)
SpreadXY-2	$\mathbf{-87.74 \pm 0.50}$	-93.92 ± 8.46	$\mathbf{-85.86 \pm 3.85}$	-90.12 ± 5.34
SpreadXY-4	-286.97 ± 7.73	$\mathbf{-286.01 \pm 4.51}$	$\mathbf{-286.80 \pm 6.34}$	-289.11 ± 3.72
SpreadBlindfold	$\mathbf{-134.85 \pm 0.68}$	-140.95 ± 5.87	$\mathbf{-136.80 \pm 4.76}$	-138.80 ± 5.38

Effect of local visibility. Local visibility determines the area around an agent within which it can observe teammates and food. On LBF-10×10, we vary the sight range $s \in \{1, 2, 3\}$, corresponding to maximum observation windows of 3×3, 5×5, and 7×7, respectively. Larger windows provide more local information, while the task dynamics remain unchanged. RCD achieves higher mean return than Local at all three sight ranges, both with and without history (Table 4). The largest gains occur at $s = 2$: 0.1176 without history and 0.1186 with history. The gains are smaller at both $s = 1$ and $s = 3$, showing that RCD's advantage is greatest at intermediate visibility in this experiment. MA^{2}E exceeds Local only at $s = 1$ without history. Appendix F.3 provides further analysis.

7.4 Additional Experimental Results

History encoder. We compare MLP and GRU encoders across four methods on three MPE tasks. The MLP achieves higher mean return in $11/12$ comparisons; the exception is RCD on SpreadXY-4, where the GRU is higher by 1.02. For RCD, the MLP gains are 12.86 on SpreadXY-2 and 26.01 on SpreadBlindfold. These results support the MLP choice in our experiments. They compare the implemented encoders and their temporal context—a fixed history window for the MLP and a recurrent state for the GRU—rather than establishing a general advantage of feed-forward policies (Appendix G.1).

Table 4: Local-visibility study on LBF-10×10: final return (mean ± SD, five seeds). Bold: best decentralized mean per row. All values retain the supplied precision.

Input	Sight	Joint	Local	MA^{2}E	RCD
Current observation	$s = 1$	0.3623 ± 0.0758	0.1443 ± 0.0266	0.1626 ± 0.0106	$\mathbf{0.2144 \pm 0.0352}$
	$s = 2$	0.5907 ± 0.0286	0.3769 ± 0.0249	0.2364 ± 0.0075	$\mathbf{0.4945 \pm 0.0097}$
	$s = 3$	0.6842 ± 0.0208	0.5329 ± 0.0181	0.3572 ± 0.0315	$\mathbf{0.6057 \pm 0.0224}$
Five-step history	$s = 1$	0.6911 ± 0.0138	0.4980 ± 0.0364	0.3465 ± 0.0906	$\mathbf{0.5556 \pm 0.0269}$
	$s = 2$	0.8374 ± 0.0260	0.7138 ± 0.0303	0.4558 ± 0.0225	$\mathbf{0.8324 \pm 0.0175}$
	$s = 3$	0.8521 ± 0.0107	0.8332 ± 0.0174	0.5970 ± 0.0364	$\mathbf{0.8793 \pm 0.0182}$

Capacity and distillation controls. We compare RCD with ExtraMLP (additional actor capacity without teacher supervision), Policy-KD (teacher action-distribution matching), and Feature-KD (teacher representation matching), using the same MAPPO backbone, training budget, and five matched seeds. RCD achieves the highest mean return on SpreadXY-2 and SpreadBlindfold with and without history. Without history, its gains over Local are 4.65 and 16.29, respectively. ExtraMLP improves over Local only on SpreadXY-4 with history; Policy-KD helps on SpreadBlindfold but reduces return on both SpreadXY tasks; Feature-KD provides no consistent gain. Thus, these controls do not consistently reproduce RCD's improvements. However, Local leads on SpreadXY-4 without history and ExtraMLP leads with history, demonstrating that the benefits are task dependent. The comparisons assess RCD's combined design rather than isolating each component (Appendix G.2).

Coordination-effect target. RCD predicts the residual coordination effect, whereas RCD-FullLogit predicts centered privileged teacher logits with the remaining architecture fixed. RCD performs better on SpreadXY-2 and SpreadBlindfold in both input settings, with gains up to 10.63 points, while FullLogit performs better on SpreadXY-4. This shows that the residual target contributes beyond the shared architecture, although its advantage is task-dependent (Appendix G.3).

Explicit reliability modulation. ConstantReliability keeps the same predictor and learned gate but fixes the reliability factor to one. Explicit reliability gives clear gains on SpreadXY-2 with history and SpreadBlindfold without history, but is neutral or worse on other settings, especially SpreadXY-4. This suggests that reliability-aware attenuation is useful when prediction uncertainty provides information beyond what the learned gate captures (Appendix G.4).

Comparison with ADVISOR and MAGPO. The comparison shows that centralized-to-decentralized transfer is strongly task dependent. RCD performs best among the transfer methods on SpreadBlindfold and improves over Local on SpreadXY-2 and Foraging-8×8, while remaining competitive with MAGPO and ADVISOR on several other tasks. However, it does not uniformly dominate: on SpreadXY-4, none of the transfer approaches improves over Local, and on Foraging-10×10, ADVISOR achieves the best decentralized result. These differences suggest that the effectiveness of privileged teacher guidance depends on how much of the teacher's useful coordination information can be inferred from the student's execution-time observations.

The RWARE results further highlight this limitation. Although Joint substantially outperforms all decentralized methods, RCD, ADVISOR, and MAGPO remain close to or below the Local baseline, indicating that much of the centralized advantage may rely on information that is not recoverable from local observations. Overall, the results support the motivation of RCD: rather than attempting to reproduce the centralized teacher unconditionally, RCD exploits only the coordination effect that can be predicted from information available to the decentralized policy. Its strongest gains therefore arise when privileged information is missing at execution time but its decision-relevant effect remains locally recoverable.(More details can be found in Appendix G.5)

8 Conclusion

We proposed Recoverable Coordination Distillation (RCD) as an alternative to trajectory reconstruction at deployment. RCD predicts the locally recoverable effect of privileged information on a teacher's action preferences and estimates its uncertainty. A learned gate regulates how this guidance influences a decentralized policy trained for team return. Our analysis characterizes the local information limit and bounds return sensitivity to prediction errors. Across seven tasks, RCD achieves higher mean final return than our MAPPO-based MA^{2}E baseline in all 14 task–input comparisons. The mean-only ablation further supports the combined use of uncertainty prediction and gated fusion, without establishing their individual contributions or uncertainty calibration. These results support predicting decision-relevant coordination effects as a useful alternative to reconstructing missing teammate trajectories.

9 AI USE STATEMENT

In this work, we used generative AI tools to aid or polish writing, for formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs. We have not used generative AI tools for other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work.

REFERENCES

Filippos Christianos, Lukas Schäfer, and Stefano Albrecht. Shared experience actor-critic for multi-agent reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 10707–10717. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/7967cc8e3ab559e68cc944c44b1cf3e8-Paper.pdf.

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 N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *AAAI Conference on Artificial Intelligence*, 2018.

Matthew Hausknecht and Peter Stone. Deep recurrent q-learning for partially observable MDPs. In *AAAI Fall Symposium on Sequential Decision Making for Intelligent Agents*, 2015. URL https://arxiv.org/abs/1507.06527.

Sehyeok Kang, Yongsik Lee, Gahee Kim, Song Chong, and Se-Young Yun. MA2E: Addressing partial observability in multi-agent reinforcement learning with masked auto-encoder. In *International Conference on Learning Representations*, 2025.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html.

Karol Kurach, Anton Raichuk, Piotr Stańczyk, Michał Zając, 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.

Yueheng Li, Guangming Xie, and Zongqing Lu. Multi-agent guided policy optimization. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/html/2507.18059v2.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, 2017.

Shayegan Omidshafiei, Jason Pazis, Christopher Amato, Jonathan P. How, and John Vian. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2681–2690. PMLR, 2017. URL https://proceedings.mlr.press/v70/omidshafiei17a.html.

Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano Albrecht. Benchmarking Multi-Agent Deep Reinforcement Learning Algorithms in Cooperative Tasks. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/a8baa56554f96369ab93e4f3bb068c22-Paper-round1.pdf.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International Conference on Machine Learning*, 2018.

Mikayel Samvelyan, Tabish Rashid, Christian Schroeder De Witt, Gregory Farquhar, Nantas Nardelli, Tim GJ Rudner, Chia-Man Hung, Philip HS Torr, Jakob Foerster, and Shimon Whiteson. The starcraft multi-agent challenge. *arXiv preprint arXiv:1902.04043*, 2019.

Pedro P. Santos, Diogo S. Carvalho, Miguel Vasco, Alberto Sardinha, Pedro A. Santos, Ana Paiva, and Francisco S. Melo. Centralized training with hybrid execution in multi-agent reinforcement learning via predictive observation imputation. *Artificial Intelligence*, 348:104404, 2025.

Luca Weihs, Unnat Jain, Iou-Jen Liu, Jordi Salvador, Svetlana Lazebnik, Aniruddha Kembhavi, and Alexander Schwing. Bridging the imitation gap by adaptive insubordination. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2007.12173.

Chao Yu, Sandeep Velu, Eugene Vinitsky, Yu Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Advances in Neural Information Processing Systems*, 2022.

Jian Zhao, Xunhan Hu, Mingyu Yang, Wengang Zhou, Jiangcheng Zhu, and Houqiang Li. CTDS: Centralized teacher with decentralized student for multi-agent reinforcement learning. *arXiv preprint arXiv:2203.08412*, 2022. URL https://arxiv.org/abs/2203.08412.

A Local Information and Recoverability

A.1 Proof of Proposition 1

Proof. Fix $X_i = x_i$. Since $R_i = h(X_i)$ is deterministic, $R_i = h(x_i)$. Define

$$\widetilde{\pi}_i(a_i \mid x_i) = \pi_i(a_i \mid x_i, h(x_i)).$$

Then, for every x_i,

$$\pi_i(a_i \mid X_i = x_i, R_i = h(x_i)) = \widetilde{\pi}_i(a_i \mid x_i),$$

so conditioning on R_i does not change the set of action distributions representable from X_i. Equivalently, $\sigma(X_i, R_i) = \sigma(X_i)$ because R_i is measurable with respect to X_i. Hence, a deterministic local prediction introduces no additional execution information. □

Proposition 1 concerns information rather than learning efficiency or representation quality. Privileged supervision may help h extract decision-relevant structure from X_i that a finite policy would otherwise be difficult to learn. However, if two privileged situations share the same local input but induce different teacher effects,

$$C_i^T(x_i, u_i^{(1)}) \neq C_i^T(x_i, u_i^{(2)}), \tag{18}$$

any deterministic local predictor must produce the same output for both and therefore cannot recover both realized effects exactly. If the two privileged situations instead induce the same effect, that effect can remain recoverable even though the underlying privileged states are indistinguishable. This motivates predicting the teacher's decision effect rather than reconstructing U_i itself.

A.2 Squared-Error Decomposition

For a fixed agent, let $C = C_i^T$, $X = X_i$, $\mu(X) = \mathbb{E}[C \mid X]$, and $\varepsilon = C - \mu(X)$. Assume C and $h(X)$ are square-integrable. Then

$$\begin{aligned}\mathbb{E}\|C - h(X)\|_2^2 = \mathbb{E}\|\varepsilon\|_2^2 + \mathbb{E}\|\mu(X) - h(X)\|_2^2 \\ + 2\mathbb{E}\left[\varepsilon^\top(\mu(X) - h(X))\right].\end{aligned}$$

The cross term is zero because

$$\begin{aligned}\mathbb{E}\left[\varepsilon^\top(\mu(X) - h(X))\right] &= \mathbb{E}\left[\mathbb{E}[\varepsilon \mid X]^\top(\mu(X) - h(X))\right] \\ &= 0,\end{aligned}$$

where $\mathbb{E}[\varepsilon \mid X] = 0$ by definition. Hence,

$$\mathbb{E}\|C - h(X)\|_2^2 = \mathbb{E}\|\varepsilon\|_2^2 + \mathbb{E}\|\mu(X) - h(X)\|_2^2,$$

which proves Equation (3).

Therefore, $\mu(X) = \mathbb{E}[C \mid X]$ minimizes the expected squared prediction error, uniquely up to almost-sure equality. The minimum achievable error is

$$\mathbb{E}\|\varepsilon\|_2^2 = \mathbb{E}[\mathrm{tr}(\mathrm{Var}(C \mid X))],$$

which represents the prediction uncertainty that remains unresolved from X.

This decomposition is a standard conditional-expectation result applied to the teacher coordination effect. It characterizes what can be predicted from local information, but does not imply that the conditional mean is itself an optimal control signal. How the predicted effect should influence action selection is learned separately by the reinforcement-learning pathway.

A.3 Effect of Removing Local Information

Proposition 2 (Coarsening cannot improve optimal prediction)**.** *Let C be a fixed square-integrable teacher target and let $X^{(w)} = g(X^{(s)})$ be a deterministic coarsening of a more informative input $X^{(s)}$. Under the same joint distribution,*

$$\mathbb{E}\|C - \mathbb{E}[C \mid X^{(w)}]\|_2^2 = \mathbb{E}\|C - \mathbb{E}[C \mid X^{(s)}]\|_2^2 + \mathbb{E}\|\mathbb{E}[C \mid X^{(s)}] - \mathbb{E}[C \mid X^{(w)}]\|_2^2. \tag{19}$$

Consequently, coarsening the local input cannot reduce the population-optimal squared prediction error.

Proof. Let

$$\mu_s = \mathbb{E}[C \mid X^{(s)}], \qquad \mu_w = \mathbb{E}[C \mid X^{(w)}].$$

Because $X^{(w)} = g(X^{(s)})$, μ_w is also measurable with respect to $X^{(s)}$. We can therefore write

$$C - \mu_w = (C - \mu_s) + (\mu_s - \mu_w).$$

Expanding the squared norm and taking expectations gives

$$\mathbb{E}\|C - \mu_w\|_2^2 = \mathbb{E}\|C - \mu_s\|_2^2 + \mathbb{E}\|\mu_s - \mu_w\|_2^2 + 2\mathbb{E}\big[(C - \mu_s)^\top(\mu_s - \mu_w)\big] .$$

The cross term vanishes since

$$\mathbb{E}[C - \mu_s \mid X^{(s)}] = 0$$

and $\mu_s - \mu_w$ is a function of $X^{(s)}$. This yields Equation (19). Since the final term is nonnegative, the coarser input cannot achieve a smaller population-optimal prediction error. □

This result concerns a fixed teacher target and a fixed sampling distribution. If changing the observation channel also changes the teacher, restricted evaluation, or target definition, the comparison in Equation (19) no longer applies directly. Likewise, the proposition does not imply an ordering of finite-sample learning performance or policy return.

B Teacher-Target Construction

Centering and action preferences. For a logit vector $v = (2, 4, 6)$, $\mathrm{ctr}(v) = (-2, 0, 2)$. Adding a common constant b to all logits leaves the softmax unchanged:

$$\frac{\exp(\ell_k + b)}{\sum_j \exp(\ell_j + b)} = \frac{\exp(\ell_k)}{\sum_j \exp(\ell_j)}.$$

Centering therefore removes the action-independent degree of freedom and retains only relative action preferences. For strictly positive action probabilities, the teacher effect can equivalently be written as

$$C_i^T = \mathrm{ctr}\big(\log \pi_i^T(\cdot \mid X_i, U_i) - \log \pi_i^T(\cdot \mid X_i)\big) ,$$

because the different softmax normalizers vanish after centering.

Both teacher evaluations must use the same action coordinates and logit temperature. If action masking is used, the same execution-available mask should be applied in both passes; centering, prediction loss, and reliability computation should then be restricted to the common legal actions.

Restricted information path. The privileged and restricted evaluations use the same frozen teacher parameters and the same underlying transition. The restricted evaluation must depend only on X_i. It is therefore insufficient to remove U_i only at the final layer if privileged information has already entered through earlier cross-agent computation. For a recurrent teacher, the restricted hidden state must likewise be constructed solely from the restricted history rather than copied from a privileged rollout. These conditions ensure that $\ell_i^T(X_i)$ is genuinely free of privileged information.

Interpretation and limitations. The target C_i^T measures how the specified teacher's relative action preferences change when privileged information is removed. It should not be interpreted as a causal effect on environmental reward, and the restricted evaluation is not assumed to represent an optimal decentralized policy.

The target also depends on the intervention used to remove privileged information. If masking creates inputs far from those encountered during teacher training, part of the observed logit difference may reflect the teacher's response to this intervention rather than useful coordination information. Consequently, the masking procedure, recurrent-state handling, and teacher quality all affect the meaning of C_i^T. Freezing the teacher makes the target function stable during student training, but does not by itself guarantee that the target is informative.

C Conditional Gaussian Prediction

Action-wise Gaussian objective. For agent i, the coordination effect $C_i^T \in \mathbb{R}^{d_i}$ has one entry for each action in $\mathcal{A}_i$. Specifically, $C_{i,k}^T$ represents the change in the teacher's relative preference for action k caused by privileged information. We model each action dimension with a scalar Gaussian loss.

For a scalar target C, predicted mean m, and variance $v > 0$, the Gaussian negative log-likelihood is

$$\ell(m, v; C) = \frac{1}{2}\left[\frac{(C - m)^2}{v} + \log v\right] .$$

For a fixed predictor input z, define

$$m^*(z) = \mathbb{E}[C \mid z], \qquad v^*(z) = \mathrm{Var}(C \mid z).$$

Then

$$\mathbb{E}[\ell(m, v; C) \mid z] = \frac{1}{2}\left[\frac{v^*(z) + (m - m^*(z))^2}{v} + \log v\right].$$

For any $v > 0$, this objective is minimized over m by $m = m^*(z)$. With this mean, the unconstrained optimum is $v = v^*(z)$ whenever $v^*(z) > 0$. Under $v \in [\sigma^2_{\min}, \sigma^2_{\max}]$, the optimum is the clipped conditional variance. If $v^*(z) = 0$, the lower variance bound gives a finite optimum. More generally, when the predicted mean is imperfect, the optimal variance reflects the conditional residual second moment $\mathbb{E}[(C - m)^2 \mid z]$, which includes both intrinsic variation and squared mean error.

These are population-optimal statements. Finite data, limited predictor capacity, and imperfect optimization may prevent the learned model from attaining these conditional moments.

Conditioning on the student representation. The predictor receives $z_i = f_S(X_i)$ rather than the full local input X_i. For a fixed encoder,

$$\mathbb{E}[C_i^T \mid z_i] = \mathbb{E}[\mu_i^*(X_i) \mid z_i],$$

and the law of total variance gives

$$\mathrm{Var}(C_i^T \mid z_i) = \mathbb{E}[\Sigma_i^*(X_i) \mid z_i] + \mathrm{Var}(\mu_i^*(X_i) \mid z_i).$$

The first term reflects ambiguity that remains even with the full local information X_i. The second captures additional ambiguity introduced when distinct local inputs are mapped to the same representation z_i. Hence, the variance predicted from z_i need not equal the irreducible uncertainty conditioned on the full X_i.

During training, the encoder changes through reinforcement learning, so the predictor is fitted to a changing representation even though the frozen teacher defines a fixed target function.

Diagonal approximation and reliability. Because centering imposes $\mathbf{1}^\top C_i^T = 0$, the action dimensions of C_i^T are statistically dependent and the exact conditional covariance is singular. We therefore use a diagonal Gaussian as a tractable approximation to the action-wise conditional means and variances. It does not model the full joint covariance or multimodal uncertainty.

The reliability function in Equation (7) can equivalently be written as

$$F_i(v) = \left(\prod_{k=1}^{d_i}(1 + v_k)\right)^{-1/d_i}.$$

It decreases monotonically as any action-wise variance increases and aggregates uncertainty using the geometric mean of $1 + v_k$, avoiding direct scaling with the number of actions. Its magnitude nevertheless depends on the scale of the teacher logits.

Low predicted variance indicates that the coordination effect is predicted consistently, but does not by itself imply that the resulting guidance is useful for control. The learned fusion gate therefore determines whether, and how strongly, the predicted guidance should influence the policy.

D Training Sequence and Gradient Routing

Stage 1: teacher learning. Train the privileged teacher with the MARL objective using both X_i and U_i, then freeze all teacher parameters. During student training, the same frozen teacher is evaluated with and without privileged information to construct the coordination-effect target. No student gradient is propagated to the teacher.

Stage 2: student learning. Each student training iteration proceeds as follows:

1. Collect transitions using the decentralized student policy with only execution-available inputs X_i. Privileged information U_i is retained by the training system for teacher-target construction and centralized critic learning.
2. Evaluate the frozen teacher on the same transition with privileged information enabled and removed, and compute C_i^T using Equation (1).
3. Encode the local input as $z_i = f_S(X_i)$ and evaluate $q_\phi(\mathrm{sg}(z_i))$. Optimize the predictor using the Gaussian objective in Equation (6) to estimate the action-wise mean and clipped variance of C_i^T.

4. Detach the predicted mean and reliability before passing them to the projection and fusion modules. Optimize the actor, centralized critic, and entropy objective using the chosen MARL algorithm.

Prediction and reinforcement-learning losses may be evaluated on the same minibatch, provided their gradient paths remain separated. The method does not require a separate predictor warm-up stage or a particular optimizer.

Gradient routing and parameter updates. The stop-gradient at the predictor input prevents $\mathcal{L}_{\mathrm{RCD}}$ from updating the student encoder. Thus, the prediction loss updates only the predictor parameters ϕ. Conversely, the predicted mean and reliability are detached before entering the policy path, so $\mathcal{L}_{\mathrm{RL}}$ updates the encoder, projection, gate, policy head, and critic, but not ϕ. Both quantities must be detached: detaching only the mean would still allow policy gradients to reach the predictor through the variance-to-reliability path, violating Equation (12).

The coefficient λ_C scales the predictor update. Because the two objectives act on disjoint parameter blocks, it does not directly trade off prediction and policy gradients on the same parameters.

Dimensions and initialization. The predicted mean $\widehat{\mu}_i$ has d_i action dimensions, whereas the student feature z_i has dimension d_z. A linear projection P_η therefore maps $\mathbb{R}^{d_i}$ to $\mathbb{R}^{d_z}$, while the feature-wise gate operates in $\mathbb{R}^{d_z}$. The scalar reliability $\widehat{\xi}_i$ is broadcast across feature dimensions. We initialize the gate with a negative bias so that its sigmoid output is initially small, keeping the student close to the local pathway $\mathrm{LN}(z_i)$ while the predictor is still inaccurate. The gate subsequently learns from return signals how much predicted guidance to use.

Deployment. At execution, retain only the student encoder, local predictor, projection, fusion gate, and policy head. The teacher, privileged information, teacher targets, and centralized critic are discarded. All remaining computations depend only on X_i, including any communication already permitted by the execution protocol. No realized teacher coordination effect is required at deployment.

E Proof and Scope of the Return Bound

E.1 Histories and Discounted Occupancy

Let H_t denote the complete joint action–observation history, together with any additional state required to describe the controlled process. Treating histories as states allows us to use a Markov value-function argument without requiring the local policy input X_i itself to be Markov. Let $X(H_t)$ denote the corresponding joint student input and define the normalized discounted occupancy measure

$$d^H_{\widehat{\pi}}(h) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr_{\widehat{\pi}}(H_t = h).$$

For general state spaces, $d^H_{\widehat{\pi}}$ is understood as a probability measure. The distribution $d_{\widehat{\pi}}$ used in the main text is its pushforward through $X(H_t)$. We assume that the conditional-moment functions and both comparison policies are well defined on the histories relevant to this occupancy measure.

E.2 Proof of Theorem 1

Define

$$\delta(X) = D_{\mathrm{TV}}\big(\widehat{\pi}(\cdot \mid X), \pi_C(\cdot \mid X)\big).$$

By the Lipschitz assumption in Equation (15),

$$\delta(X) \le L_\mu \sum_i \|\widehat{\mu}_i(X_i) - \mu^*_i(X_i)\|_2 + L_\xi \sum_i |\widehat{\xi}_i(X_i) - \xi^*_i(X_i)|. \tag{20}$$

Let $Q^{\pi_C}(h,a)$ and $V^{\pi_C}(h)$ denote the action-value and value functions of the reference policy π_C. Since $|r_t| \le R_{\max}$,

$$|Q^{\pi_C}(h,a)| \le \frac{R_{\max}}{1-\gamma}.$$

Define the one-step discrepancy

$$b(h) = \mathbb{E}_{a\sim\widehat{\pi}(\cdot|X(h))} Q^{\pi_C}(h,a) - V^{\pi_C}(h).$$

Using

$$V^{\pi_C}(h) = \mathbb{E}_{a\sim\pi_C(\cdot|X(h))} Q^{\pi_C}(h,a),$$

together with the standard total-variation bound for expectations of a bounded function, gives

$$|b(h)| \leq \frac{2R_{\max}}{1-\gamma}\,\delta(X(h)). \tag{21}$$

The Bellman equation further gives

$$b(H_t) = \mathbb{E}_{\widehat{\pi}}\left[r_t + \gamma V^{\pi_C}(H_{t+1}) - V^{\pi_C}(H_t) \mid H_t\right].$$

Multiplying by γ^t, taking expectations under $\widehat{\pi}$, and summing over t causes the value terms to telescope. Because V^{π_C} is bounded and $\gamma < 1$, the limiting term vanishes. Under the common initial distribution,

$$J(\widehat{\pi}) - J(\pi_C) = \frac{1}{1-\gamma}\mathbb{E}_{h\sim d^H_{\widehat{\pi}}}[b(h)].$$

Therefore,

$$|J(\widehat{\pi}) - J(\pi_C)| \leq \frac{2R_{\max}}{(1-\gamma)^2}\mathbb{E}_{X\sim d_{\widehat{\pi}}}[\delta(X)].$$

Substituting Equation (20) and the definitions of e_μ and e_ξ in Equation (16) yields

$$|J(\widehat{\pi}) - J(\pi_C)| \leq \frac{2R_{\max}}{(1-\gamma)^2}\left(L_\mu e_\mu + L_\xi e_\xi\right),$$

which proves Equation (17). □

E.3 Scope of the Bound

The reference policy π_C uses the same fixed fusion and policy parameters as the learned student, replacing only the predicted effect and reliability with their exact conditional counterparts. It is therefore a comparison policy rather than a separately trained model, an optimal decentralized policy, or a privileged oracle.

The theorem applies to any fixed student satisfying its assumptions. It does not imply that training drives e_μ or e_ξ to zero. In particular, a small prediction loss on the training distribution does not guarantee small errors under the discounted occupancy of $\widehat{\pi}$ without additional coverage or generalization assumptions.

The reference reliability ξ_i^* is defined from the population conditional variance in Equation (13). If variance clipping prevents the learned predictor from representing this quantity exactly, the resulting discrepancy is included in e_ξ. One could instead define a reference using clipped population variances; the same argument would apply with the corresponding definition of the reliability error.

Finally, Equation (3) and Theorem 1 address different sources of error. The former characterizes the information-dependent prediction floor caused by privileged distinctions that cannot be resolved from local information. The latter measures the additional return deviation caused by imperfectly estimating the recoverable effect and its reliability within a fixed policy network. It does not bound the full teacher–student return gap or the error caused by suboptimal use of otherwise accurate guidance. Even with exact conditional moments, the decentralized student cannot generally reproduce sample-specific teacher decisions that depend on the realized privileged information U_i. Empirical evaluation is therefore still required to determine whether the predicted guidance improves team return.

F Experiment Details

The experiments were conducted on a workstation equipped with a 32-Core AMD Ryzen Threadripper PRO 7975WX processor, 128 GB of RAM, and an NVIDIA RTX 4080 GPU (16 GB VRAM).

F.1 Choice of Evaluation Environments

Our environment suite is designed to isolate the role of coordination-relevant information under decentralized execution. In particular, we seek environments in which the difference between local and joint information can be identified and interpreted without being dominated by unrelated sources of task complexity.

This consideration makes some commonly used MARL benchmarks less suitable as diagnostic environments for our specific research question. In the original StarCraft Multi-Agent Challenge (SMAC) (Samvelyan et al., 2019), for example, Ellis et al. (2023) show that open-loop policies conditioned only on the timestep can achieve non-trivial performance in several scenarios, indicating that nominal partial observability does not necessarily require agents to recover missing information through closed-loop decision making. SMACv2 (Ellis et al., 2023) substantially improves upon this limitation through

Table 5: Evaluation environments and their information structures. The suite covers controlled information removal, spatial partial observability, and distributed multi-agent sensing.

Environment	Missing information	Information structure	Primary evaluation role
SpreadXY-2	Complementary x/y motion states	Controlled complementary split	Basic recoverability
SpreadXY-4	Complementary x/y motion states	Controlled complementary split	Larger-team coordination
SpreadBlindfold	Teammate motion states	Controlled state removal	Missing-teammate recovery
LBF-3p	Distant agents and food	Spatial partial observability	Context-dependent cooperation
LBF-2p-coop	Distant teammate and food	Spatial partial observability	Required cooperation
RWARE-tiny-4ag	Distant robots and shelves	Distributed local sensing	Multi-robot coordination
RWARE-small-4ag	Distant robots and shelves	Stronger spatial separation	Coordination under reduced coverage

procedural generation, randomized starting configurations and unit compositions, and stronger partial observability. However, these properties also introduce several simultaneous sources of variation, including heterogeneous unit types, enemy states, visibility and attack ranges, and combat dynamics. Consequently, a performance difference between local and joint policies can be difficult to attribute specifically to the recoverability of coordination-relevant information.

Google Research Football (GRF) (Kurach et al., 2020) presents a different difficulty. Its commonly used Football Academy tasks are built around predefined football situations in which the initial locations and roles of the ball, controlled players, and opponents are specified by the scenario. Such structured initialization is useful for isolating particular football skills, but also introduces strong task-specific regularities. Moving to full-game GRF reduces this concern while introducing substantially greater complexity from opponent behavior, ball possession, passing, long-horizon tactics, and large-scale training.

These properties do not make SMAC, SMACv2, or GRF unsuitable as general MARL benchmarks. Rather, they make it more difficult to identify whether an observed performance difference specifically arises from recovering coordination-relevant information that is unavailable locally. We therefore focus on MPE, LBF, and RWARE, for which the source and coordination role of missing information can be characterized more directly.

Table 5 summarizes the resulting experimental axis. The suite progresses from explicitly controlled information partitions in MPE, through spatially induced partial observability in LBF, to distributed sensing and multi-robot coordination in RWARE.

F.2 Calculation of Learning-Curve AUC

We compute AUC separately for each seed by trapezoidal integration over the observed evaluation checkpoints, divided by the 10M-step training horizon. If r_j is the evaluation return at checkpoint t_j, then

$$\mathrm{AUC} = \frac{1}{T}\sum_{j=1}^{m-1}(t_{j+1}-t_j)\frac{r_j+r_{j+1}}{2}, \qquad T = 10^7.$$

This uses the observed checkpoints without assuming unreported endpoint values. AUC measures return accumulated during learning, whereas the primary metric measures performance near the end. Returns are not normalized across tasks, so AUC values should be compared only within the same environment.

The full AUC results and their comparison with final return appear in Section 7.2 and Table 2 of the main paper. The table reports the mean and sample standard deviation of the five seed-level AUC values. Each seed is integrated separately before this aggregation; no significance test is inferred from these summary statistics.

F.3 Effect of Local Observation Range

We vary the sight range $s \in \{1, 2, 3\}$ in LBF-10×10, keeping the grid, three agents, three food items, and task dynamics fixed. The configurations are `1s-10x10-3p-3f`, `2s-10x10-3p-3f`, and `3s-10x10-3p-3f`. For agent i at $p_i = (r_i, c_i)$, the visible region is

$$\mathcal{V}_i^{(s)} = \{(r, c) : |r - r_i| \le s,\ |c - c_i| \le s\} \cap \mathcal{G},$$

where $\mathcal{G}$ is the grid. The maximum observation windows are 3×3, 5×5, and 7×7, truncated at map boundaries. Thus, this study changes local visibility while preserving the task.

Table 4 in the main paper reports the full results. Mean return increases with sight range for every reported method. RCD exceeds its corresponding Local baseline at all three ranges, with and without history. Its absolute gains are 0.0701, 0.1176, and 0.0728 without history, and 0.0576, 0.1186, and 0.0461 with history. The largest gain in both settings occurs at the intermediate range $s = 2$.

At $s = 2$, the fraction of the observed Joint–Local mean-return gap closed by RCD is

$$\frac{J_{\mathrm{RCD}} - J_{\mathrm{Local}}}{J_{\mathrm{Joint}} - J_{\mathrm{Local}}},$$

which is approximately 55.0% without history and 96.0% with history, using Table 4. Here the J values denote the reported sample mean returns. This is a descriptive return comparison, not a measurement of the conditional-mean recoverability defined in the theory.

MA^{2}E exceeds Local only at $s = 1$ without history; it is below Local at $s = 2, 3$ and at every range with history. The results favor RCD over this reconstruction baseline across the tested visibility settings. They do not identify which part of the gain comes from the prediction target, teacher supervision, or differences in model capacity.

F.4 Details of Selected Environments

Multi-Agent Particle Environment (MPE). In the three MPE-based tasks, agents move in a two-dimensional world and coordinate to cover landmarks while avoiding collisions. Each agent selects one of five discrete actions,

$$\mathcal{A}_{\mathrm{MPE}} = \{\texttt{no-op}, \texttt{left}, \texttt{right}, \texttt{down}, \texttt{up}\}.$$

The environments differ in team size and, more importantly, in the structure of the information available to each agent.

SpreadXY-2. SpreadXY-2 contains two agents and two landmarks. The two agents receive complementary local observations: one observes the x-axis positions and velocities of both agents, while the other observes the corresponding y-axis quantities. Both agents additionally observe the positions of all landmarks. Thus, neither local observation contains the complete two-dimensional motion state of the team, whereas combining both observations provides access to the complementary spatial information. This environment provides a minimal setting in which coordination-relevant information is systematically split across agents.

SpreadXY-4. SpreadXY-4 extends this structure to four agents and four landmarks, organized as two teams of two agents. Within each pair, one agent observes the x-axis positions and velocities of both agents in the pair, while the other observes the corresponding y-axis quantities. All agents observe the landmark positions. This environment increases the team size and landmark-assignment complexity while retaining the complementary axis-wise observation structure of SpreadXY-2. It therefore tests whether the same recovery mechanism remains effective as the joint coordination problem becomes larger.

SpreadBlindfold. SpreadBlindfold contains three agents and three landmarks. Each agent observes its own position and velocity together with the landmark positions, but does not observe the positions or velocities of the other agents. Consequently, local decision making has access to the agent's own complete motion state but lacks the current motion states of its teammates. Unlike the axis-wise decomposition in SpreadXY, the missing information here corresponds directly to teammate states, providing a qualitatively different form of decentralized uncertainty.

Level-Based Foraging (LBF). We further evaluate on Level-Based Foraging (LBF) (Papoudakis et al., 2021), a grid-world domain in which agents with heterogeneous levels navigate toward food items and decide when and with whom to forage. A food item can be collected when the combined level of the agents simultaneously attempting to load it is sufficient for the food level. Each agent selects from a discrete action space consisting of movement, loading, and no-op actions.

We use partial-observation variants with sight range two, denoted by the `2s` component of the environment name. Consequently, each agent observes only a limited neighborhood and may lack current information about distant teammates and food items. Unlike the deliberately constructed information partitions in the MPE tasks, the missing information in LBF arises from spatial separation and changes dynamically as agents move.

10x10-3p-3f contains three agents and three food items on a 10×10 grid. Because cooperation is not forced for every food item, agents may need to decide whether locally feasible individual collection or coordinated collection with other agents is preferable. This setting therefore evaluates coordination recovery when the relevance of other agents' information is context dependent.

8x8-2p-2f-2s-c contains two agents and two food items on an 8×8 grid and uses the forced-cooperation variant. Successful food collection consequently depends directly on coordinated participation. The two LBF configurations therefore contrast context-dependent cooperation with a setting in which cooperation is an explicit requirement of the task.

Multi-Robot Warehouse (RWARE). Finally, we evaluate on the Multi-Robot Warehouse environment (RWARE) (Papoudakis et al., 2021; Christianos et al., 2020). RWARE models a warehouse in which multiple robots must locate requested shelves, transport them to delivery workstations, and subsequently return the shelves to available storage locations. Rewards

are sparse, since successful behavior requires a sequence of navigation, pickup, transport, and delivery decisions before positive feedback is obtained.

RWARE is partially observable by construction. Under the standard local observation setting, each robot observes only a small spatial neighborhood centered on itself, including nearby robots and shelves. Consequently, information about robots and requested shelves outside an agent's sensing region is unavailable to its local policy but may simultaneously be available in another agent's observation.

This information structure is particularly relevant to our setting. Unlike SpreadXY, RWARE does not explicitly partition a known global state between agents. Instead, complementary information emerges dynamically from the agents' spatial locations. The collection of all agents' observations can therefore contain coordination-relevant information that is absent from an individual observation, while still not necessarily constituting a complete global description of the warehouse. RWARE thus tests whether coordination effects can be recovered when information asymmetry arises naturally through distributed sensing.

Each robot selects from five discrete actions corresponding to no-op, moving forward, rotating left, rotating right, and loading or unloading a shelf,

$$\mathcal{A}_{\text{RWARE}} = \{\texttt{no-op}, \texttt{forward}, \texttt{turn-left}, \texttt{turn-right}, \texttt{load/unload}\}.$$

Coordination is required not only for allocating requested shelves across robots but also for navigating shared corridors. Because robots occupy the same physical space, the actions of one robot can affect the feasible motion of nearby teammates, making other agents' locations and behavior directly relevant to local decision making.

We evaluate two RWARE configurations. **tiny-4ag** contains four robots in a 10×11 warehouse. The relatively compact layout produces frequent interaction among robots operating with only local observations, providing a setting in which distributed information and multi-robot coordination are both important.

small-4ag also contains four robots but uses a larger 10×20 warehouse. Because the number of agents is held fixed while the spatial extent of the environment increases, each local observation covers a smaller fraction of the overall task space. Relative to **tiny-4ag**, this configuration therefore increases spatial separation without changing team size, providing a controlled test of coordination recovery under reduced local information coverage.

Together, the two RWARE configurations isolate the effect of increasing spatial extent while keeping the number of agents fixed:

$$\texttt{tiny-4ag} \longrightarrow \texttt{small-4ag}.$$

This comparison complements the MPE and LBF tasks by evaluating whether recoverable coordination effects remain useful when task-relevant information is distributed dynamically across multiple locally observing agents.

F.5 Compared Methods and MAPPO Implementation Details

All methods use a common MAPPO-style actor–critic backbone (Yu et al., 2022). The centralized critic is used only during training and discarded at execution. The comparisons therefore differ primarily in the information available to the actor and in how that information is used.

We distinguish between *privileged* policies, whose actors may access cross-agent information, and *decentralized* policies, whose actors must satisfy the execution-time information constraint. Joint and Joint-History serve as privileged references and RCD teachers, whereas Local, RCD, and MA^{2}E variants are decentralized. Actor parameters are shared across agents within each method, with each forward pass corresponding to a focal agent. The centralized critic is also shared across agents and receives the complete team observation `share_obs` during training.

F.5.1 Information Contracts and Teacher Intervention

History-free information contract. For the history-free comparison, every decentralized actor receives only the current observation of the focal agent:

$$X_{i,t} = o_t^i, \qquad U_{i,t} = o_t^{-i}, \tag{22}$$

where $U_{i,t}$ denotes cross-agent information unavailable to agent i at decentralized execution. No previous action, explicit observation history, recurrent state, or privileged teammate feature is included in $X_{i,t}$. The centralized critic may still use the complete team observation during training; this does not alter the actor's execution-time information.

Shared-Mean Joint teacher. The history-free Joint actor uses a shared per-agent observation encoder. Given

$$O_t = (o_t^1, \ldots, o_t^N), \tag{23}$$

it computes

$$h_{j,t}^T = E_T(o_t^j), \qquad j = 1, \dots, N, \tag{24}$$

using the same encoder E_T for every agent. For focal agent i, the representations of the other agents are aggregated by the leave-one-out mean

$$m_{i,t}^T = \frac{1}{N-1} \sum_{j \neq i} h_{j,t}^T, \tag{25}$$

followed by LayerNorm:

$$c_{i,t}^T = \mathrm{LN}\big(m_{i,t}^T\big) . \tag{26}$$

The teacher logits under intervention $m \in \{0, 1\}$ are

$$\ell_{i,t}^T(m) = H_T\big(F_T\big([h_{i,t}^T;\, m c_{i,t}^T]\big)\big) , \tag{27}$$

where F_T and H_T denote the fusion module and policy head, respectively. The Joint model is trained with the full teammate context ($m = 1$) and frozen before being used as the RCD teacher.

Controlled teacher intervention. RCD evaluates the same frozen teacher twice on each transition:

$$\ell_{i,J,t}^T = \ell_{i,t}^T(1), \qquad \ell_{i,L,t}^T = \ell_{i,t}^T(0). \tag{28}$$

The restricted pass disables the privileged teammate pathway of the same teacher; it is not the output of a separately trained Local policy. The coordination-effect target is

$$C_{i,t}^T = \mathrm{ctr}\big(\ell_{i,J,t}^T - \ell_{i,L,t}^T\big) , \tag{29}$$

which removes action-independent logit shifts and retains the change in relative action preference induced by privileged teammate information. The teacher remains frozen and supplies supervision only.

F.5.2 History-Free Methods

Joint. Joint is the privileged MAPPO reference defined by Equations (24)–(27). Its actor uses all agents' current observations through encoded teammate features and the Shared-Mean context. It therefore does not satisfy the decentralized execution constraint.

Local. Local is the direct decentralized MAPPO baseline:

$$a_t^i \sim \pi_{\mathrm{Local}}\big(a_t^i \mid o_t^i\big) . \tag{30}$$

Its local representation is

$$z_{i,t} = E_{\mathrm{local}}(o_t^i), \tag{31}$$

which is mapped directly to action logits. The actor does not use teammate features, reconstruction, history, or teacher guidance. The centralized critic nevertheless receives the complete `share_obs` during training.

Local+RCD. Local+RCD augments the Local actor without changing its legal execution-time input. The student first computes

$$z_{i,t} = f_S(o_t^i). \tag{32}$$

The local predictor estimates the action-wise mean and variance of the frozen teacher coordination effect:

$$\big(\widehat{\mu}_{i,t}, \log \widehat{\sigma}_{i,t}^2\big) = q_\phi(\mathrm{sg}(z_{i,t})) , \tag{33}$$

where sg denotes stop-gradient. Thus, the RCD prediction objective does not update the student encoder.

The predicted variance is summarized by the reliability score

$$\widehat{\xi}_{i,t} = \exp\left[-\frac{1}{|\mathcal{A}_i|} \sum_{k=1}^{|\mathcal{A}_i|} \log\big(1 + \widehat{\sigma}_{i,t,k}^2\big) \right] , \tag{34}$$

so that larger predicted uncertainty gives lower reliability.

The predicted coordination effect is projected into the student feature space:

$$u_{i,t} = P_\eta(\mathrm{sg}(\widehat{\mu}_{i,t})) , \tag{35}$$

and we define

$$\bar{\xi}_{i,t} = \mathrm{sg}(\widehat{\xi}_{i,t}). \tag{36}$$

The feature-wise reliability-aware gate is

$$g_{i,t} = \bar{\xi}_{i,t}\,\sigma\big(W_g[z_{i,t}; u_{i,t}; \bar{\xi}_{i,t}] + b_g\big)\,, \tag{37}$$

and the fused representation is

$$\widetilde{z}_{i,t} = \mathrm{LN}(z_{i,t} + g_{i,t} \odot u_{i,t})\,. \tag{38}$$

The decentralized policy is

$$\pi_i^S(a_t^i \mid o_t^i) = \mathrm{PolicyHead}_S(\widetilde{z}_{i,t}). \tag{39}$$

Both $\widehat{\mu}_{i,t}$ and $\widehat{\xi}_{i,t}$ are detached before entering the policy pathway. Consequently, policy gradients update the student encoder, projection, gate, and policy head, but do not update the RCD predictor. At deployment, the teacher and o_t^{-i} are absent, and the policy remains local-observation-only.

Local+MA^{2}E. We integrate MA^{2}E (Kang et al., 2025) with the same MAPPO backbone. The history-free variant has an effective MA^{2}E window length $T_{\mathrm{MA^2E}} = 1$. The online actor receives only o_t^i; action slots in the reconstruction representation are structural placeholders and do not provide temporal memory.

For focal agent i, the MA^{2}E branch masks unavailable team observations and reconstructs them using its Transformer encoder–decoder. Denoting the reconstruction of agent j by $\widetilde{o}_t^{j|i}$, define

$$\bar{o}_t^{j|i} = \begin{cases} o_t^i, & j = i, \\ \widetilde{o}_t^{j|i}, & j \neq i, \end{cases} \qquad \bar{O}_t^{(i)} = \left(\bar{o}_t^{1|i}, \ldots, \bar{o}_t^{N|i}\right). \tag{40}$$

The filled team representation is processed by the downstream attention/controller path,

$$c_{i,t}^{\mathrm{MA^2E}} = \mathrm{Attn}_i\left(\bar{O}_t^{(i)}\right). \tag{41}$$

Reconstructed teammate information is a learned prediction; raw teammate observations are never exposed to the decentralized actor.

The reconstructed features are detached before entering the policy pathway. MAPPO therefore updates the downstream controller and policy head without updating the MA^{2}E reconstruction branch, which is optimized by its own reconstruction objective. The history-free variant uses no GRU, recurrent state, or BPTT.

The resulting history-free comparison is

$$\text{Local:} \quad o_t^i \longrightarrow a_t^i, \tag{42}$$

$$\text{Local+MA}^2\text{E:} \quad o_t^i \longrightarrow \widetilde{O}_t^{-i} \longrightarrow a_t^i, \tag{43}$$

$$\text{Local+RCD:} \quad o_t^i \longrightarrow (\widehat{\mu}_{i,t}, \widehat{\sigma}_{i,t}^2) \longrightarrow a_t^i. \tag{44}$$

Thus, all decentralized history-free methods receive the same raw actor input and differ only in how they exploit it.

F.5.3 History-Conditioned Methods

We additionally consider a finite causal local history while preserving decentralized execution. All reported history-conditioned actors are feed-forward fixed-window policies: the complete window is part of each sample, and no hidden state is carried between environment steps.

Matched temporal information. Let $T_{\mathrm{hist}} = 5$. The token for agent i at time s is

$$x_{i,s} = \left[o_s^i;\, \mathrm{onehot}(a_{s-1}^i)\right], \tag{45}$$

and the causal window used at time t is

$$H_{i,t} = (x_{i,t-4}, x_{i,t-3}, x_{i,t-2}, x_{i,t-1}, x_{i,t})\,. \tag{46}$$

Unavailable episode-start positions are zero-filled and marked invalid by $v_{i,t}$. The history is reset at episode boundaries. The current action a_t^i is never included before it is selected; it becomes the previous-action token only at the next environment step.

The implementation encodes this fixed window using feed-forward branches, including validity indicators. There is no recurrent hidden state, recurrent sequence generator, or BPTT.

For consistency with the history-free notation, define

$$X_{i,t}^{\mathrm{hist}} = (H_{i,t}, v_{i,t}), \qquad U_{i,t}^{\mathrm{hist}} = \{(H_{j,t}, v_{j,t})\}_{j \neq i}\,. \tag{47}$$

Every decentralized history-conditioned actor receives only $X_{i,t}^{\mathrm{hist}}$ at execution.

Joint-History. Joint-History is the privileged fixed-window counterpart of Joint. For each agent j, the shared history encoder computes

$$h_{j,t}^{T,\text{hist}} = E_{T,\text{hist}}(H_{j,t}, v_{j,t}). \tag{48}$$

For focal agent i, teammate representations are aggregated as

$$m_{i,t}^{T,\text{hist}} = \frac{1}{N-1}\sum_{j\neq i} h_{j,t}^{T,\text{hist}}, \qquad c_{i,t}^{T,\text{hist}} = \text{LN}\Big(m_{i,t}^{T,\text{hist}}\Big). \tag{49}$$

The teacher logits under intervention $m \in \{0,1\}$ are

$$\ell_{i,t}^{T,\text{hist}}(m) = H_{T,\text{hist}}\Big(F_{T,\text{hist}}\Big([h_{i,t}^{T,\text{hist}}; mc_{i,t}^{T,\text{hist}}]\Big)\Big). \tag{50}$$

Joint-History is trained with $m = 1$ and then frozen before being used as the teacher for Local-History+RCD.

Local-History and Local-History+RCD. Local-History is the decentralized fixed-window baseline:

$$z_{i,t}^{\text{hist}} = E_{S,\text{hist}}(H_{i,t}, v_{i,t}), \qquad \pi_i = \text{PolicyHead}_S(z_{i,t}^{\text{hist}}). \tag{51}$$

It uses only the focal agent's five-token causal history and does not pool teammate histories or query a teacher.

Local-History+RCD uses the same legal execution-time input and student encoder but is supervised by the frozen Joint-History teacher. Holding the focal sample fixed, the teacher is evaluated with its privileged teammate-history path enabled and disabled:

$$C_{i,t}^{T,\text{hist}} = \text{ctr}\Big(\ell_{i,t}^{T,\text{hist}}(1) - \ell_{i,t}^{T,\text{hist}}(0)\Big). \tag{52}$$

The student then applies the same coordination-effect predictor, Gaussian uncertainty model, reliability function, projection, gate, gradient routing, and residual fusion as Local+RCD, replacing $z_{i,t}$ with $z_{i,t}^{\text{hist}}$. No recurrent state, RNN, GRU, or BPTT is used.

Local-History+MA^{2}E. Local-History+MA^{2}E combines the same local fixed window with the masked reconstruction branch. Its actor-side history is exactly the causal $T_{\text{hist}} = 5$ observation–previous-action window in Equations (45)–(46), including episode-start padding and validity masks.

The MA^{2}E reconstruction branch masks team-history slots and predicts observation and action tokens over valid positions. Realized actions may be used as auxiliary reconstruction targets only after the corresponding environment step; they are not available to the actor before action selection. Reconstructed features are detached before entering the policy pathway, so MAPPO does not update the MA^{2}E reconstruction module.

Consequently, Local-History, Local-History+MA^{2}E, and Local-History+RCD receive exactly the same raw execution-time local history. They differ only in whether that history is used directly, used to reconstruct unavailable teammate information, or used to predict the recoverable coordination effect.

F.5.4 Optimization and Gradient Routing

MAPPO objective. Let $I_{i,t}$ denote the legal local actor input:

$$I_{i,t} = \begin{cases} o_t^i, & \text{history-free,} \\ (H_{i,t}, v_{i,t}), & \text{history-conditioned.} \end{cases} \tag{53}$$

The per-agent PPO ratio is

$$r_{i,t}(\theta) = \frac{\pi_\theta(a_{i,t} \mid I_{i,t})}{\pi_{\theta_{\text{old}}}(a_{i,t} \mid I_{i,t})}, \tag{54}$$

and the actor is optimized with the standard clipped surrogate

$$\mathcal{L}_{\text{actor}} = -\mathbb{E}_t\left[\min\Big(r_{i,t}(\theta)\widehat{A}_{i,t}, \text{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon)\,\widehat{A}_{i,t}\Big)\right]. \tag{55}$$

The centralized critic receives the complete team `share_obs` during training for all methods. This value-function input is separate from actor visibility and does not provide additional information to a decentralized actor at execution. All fixed-window history methods use ordinary feed-forward PPO minibatches.

RCD objective and gradient routing. The RCD predictor is supervised by the frozen teacher target $C_{i,t}^{T}$, or $C_{i,t}^{T,\mathrm{hist}}$ in the history-conditioned setting. Its Gaussian prediction objective is

$$\mathcal{L}_{\mathrm{RCD}} = \frac{1}{|\mathcal{A}_i|} \sum_{k=1}^{|\mathcal{A}_i|} \left[\frac{(C_{i,t,k}^{T} - \widehat{\mu}_{i,t,k})^2}{2\widehat{\sigma}_{i,t,k}^2} + \frac{1}{2} \log \widehat{\sigma}_{i,t,k}^2 \right], \tag{56}$$

with $C_{i,t}^{T,\mathrm{hist}}$ substituted in the history-conditioned case.

Let θ_S denote the student encoder, projection, gate, and policy-head parameters. Gradient routing satisfies

$$\nabla_{\theta_S} \mathcal{L}_{\mathrm{RCD}} = 0, \qquad \nabla_{\phi} \mathcal{L}_{\mathrm{RL}} = 0. \tag{57}$$

The stop-gradient at the predictor input ensures that $\mathcal{L}_{\mathrm{RCD}}$ updates the predictor parameters ϕ but not the student encoder. Conversely, the predicted mean and reliability are detached before entering the policy pathway, so the MAPPO objective updates the student encoder, projection, gate, policy head, and centralized critic without modifying the predictor.

The teacher remains frozen throughout student training and receives no gradient. The coefficient λ_C scales the predictor objective; because the prediction and reinforcement-learning objectives act on separate parameter blocks, it does not directly trade off their gradients on the same parameters.

MA^{2}E optimization. Let $\theta_{\mathrm{MA^2E}}$ denote the MA^{2}E Transformer and reconstruction parameters. Its auxiliary objective is

$$\mathcal{L}_{\mathrm{MA^2E}} = \mathcal{L}_{\mathrm{obs}} + \mathcal{L}_{\mathrm{act}}, \tag{58}$$

where observation and action reconstruction losses are applied to valid masked positions. The reconstruction branch uses a separate optimizer. Its outputs are detached before entering the downstream policy/controller path, so the MAPPO objective does not update the reconstruction module.

For both Local+MA^{2}E and Local-History+MA^{2}E, one reconstruction update is performed in the first PPO minibatch of each PPO update cycle. The history-free variant uses $T_{\mathrm{MA^2E}} = 1$, whereas the history-conditioned variant uses $T_{\mathrm{MA^2E}} = T_{\mathrm{hist}} = 5$. Neither variant uses a recurrent controller, persistent hidden state, or BPTT.

Controlled comparison. Within each decentralized comparison group, the raw execution-time actor input is held fixed:

$$\text{Local:} \quad X_i \longrightarrow a_i, \tag{59}$$

$$\text{MA}^2\text{E:} \quad X_i \longrightarrow \widehat{O}_{-i} \longrightarrow a_i, \tag{60}$$

$$\text{RCD:} \quad X_i \longrightarrow (\widehat{\mu}_i, \widehat{\sigma}_i^2) \longrightarrow a_i. \tag{61}$$

For the history-free group, $X_i = o_t^i$; for the history-conditioned group, $X_i = (H_{i,t}, v_{i,t})$. Thus, MA^{2}E uses the legal local input to reconstruct unavailable teammate information, whereas RCD uses the same input to predict the locally recoverable decision effect of a frozen privileged teacher. The MAPPO backbone, centralized critic, and raw execution-time information are otherwise matched within each comparison group.

G Additional Experimental Results

This appendix specifies the AUC calculation, provides further analysis of the local-visibility study, and reports the history-encoder comparison. The environment definitions and MAPPO implementation remain in Appendices F.4 and F.5. Unless stated otherwise, the studies use the same environment configuration, MAPPO backbone, interaction budget, optimization settings, and matched seeds as the corresponding main comparison. Final return is averaged over the final 20% of training within each seed, then summarized across five seeds. All numerical comparisons are descriptive comparisons of sample means.

G.1 History Encoder: Feed-Forward MLP versus GRU

We compare MLP and GRU variants of Joint-History, Local-History, Local-History+MA^{2}E, and Local-History+RCD on the three MPE tasks. The MLP processes an explicit fixed-length history, while the GRU maintains a recurrent hidden state and is trained using backpropagation through time. Joint uses the corresponding joint-information inputs; the MA^{2}E reconstruction and RCD coordination-prediction mechanisms are retained. Within each pair, the policy head, critic, optimization settings, interaction budget, and seeds are held fixed as reported in the experimental setup.

The MLP variant has higher mean return in $11/12$ comparisons. The sole exception is RCD on SpreadXY-4, where GRU is higher by 1.02 (-285.78 versus -286.80). For RCD, the MLP advantage is 12.86 on SpreadXY-2 and 26.01 on

Table 6: History-encoder comparison: final return (mean ± SD, five matched seeds). All actors use history; the GRU variants use BPTT. MA^{2}E and RCD abbreviate Local-History+MA^{2}E and Local-History+RCD. Bold: higher mean within each MLP–GRU pair.

Method	Encoder	SpreadXY-2	SpreadXY-4	SpreadBlindfold
Joint	MLP	$\mathbf{-76.05 \pm 0.45}$	$\mathbf{-235.50 \pm 19.23}$	$\mathbf{-135.13 \pm 0.42}$
	GRU	-77.37 ± 0.58	-246.47 ± 29.73	-149.90 ± 8.90
Local	MLP	$\mathbf{-87.68 \pm 4.25}$	$\mathbf{-283.92 \pm 11.97}$	$\mathbf{-147.79 \pm 2.09}$
	GRU	-108.31 ± 3.78	-305.84 ± 13.46	-169.70 ± 9.74
MA^{2}E	MLP	$\mathbf{-98.12 \pm 10.44}$	$\mathbf{-286.82 \pm 5.71}$	$\mathbf{-155.08 \pm 6.97}$
	GRU	-107.33 ± 5.22	-296.26 ± 11.05	-170.61 ± 4.20
RCD	MLP	$\mathbf{-85.86 \pm 3.85}$	-286.80 ± 6.34	$\mathbf{-136.80 \pm 4.76}$
	GRU	-98.72 ± 12.17	$\mathbf{-285.78 \pm 11.64}$	-162.81 ± 14.98

SpreadBlindfold. These results support the feed-forward choice for the reported experiments, without establishing a general advantage over recurrent policies.

A persistent GRU state can retain information preceding a fixed input window. The comparison therefore concerns the complete encoder implementations, including their treatment of temporal context.

Table 7: Capacity and distillation controls on the MPE tasks. Results are mean ± sample SD over five matched seeds, computed from the average return over the final 20% of training. Higher is better; bold marks the best mean per row.

Input	Task	Local	ExtraMLP	Policy-KD	Feature-KD	RCD
Current observation	SpreadXY-2	-92.39 ± 9.20	-93.07 ± 5.18	-131.52 ± 13.40	-91.59 ± 6.06	$\mathbf{-87.74 \pm 0.50}$
	SpreadXY-4	$\mathbf{-277.90 \pm 11.71}$	-291.21 ± 2.58	-359.75 ± 29.90	-280.59 ± 7.82	-286.97 ± 7.73
	SpreadBlindfold	-151.14 ± 4.71	-154.33 ± 6.66	-139.92 ± 5.67	-155.41 ± 9.87	$\mathbf{-134.85 \pm 0.68}$
Five-step history	SpreadXY-2	-87.68 ± 4.25	-96.48 ± 7.24	-106.39 ± 0.98	-88.07 ± 3.27	$\mathbf{-85.86 \pm 3.85}$
	SpreadXY-4	-283.92 ± 11.97	$\mathbf{-279.75 \pm 8.45}$	-295.06 ± 12.63	-285.41 ± 7.76	-286.80 ± 6.34
	SpreadBlindfold	-147.79 ± 2.09	-150.80 ± 10.61	-141.99 ± 10.66	-147.67 ± 8.70	$\mathbf{-136.80 \pm 4.76}$

G.2 Actor Capacity and Distillation Controls

We examine whether the gains of RCD can be explained by increased model capacity or generic supervision from the centralized teacher. To control for capacity, **ExtraMLP** augments the local actor with additional trainable parameters but receives no teacher supervision. To control for generic distillation, **Policy-KD** trains the local policy to match the teacher's action distribution, whereas **Feature-KD** directly regresses the local representation toward the corresponding teacher representation. Unlike RCD, these baselines transfer the teacher's policy or representation directly rather than predicting the coordination effect induced by privileged information. All variants use the same MAPPO backbone, training budget, environment configuration, and five matched random seeds. We report mean episodic return over the final 20% of training (8M–10M environment steps); higher is better.

Table 7 shows that neither additional capacity nor generic teacher supervision consistently reproduces the gains of RCD. ExtraMLP provides no improvement over Local on any history-free task and improves only SpreadXY-4 in the history-conditioned setting. Similarly, Policy-KD is highly task dependent: it improves over the local baseline on SpreadBlindfold but substantially degrades performance on both SpreadXY tasks. Feature-KD remains close to the corresponding local policy and yields no consistent improvement.

In contrast, RCD attains the highest mean return on SpreadXY-2 and SpreadBlindfold under both information regimes. In the history-free setting, RCD improves over Local by 4.65 points on SpreadXY-2 and 16.29 points on SpreadBlindfold, while the corresponding ExtraMLP variants do not improve over Local. On SpreadBlindfold, RCD also outperforms both conventional distillation controls, suggesting that the benefit is not explained by the mere presence of teacher supervision.

RCD does not improve upon the strongest local control on SpreadXY-4. Local obtains the best history-free mean, while Local-History+ExtraMLP obtains the best history-conditioned mean. We therefore do not interpret these ablations as evidence that RCD universally improves decentralized control. Rather, they show that the gains observed on SpreadXY-2 and SpreadBlindfold cannot be reduced to either additional actor capacity or generic teacher distillation, while also identifying SpreadXY-4 as a setting in which the proposed correction provides limited benefit.

Table 8: Auxiliary-target ablation. RCD predicts the residual coordination effect, while RCD-FullLogit predicts centered privileged teacher logits. Mean ± SD over matched seeds; higher is better.

Input	Method	SpreadXY-2	SpreadXY-4	SpreadBlindfold
Current observation	Joint	-76.16 ± 0.38	-241.44 ± 9.38	-137.89 ± 4.83
	Local	-92.39 ± 9.20	$\mathbf{-277.90 \pm 11.71}$	-151.14 ± 4.71
	Local+RCD	$\mathbf{-87.74 \pm 0.50}$	-286.97 ± 7.73	$\mathbf{-134.85 \pm 0.68}$
	Local+RCD-FullLogit	-89.79 ± 1.74	-284.04 ± 11.13	-140.62 ± 6.62
Five-step history	Joint-History	-76.05 ± 0.45	-235.50 ± 19.23	-135.13 ± 0.42
	Local-History	-87.68 ± 4.25	-283.92 ± 11.97	-147.79 ± 2.09
	Local-History+RCD	$\mathbf{-85.86 \pm 3.85}$	-286.80 ± 6.34	$\mathbf{-136.80 \pm 4.76}$
	Local-History+RCD-FullLogit	-88.14 ± 8.53	$\mathbf{-281.90 \pm 14.40}$	-147.43 ± 11.27

G.3 Target Design: Residual Coordination Effect vs. Full Privileged Logits

We next examine whether RCD benefits specifically from predicting the *coordination effect* of privileged information, rather than from predicting an arbitrary teacher-derived signal with the same auxiliary architecture.

Recall that RCD uses the centered change in the teacher's action preferences induced by privileged coordination information,

$$C_i^T = \mathrm{ctr}\left(\ell_i^T(X_i, U_i) - \ell_i^T(X_i)\right), \tag{62}$$

where $\ell_i^T(X_i, U_i)$ denotes the privileged teacher logits and $\ell_i^T(X_i)$ denotes the corresponding teacher output when privileged coordination information is removed. Thus, C_i^T isolates the incremental change in action preference attributable to information that is unavailable to the student at execution time.

As a target-control ablation, we introduce **RCD-FullLogit**, which replaces C_i^T with the centered privileged teacher logits

$$A_i^T = \mathrm{ctr}\left(\ell_i^T(X_i, U_i)\right). \tag{63}$$

RCD-FullLogit retains the same predictor output dimension, Gaussian prediction objective, uncertainty-based reliability estimate, projection module, learned fusion gate, local student inputs, teacher checkpoint, and training schedule as RCD. Both variants also use the same gradient-routing configuration. Consequently, the comparison changes only the semantics of the auxiliary prediction target: RCD predicts the *privileged-induced change* in teacher preferences, whereas RCD-FullLogit predicts the teacher's full privileged preference vector.

The results show a consistent target-dependent pattern across the history-free and history-conditioned settings. On SpreadXY-2, predicting the residual coordination effect improves the mean return over full-logit prediction by 2.05 without history and by 2.28 with history. The difference is substantially larger on SpreadBlindfold: RCD improves over RCD-FullLogit by 5.77 and 10.63, respectively. In the history-conditioned case, full-logit prediction provides almost no improvement over the Local-History baseline on SpreadBlindfold (-147.43 vs. -147.79), whereas residual prediction reaches -136.80.

SpreadXY-4 is an important exception. RCD-FullLogit achieves higher mean return than RCD by 2.93 in the history-free setting and by 4.90 in the history-conditioned setting. Moreover, neither target improves over the Local baseline in the history-free comparison. Thus, the residual target is not uniformly superior across environments.

Overall, however, the same qualitative pattern appears with and without temporal history: residual prediction performs better on SpreadXY-2 and SpreadBlindfold, while full-logit prediction performs better on SpreadXY-4. This comparison suggests that the choice of prediction target matters beyond the shared uncertainty estimation and gated-fusion architecture. In particular, directly predicting the full privileged teacher preferences is not sufficient to reproduce the gains of RCD. Subtracting the teacher's local-information response focuses the auxiliary prediction on the *change in action preference attributable to privileged coordination information*, rather than requiring the student to reproduce components of the teacher preference that are already present without that information. The SpreadXY-4 result nevertheless indicates that the advantage of this residualization is task-dependent, rather than a universal consequence of the architecture.

G.4 Effect of Explicit Reliability Modulation

We next isolate the contribution of the explicit reliability factor in RCD from that of the learned fusion gate. The **ConstantReliability** variant retains the same teacher target, mean–variance predictor, Gaussian prediction loss, projection, and learned gate as RCD, but replaces the predicted reliability by a constant value of one both as an input to the gate and as a multiplier on the projected coordination correction. Thus, if $r_i = P\hat{\mu}_i$ denotes the projected predicted coordination effect, the standard RCD fusion takes the form

$$\tilde{z}_i = z_i + g_\psi(z_i, r_i, \hat{\xi}_i)\,\hat{\xi}_i r_i,$$

Table 9: Reliability modulation ablation. Mean episodic return ± SD over five matched seeds. ConstantReliability fixes the fusion reliability to one. Higher is better.

Input	Method	SpreadXY-2	SpreadXY-4	SpreadBlindfold
Current observation	Local	-92.39 ± 9.20	$\mathbf{-277.90 \pm 11.71}$	-151.14 ± 4.71
	Local+RCD	$\mathbf{-87.74 \pm 0.50}$	-286.97 ± 7.73	$\mathbf{-134.85 \pm 0.68}$
	Local+RCD-ConstantReliability	-87.89 ± 0.59	-283.32 ± 5.95	-137.67 ± 4.53
Five-step history	Local-History	-87.68 ± 4.25	-283.92 ± 11.97	-147.79 ± 2.09
	Local-History+RCD	$\mathbf{-85.86 \pm 3.85}$	-286.80 ± 6.34	-136.80 ± 4.76
	Local-History+RCD-ConstantReliability	-93.58 ± 10.67	$\mathbf{-275.52 \pm 13.28}$	$\mathbf{-136.61 \pm 4.66}$

Table 10: Comparison with centralized-to-decentralized transfer methods. Results are mean ± sample standard deviation across five matched seeds. Higher returns are better. Bold indicates the best result among decentralized methods (**Joint** is excluded from this comparison).

Tasks	Joint	Local	Local+RCD	Local+ADVISOR	Local+MAGPO
SpreadXY-2	-76.16 ± 0.38	-92.39 ± 9.20	$\mathbf{-87.74 \pm 0.50}$	-99.20 ± 1.73	-88.61 ± 0.88
SpreadXY-4	-241.44 ± 9.38	$\mathbf{-277.90 \pm 11.71}$	-286.97 ± 7.73	-295.15 ± 12.16	-284.93 ± 13.31
SpreadBlindfold	-137.89 ± 4.83	-151.14 ± 4.71	$\mathbf{-134.85 \pm 0.68}$	-140.06 ± 5.83	-148.00 ± 5.89
Foraging-8×8	0.87 ± 0.02	0.73 ± 0.05	$\mathbf{0.76 \pm 0.09}$	0.54 ± 0.07	0.72 ± 0.07
Foraging-10×10	0.59 ± 0.03	0.38 ± 0.03	0.50 ± 0.01	$\mathbf{0.57 \pm 0.06}$	0.44 ± 0.03
RWARE-Tiny	35.83 ± 1.01	21.90 ± 1.53	21.81 ± 2.30	$\mathbf{21.94 \pm 3.24}$	20.33 ± 3.47
RWARE-Small	13.12 ± 1.86	8.60 ± 1.09	8.14 ± 1.20	$\mathbf{8.72 \pm 1.77}$	5.67 ± 0.89

whereas ConstantReliability uses

$$\tilde{z}_i = z_i + g_\psi(z_i, r_i, 1)\, r_i.$$

Importantly, the variance predictor and its training objective are left unchanged. This comparison therefore isolates whether explicitly using prediction uncertainty at fusion time provides information beyond what can already be captured by a generic learned gate.

The results show that explicit reliability modulation is useful in some, but not all, information structures. Without history, RCD and ConstantReliability are nearly identical on SpreadXY-2 (-87.74 vs. -87.89), while explicit reliability improves SpreadBlindfold by 2.82 return points. On SpreadXY-4, however, ConstantReliability performs better than standard RCD (-283.32 vs. -286.97), although neither variant improves over the Local baseline.

The dependence on the information regime becomes more pronounced when history is available. On SpreadXY-2, removing reliability modulation reduces the return from -85.86 to -93.58, a drop of 7.72 points, and also substantially increases variability across seeds. In contrast, ConstantReliability performs better on SpreadXY-4 (-275.52 vs. -286.80). The two variants are effectively comparable on SpreadBlindfold in terms of their reported means (-136.80 vs. -136.61).

Overall, these results do not indicate that explicit reliability modulation is uniformly beneficial. Instead, its contribution depends on the recoverability structure of the task and on the information already available to the local policy. In settings such as SpreadXY-2 with history and history-free SpreadBlindfold, uncertainty-aware attenuation provides an advantage over using the learned gate alone. In other settings, particularly SpreadXY-4, the learned gate with an unattenuated correction is competitive or better. This suggests that the learned gate and the explicit reliability factor provide partially overlapping forms of modulation, while reliability can act as an additional safeguard when uncertainty in the predicted coordination effect is informative for control.

G.5 Comparison with ADVISOR and MAGPO

We further compare RCD with two representative approaches for transferring information from privileged or centralized policies to decentralized agents: **ADVISOR** and **MAGPO**. This comparison is intended to distinguish RCD from more general teacher–student transfer mechanisms. Joint provides a privileged-reference upper baseline, whereas Local represents decentralized training without additional teacher guidance. All decentralized variants use the same execution-time local information and are evaluated under the same training budget and matched random seeds as the main experiments. We report mean episodic return over the final evaluation phase; higher is better.

Table 10 shows substantial task dependence in centralized-to-decentralized transfer. On **SpreadXY-2**, RCD improves over Local and performs slightly better than MAGPO, whereas ADVISOR degrades performance relative to the Local baseline. The distinction becomes stronger on **SpreadBlindfold**, where RCD achieves the highest return among all methods, including

Joint. This task removes information that is directly relevant to coordination, and the result is consistent with RCD benefiting when the effect of the missing information can still be inferred from the local observation without reconstructing the privileged input itself.

The picture is different on **SpreadXY-4**. None of the transfer methods improves upon Local: MAGPO is the strongest of the three transfer variants, followed by RCD, while ADVISOR performs worst. This result indicates that privileged teacher supervision is not automatically beneficial when the execution-time observation contains insufficient information to recover the teacher's coordination behavior. In such cases, the teacher–student mismatch can remain fundamentally ambiguous.

The Foraging environments exhibit a complementary pattern. On **Foraging-**8×8, RCD provides a modest improvement over Local and outperforms both ADVISOR and MAGPO. On the larger **Foraging-**10×10 environment, all three transfer approaches improve over Local, with ADVISOR obtaining 0.57 ± 0.06, RCD 0.50 ± 0.01, and MAGPO 0.44 ± 0.03. Thus, RCD does not uniformly dominate alternative teacher-guided methods; rather, different mechanisms benefit from different forms of information asymmetry.

Finally, the **RWARE** results expose a regime in which centralized information is valuable but difficult to exploit through decentralized distillation. Joint substantially outperforms every local-execution method on both RWARE-Tiny and RWARE-Small, while RCD, ADVISOR, and MAGPO remain close to or below the Local baseline. The large remaining Joint–Local gap therefore should not be interpreted as evidence that privileged information is directly recoverable from local observations. Instead, it suggests that a significant part of the centralized advantage may depend on information that cannot be reliably inferred by the decentralized policy at execution time.

Overall, these experiments show that the benefit of centralized-to-decentralized transfer depends not only on the quality of the teacher, but also on the relationship between privileged information and the student's execution-time observation. RCD is particularly effective on SpreadBlindfold and provides competitive gains on SpreadXY-2 and the Foraging tasks, but it deliberately does not eliminate the Joint–Local gap when the missing coordination information is weakly recoverable. This behavior is consistent with the motivation of RCD: the objective is not to reproduce the centralized teacher unconditionally, but to exploit only the coordination effect that can be predicted from information legally available to the decentralized policy.

H HYPERPARAMETERS

This section summarizes the hyperparameters and implementation settings used for all reported experiments. Unless otherwise stated, all methods use the same MAPPO backbone, centralized training with decentralized execution, generalized advantage estimation (GAE), value normalization, clipped Huber value loss, orthogonal initialization, Adam optimization with zero weight decay, and gradient clipping. Method-specific parameters are introduced only when required by the corresponding auxiliary objective.

H.1 COMMON MAPPO HYPERPARAMETERS

Table 11 lists the optimization and PPO settings shared across experiments unless explicitly overridden by an environment-specific configuration.

All agents share actor parameters. During training, the centralized critic receives the concatenated team observation, whereas each decentralized actor receives only the information permitted at execution time. For teacher-dependent methods, the teacher is frozen before student training and is used only to construct training targets or auxiliary supervision. Privileged teacher inputs are never concatenated to the local student observation.

H.2 MPE/SPREAD CONFIGURATION

Table 12 reports the configuration used for **SpreadXY-2**, **SpreadXY-4**, and **SpreadBlindfold**.

H.3 FORAGING AND RWARE CONFIGURATION

The Foraging and RWARE experiments use the same MAPPO objective but employ different rollout geometry and optimizer settings from MPE. Their configuration is summarized in Table 13.

For RWARE, the environment episode limit and PPO rollout length are different quantities. The environment may permit episodes of up to 500 steps, while the rollout length is fixed to 400 transitions. The former controls the maximum duration of an episode, whereas the latter determines the amount of experience collected before a PPO update.

Table 11: Common MAPPO hyperparameters.

Hyperparameter	Value
Discount factor γ	0.99
GAE parameter λ	0.95
Value loss	Clipped Huber
Huber delta	10.0
Value normalization	Enabled
Feature normalization	Enabled
Optimizer	Adam
Adam ϵ	10^{-5}
Weight decay	0
Maximum gradient norm	10.0
Activation	ReLU
Initialization	Orthogonal
Initialization gain	0.01
Entropy coefficient	0.01
PPO value clipping	Enabled
Linear learning-rate decay	Disabled
Mini-batches per PPO epoch	1

Table 12: Hyperparameters for the MPE/Spread experiments.

Hyperparameter	Value
Environment steps	10,000,000
Random seeds	1001–1005
Parallel rollout workers	128
Evaluation workers	32
Training threads	8
Rollout length	25
PPO epochs	5
Actor learning rate	7×10^{-4}
Critic learning rate	7×10^{-4}
Predictor learning rate	7×10^{-4}
PPO clip parameter	0.05
Hidden dimension	128
Evaluation episodes	100
Evaluation interval	100,000 steps
Checkpoint interval	100,000 steps

H.4 History and Teacher Configuration

The temporal and privileged-information settings used by the corresponding method variants are reported in Table 14.

The explicit `*-History-GRU` variants maintain a persistent recurrent state throughout the episode and are trained using contiguous truncated-BPTT segments of length 16. In contrast, `Local-History+MA2E` uses a fixed observation window of length $T = 5$ with a feed-forward controller. It therefore receives temporal context but does not maintain a persistent GRU state and does not use BPTT.

This distinction separates access to temporal information from the use of a recurrent policy architecture.

H.5 Method-Specific Hyperparameters

Table 15 reports parameters that are active only for the corresponding method family. All other parameters inherit the common and environment-specific configurations above.

The PPO objective continues to optimize the local actor and its encoder through the standard policy-gradient path. At deployment, the executed policy is always the decentralized student actor; the centralized teacher, teacher coordination targets, auxiliary losses, and centralized critic are removed.

Table 13: Hyperparameters for the Foraging and RWARE experiments.

Hyperparameter	**Value**
Environment steps	10,000,000
Random seeds	1001–1005
Parallel rollout workers	8
Evaluation workers	2
Training threads	1
Rollout length	400
PPO epochs	5
Actor learning rate	5×10^{-4}
Critic learning rate	5×10^{-4}
Predictor learning rate	7×10^{-4}
PPO clip parameter	0.2
Hidden dimension	128
Evaluation episodes	100
Evaluation interval	100,000 steps
Checkpoint interval	100,000 steps

Table 14: History and teacher configuration.

Hyperparameter	**Value**
History window T	5
GRU hidden dimension	128
BPTT chunk length	16
Shared-Mean pooling	Leave-one-out teammate mean
Post-pooling normalization	LayerNorm
Restricted teacher path	Privilege mask $m = 0$
Privileged teacher path	Privilege mask $m = 1$
Teacher pairing	Same environment and seed
Teacher checkpoint	Frozen `best.pt`

H.6 Evaluation Protocol

Unless explicitly identified as an ablation or diagnostic experiment, each method is trained from scratch for 10 million environment steps using five matched random seeds,

$$\{1001, 1002, 1003, 1004, 1005\}.$$

Evaluation uses deterministic action selection. For each seed, we first average the evaluation returns over the final 20% of training, corresponding to the interval from 8M to 10M environment steps. The reported result is then the mean and sample standard deviation of the five seed-level averages.

Teacher-dependent methods use teacher checkpoints trained for the same environment and paired with the corresponding random seed. Within each environment, all compared methods use the same training horizon, matched seed set, and evaluation protocol. Environment-specific rollout configurations are held fixed within each comparison.

Table 15: Method-specific hyperparameters. Parameters not shown here are shared with the corresponding MAPPO configuration.

Method	Hyperparameter	Value	Description
RCD	λ_{RCD}	1.0	Weight of the RCD prediction objective
RCD	Gate bias	-4.0	Initializes fusion near the low-residual regime
RCD	Variance range	$[10^{-4}, 100]$	Bounds the predicted uncertainty
Policy-KD	Distillation coefficient	1.0	Weight of the policy-distillation loss
Policy-KD	Temperature	2.0	Temperature for teacher-policy KL distillation
Feature-KD	Distillation coefficient	1.0	Weight of normalized feature MSE
ADVISOR	Auxiliary coefficient	4.0	Weight of auxiliary imitation objective
MAGPO	Ratio constraint δ	1.5	Policy-ratio constraint
MAGPO	Auxiliary PPO coefficient λ	1.0	Auxiliary PPO objective weight
MA2E	History window	$T = 1$ / $T = 5$	History-free / history-conditioned setting
MA2E	Reconstruction variance	$[10^{-4}, 100]$	Probabilistic reconstruction variant only
MA2E	Pretraining budget	0	No extra pretraining in controlled comparisons