TEAM CALIBRATION FOR MULTI-AGENT IMITATION UNDER DECENTRALIZED INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent imitation systems must turn centralized demonstrations into high-value decentralized behavior, because teacher information and team objectives are not preserved by separate local decisions. Overcoming this mismatch requires calibration to executable team utility, rather than accurate reproduction of agentwise action labels. Local imitation is well suited to decentralized deployment, yet its reliance on teacher marginals leaves the utility-optimal team protocol unidentified when value depends on joint behavior, especially in symmetric teams with private signals. Within that setting, joint-occupancy imitation captures demonstrated coordination through distribution matching, but its target remains tied to demonstrated behavior and cannot resolve utility interventions that preserve the demonstrations. We therefore propose Structured Team-Calibrated Projection (STCP), which constructs a fixed executable portfolio by applying public-context routing and shared-clock switching to theory-identified decentralized atoms and selects its deployed member by held-out team value; our analysis provides an exact phase characterization, a heterogeneous convex-closure result, and a finite-portfolio selection bound. Experiments on controlled utility interventions, synchronization tasks, and graph consensus show that STCP recovers the predicted protocol transitions and improves team value over representative decentralized baselines.

1 INTRODUCTION

Multi-agent imitation systems commonly learn from a centralized expert and deploy decentralized agents. The expert conditions on global state, while each deployed agent must produce high-value joint behavior from its local history. The central question is whether centralized demonstrations identify the behavior that a decentralized team should execute for its team utility.

Solving this mismatch requires calibration to executable team utility, rather than accurate reproduction of separate action labels. Such calibration must identify when private evidence should be followed and when a shared protocol creates greater team value.

Behavior cloning provides a direct route from demonstrations to local policies. Interactive aggregation corrects the distribution shift caused by those policies (Ross et al., 2011), while coordinated and centralized-controller imitation extend this route to decentralized multi-agent policies (Le et al., 2017; Lin et al., 2022). Adversarial and inverse soft-Q methods further match coupled trajectories or recover implicit objectives (Ho & Ermon, 2016; Song et al., 2018; Garg et al., 2021). Structured-prediction and multi-label theory explains why a surrogate must match its target decision loss (Osokin et al., 2017; Waegeman et al., 2014), yet an agentwise training signal identifies only local teacher statistics when the deployed utility depends on joint action. The common bottleneck is a team calibration gap between the imitation-optimal local rule and the utility-optimal executable protocol.

Joint-occupancy imitation is the closest direct response because it preserves coordination through distribution matching (Ho & Ermon, 2016; Song et al., 2018), but its target is drawn from demonstrated behavior and cannot distinguish the protocols preferred by different team utilities paired with identical demonstrations. This utility intervention isolates a gap that neither local accuracy nor demonstration matching resolves.

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.

We therefore propose Structured Team-Calibrated Projection (STCP), a single executable protocol grammar over decentralized atoms. Public-context routing and shared-clock switching compose the private-signal, convention, randomized, and time-indexed branches characterized by the theory, while held-out team value selects the deployed member. The accompanying analysis gives an exact phase characterization for anonymous binary teams, a convex closure beyond parameter sharing, and a finite-portfolio selection bound.

Experiments on controlled utility interventions, synchronization tasks, and graph consensus recover the predicted protocol transitions and improve team value over representative decentralized baselines. We make three contributions.

1. We formulate the team calibration gap that separates locally accurate imitation from utility-optimal decentralized behavior, including an identifiability obstruction under unchanged teacher marginals.
2. We derive an exact private-evidence versus convention phase characterization and prove when it remains optimal among heterogeneous agent rules. This structure yields the executable portfolio in STCP.
3. We connect the characterization to learned policy transitions, synchronization protocols, and certified graph-consensus behavior across controlled and independently maintained environments.

2 Related Work

Multi-agent imitation. Behavior cloning can compound its own distribution shift, motivating interactive aggregation and occupancy-matching alternatives (Ross et al., 2011; Ho & Ermon, 2016). Coordinated and adversarial multi-agent extensions address coupled trajectories (Le et al., 2017; Song et al., 2018; Lin et al., 2022). More recently, IQ-Learn avoids adversarial reward-policy optimization (Garg et al., 2021), while MIFQ and MAFIS factor inverse soft-Q objectives for cooperative teams (Bui et al., 2024; Li et al., 2025), and MisoDICE learns from unlabeled mixed-quality demonstrations (Bui et al., 2025). Select to Perfect uses collective desirability to identify behavior for selective imitation, while recent regret-gap theory studies robustness to strategic deviations (Franzmeyer et al., 2024; Tang et al., 2024). Our setting asks when fixed centralized demonstrations identify the utility-optimal executable protocol for a cooperative decentralized team.

Structured and multi-label calibration. Calibration theory characterizes when minimizing a surrogate recovers a target decision rule (Bartlett et al., 2006; Tewari & Bartlett, 2007; Ramaswamy & Agarwal, 2016). Structured surrogates and hypothesis-class-specific consistency extend this view beyond binary classification (Osokin et al., 2017; Awasthi et al., 2022). Bayes-optimal multi-label prediction shows that marginal and joint metrics require different statistics (Waegeman et al., 2014; Wu & Zhu, 2020; Mao et al., 2024). Our setting adds distinct local information fields and restricts the output to a deployable product policy.

Symmetric stochastic teams. Decentralized detection and stochastic-team theory formalize how information constraints restrict joint decisions (Tsitsiklis, 1993; Bernstein et al., 2002; Yüksel & Başar, 2024). Symmetry can reduce the policy class, sometimes only after admitting randomized rules (Sanjari et al., 2023). Decentralized detection exhibits both regimes. Identical rules can be asymptotically optimal in a binary special case, while finite designs admit asymmetric counterexamples (Tsitsiklis, 1988; Cherikh & Kantor, 1992). Recent work studies local optima and symmetric strategies under decentralized information (Emmons et al., 2022; Sudhakara & Nayyar, 2023). We identify a discrete convex utility class where anonymous execution is globally optimal, then use the resulting characterization to construct a finite executable portfolio.

Value factorization. Centralized training with decentralized execution includes actor-critic and counterfactual credit-assignment methods (Lowe et al., 2017; Foerster et al., 2018). VDN and QMIX impose additive or monotone factorizations, while QTRAN, Weighted QMIX, and QPLEX relax different representational limits (Sunehag et al., 2018; Rashid et al., 2018; Son et al., 2019; Rashid et al., 2020; Wang et al., 2021). STCP may include policies learned by these architectures, but its

theoretical object is validation-time selection among executable protocol classes under information constraints.

3 Problem Setup

Let $(X, Y) \sim P$, where $X = (X_1, \ldots, X_n)$ are local histories and $Y = (Y_1, \ldots, Y_n)$ is a centralized teacher action. A deterministic executable policy is

$$d = (d_1, \ldots, d_n) \in \mathcal{D}_{\mathrm{dec}}, \qquad d(X) = (d_1(X_1), \ldots, d_n(X_n)). \tag{1}$$

A separable teacher surrogate and a team risk are

$$S_P(f) = \sum_{i=1}^{n} \mathbb{E}_P[s_i(f_i(X_i), Y_i)], \qquad R_{U,P}(d) = \mathbb{E}_P[L_U(d(X), Y)]. \tag{2}$$

A link ψ maps a surrogate predictor to an executable decision $d = \psi(f) \in \mathcal{D}_{\mathrm{dec}}$. We write $\mathrm{Reg}_{U,P}(d) = R_{U,P}(d) - \inf_{d' \in \mathcal{D}_{\mathrm{dec}}} R_{U,P}(d')$.

Definition 1 (Executable team protocol)**.** *An executable team protocol is a collection of decision rules whose inputs are restricted to each agent's local history and shared public variables available at deployment. A stationary decentralized policy is a special case.*

Definition 2 (Executable team calibration)**.** *A surrogate and link are team-calibrated over distribution class $\mathcal{P}$ and executable class $\mathcal{D}_{\mathrm{dec}}$ if surrogate excess risk converging to zero implies team excess risk relative to $\mathcal{D}_{\mathrm{dec}}$ converging to zero for every $P \in \mathcal{P}$.*

The definition isolates the statistics that local imitation must identify under decentralized execution.

Proposition 1 (Marginal indistinguishability)**.** *Suppose $P_{X_i,Y_i} = Q_{X_i,Y_i}$ for every agent. Then every separable local surrogate has the same risk under P and Q. If every decision in the image of the link obeys $\max\{\mathrm{Reg}_{U,P}(d), \mathrm{Reg}_{U,Q}(d)\} \geq \eta > 0$, the surrogate and link cannot be team-calibrated on both.*

The proof is immediate from equality of each local expectation and a common near-minimizing surrogate sequence. For finite decision classes, disjoint team-optimal sets imply a positive η. Appendix C.1 gives a two-agent witness with equal local marginals, incompatible unique joint modes, and $\eta = 0.1$.

The impossibility result has a sharp positive boundary. If $L_U(a, Y) = \sum_i c_i L_i(a_i, Y_i)$ with $c_i > 0$ and each local surrogate is calibrated for L_i, the product of local Bayes decisions minimizes the team risk. Nonseparable interaction, rather than multi-agent structure alone, creates the obstruction.

Corollary 1 (Utility identifiability)**.** *Fix the demonstration distribution and executable policy class. Suppose two team utilities U, U' obey the uniform incompatibility condition*

$$\inf_{d \in \mathcal{D}_{\mathrm{dec}}} \max\{\mathrm{Reg}_{U,P}(d), \mathrm{Reg}_{U',P}(d)\} \geq \eta > 0. \tag{3}$$

Then a surrogate and link that do not receive the utility cannot be calibrated for both utilities. In a finite executable class, disjoint optimal sets imply this condition.

The demonstration statistics are identical by construction, so the surrogate and utility-blind link generate the same decision sequence under both utilities. The uniform gap prevents both regrets from vanishing. For a finite class, disjoint optimal sets yield a positive minimum gap. This corollary motivates the paired utility-phase experiment, which changes the interaction spectrum while keeping every teacher label and private observation fixed.

4 Method

Structured Team-Calibrated Projection (STCP) instantiates one executable protocol grammar through three declared objects: a training-generated set of decentralized atoms, available public-context routing variables, and a shared-clock switch grid. The grammar composes these objects into a finite portfolio, and held-out team value selects the deployed protocol. The theory characterizes the private-evidence, convention, randomized, and temporal branches that guide atom and operator declaration. We develop this construction through the static phase characterization, randomized shared rules, finite portfolio selection, and sequential protocol separation.

4.1 EXACT INTERACTION-SPECTRUM CHARACTERIZATION

The identifiability obstruction applies to the general product class $\mathcal{D}_{\mathrm{dec}}$. For an exact structural characterization, we specialize to anonymous execution, where exchangeable agents share one rule and have no agent identifier. Write $\mathcal{D}_{\mathrm{shared}} = \{(d, \dots, d) : d : \{0,1\} \to \{0,1\}\}$. For $n \geq 2$, let $Z \sim \mathrm{Bernoulli}(1/2)$. Conditional on Z, agents independently observe binary signals satisfying

$$\Pr(X_i = Z \mid Z) = p, \qquad p \geq \tfrac{1}{2}. \tag{4}$$

A policy in $\mathcal{D}_{\mathrm{shared}}$ applies the same deterministic rule $d : \{0,1\} \to \{0,1\}$ at every agent. If $K(d) = \sum_i \mathbf{1}\{d(X_i) = Z\}$ agents are correct, the team obtains monotone utility $g(K)$.

Define

$$B_g(p) = \sum_{k=0}^{n} g(k) \binom{n}{k} p^k (1-p)^{n-k}, \qquad C_g = \frac{g(0) + g(n)}{2}. \tag{5}$$

Theorem 1 (Anonymous phase and convex closure)**.** *For every nonconstant monotone g, the optimum in the shared deterministic binary policy class is*

$$J^*_{\mathrm{shared-det}}(p) = \max\{B_g(p), C_g\}. \tag{6}$$

If $g(0) < g(n)$, the unique threshold $p_c \in (0,1)$ satisfying $B_g(p_c) = C_g$ separates the identity and common-convention branches on the full signal-accuracy interval. When restricted to the operational domain $p \geq 1/2$, a threshold below $1/2$ means that identity is optimal throughout. If g is additionally discrete convex, meaning $\Delta^2 g(k) \geq 0$, Equation (6) *is also the optimum over all agent-specific deterministic product rules.*

The convex closure is sharp. Outside this regime, role-split conventions can form an additional heterogeneous branch; the appendix gives an exact quorum witness that this branch can dominate every shared rule.

Sketch. There are four deterministic rules, comprising identity, flip, and two constants. Their values are $B_g(p)$, $B_g(1-p)$, and C_g. Monotonicity gives $B_g(p) \geq B_g(1-p)$ for $p \geq 1/2$, while

$$B'_g(p) = n \sum_{k=0}^{n-1} \Delta g(k) \binom{n-1}{k} p^k (1-p)^{n-1-k} > 0 \tag{7}$$

on the open interval for nonconstant monotone g. For convex closure, replace every flip rule by identity, align all constant agents by convexity, and let a be the number of identity agents. The resulting value is convex in a, so its maximum occurs at $a = 0$ or $a = n$, giving C_g and $B_g(p)$.

Under the identity rule, let $C_i = \mathbf{1}\{X_i = Z\}$ and $\mu_r = \Delta^r g(0)$. Then the C_i are conditionally independent Bernoulli(p) variables, and the symmetric pseudo-Boolean expansion and its expectation yield

$$g\left(\sum_i C_i\right) = \sum_{S \subseteq [n]} \mu_{|S|} \prod_{i \in S} C_i, \qquad B_g(p) = \sum_{r=0}^{n} \binom{n}{r} \mu_r p^r. \tag{8}$$

At the uninformative boundary, the spectrum gives

$$B_g(1/2) - C_g = -\sum_{r=2}^{n} \binom{n}{r} \mu_r (1/2 - 2^{-r}). \tag{9}$$

Thus nonnegative higher-order coefficients, with at least one positive, place a convention region next to the uninformative boundary, while the full spectrum determines when private information overcomes that region.

Corollary 2 (Curvature and scale)**.** *For nonconstant affine g, $p_c = 1/2$; for strictly discrete-convex g, $p_c > 1/2$; and for strictly discrete-concave g, $p_c < 1/2$. For $g_n(k) = (k/n)^r$, $p_c(n) \to 2^{-1/r}$; for all-match utility, $p_c = 2^{-1/n} \to 1$; and for proportional quorum $g(k) = \mathbf{1}\{k \geq \lceil \rho n \rceil\}$, $p_c(n) \to \rho$.*

4.2 RANDOMIZED SHARED RULES

Suppose agents conditionally sample actions with independent private random coins, and let $q_x = \Pr(A_i = 1 \mid X_i = x)$. Define

$$r_0 = p(1 - q_0) + (1 - p)(1 - q_1), \qquad r_1 = (1 - p)q_0 + pq_1. \tag{10}$$

The exact randomized value is

$$J(q_0, q_1) = \tfrac{1}{2}\{B_g(r_0) + B_g(r_1)\}. \tag{11}$$

Theorem 2 (A boundary for randomization)**.** *If B_g is convex on $[0, 1]$, a maximizing shared randomized rule exists at a vertex $(q_0, q_1) \in \{0, 1\}^2$. If B_g is concave, the identity rule is optimal. For general monotone mixed-curvature utilities, randomization can strictly improve on all four deterministic rules.*

Sketch. The convex claim follows because a convex function on the policy square has a maximizer at a vertex. In the concave case, Jensen's inequality and $(r_0 + r_1)/2 \le p$ give $J(q_0, q_1) \le B_g(p)$. We provide an explicit $n = 8$ mixed-curvature counterexample in Appendix C.8. With agent-specific randomized product rules, the value is multi-affine in each agent's action probabilities, so a deterministic heterogeneous maximizer always exists. The interior gain above is specific to parameter sharing.

4.3 EXECUTABLE PROTOCOL PROJECTION

Operationally, STCP fixes before validation a finite set $\mathcal{A}$ of decentralized atoms, the public routing variables, and a switch-time grid $\mathcal{T}$. The atoms are training-generated local or utility-aware learners and analytic evidence or convention endpoints. The portfolio is $\mathcal{A}$ together with public-context routings of atoms and $\mathrm{Switch}_t(\pi, \pi')$ for $\pi, \pi' \in \mathcal{A}$ and $t \in \mathcal{T}$. Routing and switching use only public deployment variables, so every candidate remains executable. Appendix Table 11 lists each finite instantiation.

Let the training set contain teacher demonstrations, let $\widehat{Q}(s, \mathbf{a})$ be a bounded centralized team critic available only during training, and let

$$\Pi_{\mathrm{dec}} = \left\{ \pi : \pi(\mathbf{a} \mid \boldsymbol{\tau}) = \prod_i \pi_i(a_i \mid \tau_i) \right\} \tag{12}$$

be the deployable class. We train a local candidate and a fidelity-indexed candidate family.

$$\widehat{\pi}_{\mathrm{local}} \in \arg\min_{\pi \in \Pi_{\mathrm{dec}}} \widehat{L}_{\mathrm{BC}}(\pi), \tag{13}$$

and, for a fixed set Λ,

$$\widehat{\pi}_\lambda \in \arg\min_{\pi \in \Pi_{\mathrm{dec}}} \left[-\widehat{J}_{\widehat{Q}}(\pi) + \lambda \widehat{L}_{\mathrm{BC}}(\pi) \right]. \tag{14}$$

The atom generators, optimizer families, routing variables, and switch-time grid are fixed before validation. The selection theorem is optimizer-agnostic, and the cost of generating the full portfolio is reported.

The procedure trains a local imitation candidate, a finite set of team-value and fidelity candidates, and any theory-derived executable protocols declared before evaluation. It scores every candidate on independent validation data and deploys the one with the largest lower-confidence team value. Full pseudocode appears in Appendix D.

When every candidate is evaluated on the same number of bounded validation episodes, the Hoeffding confidence radius below is common across candidates. Maximizing the lower confidence bound is therefore exactly equivalent to maximizing empirical validation value. Candidate-dependent sample counts would instead require candidate-dependent radii. Any bounded raw return in $[a, b]$ can be affinely normalized to $[0, 1]$ without changing candidate rankings; multiplying the bound by $b - a$ returns it to the original units.

For small action spaces, the product-policy expectation can be exact. For larger spaces, we use sampled joint actions or a centralized actor-critic gradient. The deployment policy always uses only local history.

Theorem 3 (Finite candidate selection). *Let $\mathcal{C}$ contain M candidates fixed independently of m independent and identically distributed validation instances $\xi_1, \ldots, \xi_m \sim \nu$. For a candidate π, let $V(\pi;\xi) \in [0,1]$ denote its true normalized value on instance ξ. Set $\bar{Q} = \mathrm{clip}(\widehat{Q}, 0, 1)$, and let $V_{\bar{Q}}(\pi;\xi)$ be its clipped-critic value with $\sup_{\pi,\xi} |V_{\bar{Q}}(\pi;\xi) - V(\pi;\xi)| \le \epsilon_Q$. Write $J(\pi) = \mathbb{E}_{\xi\sim\nu} V(\pi;\xi)$ for population value and $\widehat{J}_{\bar{Q}}(\pi) = m^{-1}\sum_{j=1}^{m} V_{\bar{Q}}(\pi;\xi_j)$ for the validation average. If $\widehat{\pi}$ maximizes $\widehat{J}_{\bar{Q}}$, then with probability at least $1-\delta$,*

$$\max_{\pi\in\mathcal{C}} J(\pi) - J(\widehat{\pi}) \le 2\epsilon_Q + 2\sqrt{\frac{\log(2M/\delta)}{2m}}. \tag{15}$$

Relative to the best policy $\pi^* \in \arg\max_{\pi\in\Pi_{\mathrm{dec}}} J(\pi)$, add the candidate approximation error $\epsilon_{\mathrm{app}} = J(\pi^*) - \max_{\pi\in\mathcal{C}} J(\pi)$.

Sketch. A union bound makes the validation deviation uniform over the M fixed candidates; empirical optimality then costs two confidence radii, and replacing the clipped critic by the true value costs $2\epsilon_Q$. The guarantee separates selection error from candidate approximation. When validation uses observed normalized returns, $\epsilon_Q = 0$; when the local candidate belongs to $\mathcal{C}$, the same inequality also controls degradation relative to local imitation. In the deterministic phase class, the identity and convention branches give $\epsilon_{\mathrm{app}} = 0$. Appendix C proves the phase, curvature, randomization, selector, and synchronization results, and states explicitly where strict monotonicity, independent validation, and bounded returns enter.

4.4 Sequential protocol separation

With $t = 0$ at initialization, consider $n \ge 2$ agents at private distances $x_i \in \{1, \ldots, L\}$ from a terminal boundary. A left action reduces a positive distance by one, wait leaves it unchanged, and right increases it up to L. Choosing left at distance one causes arrival. The team succeeds only if all agents arrive on the same step; arrival by a strict subset terminates in failure.

Proposition 2 (Stationary local impossibility). *No deterministic stationary local rule $d(x)$ guarantees success from every initial state.*

From the all-one state, a successful stationary rule must choose left at position one. Waiting is absorbing, while moving right cannot later cross one because revisiting the same local state repeats the same action. In a state containing distances one and two, choosing left at one causes a strict subset to arrive immediately, a contradiction.

Proposition 3 (Worst-case optimal time-aware protocol). *The local time-aware rule*

$$\pi_t(x) = \text{left} \quad \Longleftrightarrow \quad t \ge L - x \tag{16}$$

makes every agent arrive exactly at time L from every initial state and is worst-case completion-time optimal.

An agent at distance x waits $L - x$ steps and then moves left for x steps. The all-L initial state gives an L-step lower bound for every policy.

5 Experiments

Our experiments span controlled binary utility interventions, action-dependent sequential tasks, the independently maintained COGNAC consensus benchmark, and CASEC Hallway synchronization. They address four research questions. **RQ1.** Do learned policies recover the static phase predicted by the theory? **RQ2.** Can utility-conditioned selection recover distinct executable protocols from identical demonstrations, including under sequential dynamics? **RQ3.** Do structured protocol classes transfer to graph consensus with finite-state certificates? **RQ4.** Does time-indexed execution resolve the predicted synchronization obstruction? All learned results use five random seeds. Hallway and the reported COGNAC certificates additionally use exhaustive evaluation.

5.1 Setup

All candidates are fixed using training data and selected using validation return before test evaluation. Baselines share the deployable information structure within each environment. Tables report

Table 1: Controlled utility intervention. Best in bold; tied baselines are shown separately.

Agents	p	Local BC	Conditional occupancy	Agreement	Global branch	STCP
4	.70	$.6161 \pm .0005$	$.5014 \pm .0022$	$.6161 \pm .0005$	$.6161 \pm .0005$	$\mathbf{.7466 \pm .0012}$
6	.70	$.5586 \pm .0008$	$.4981 \pm .0017$	$.5586 \pm .0008$	$.5586 \pm .0008$	$\mathbf{.7487 \pm .0009}$
8	.75	$.5502 \pm .0010$	$.5016 \pm .0018$	$.5502 \pm .0010$	$.5502 \pm .0010$	$\mathbf{.7508 \pm .0009}$

sample standard deviations across five random seeds for learned results; Appendix Table 4 reports return dispersion across held-out initial states for a fixed selected protocol. The best mean in each comparable group is shown in **bold**, and a unique second-best mean is underlined.

5.2 Recovery of the predicted static phase

To answer the first research question, we compare learned local policies with the exact static branch across convex, concave, affine, and threshold utilities. Across these utilities, neural local policies reproduce the predicted deterministic branch in 40 of 40 evaluated configuration runs. Across the eight utility and accuracy configurations, the validation selector has mean test-oracle regret 0.00101 and maximum regret 0.00816 under five random seeds; every nonzero regret is between conventions with equal population value.

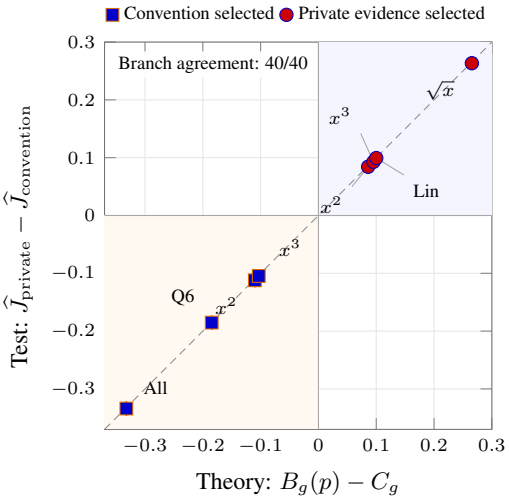

Figure 1: Static phase recovery. Shading highlights agreement regions between theory and test for the convention and private-evidence branches; markers show the validation-selected branch, and bars show sample standard deviations across five random seeds. The vertical difference uses the higher-return fixed convention on the test set.

Correlation-aware baselines account for much of the low-information gain. Joint-occupancy matching nearly matches team projection when a stable unconditional convention is sufficient, while fidelity and team-value candidates better preserve conditional information in higher-information regimes.

5.3 Utility-conditioned protocol recovery

To answer the second research question, we hold demonstrations fixed and intervene only on team utility. Every sample is paired across two public regimes with the same hidden target and private signals. One regime rewards only the event that all agents are correct; the other rewards the event that at least one agent is correct. A shared deterministic policy maps regime and private signal to action, giving only 16 policies, which we enumerate exactly.

The selected STCP candidate has the same aggregate value as team ERM. It adopts a convention for the all-correct utility and follows private information for the any-correct utility. Local imitation,

Table 2: Sequential utility-phase team value. Best mean in bold and second best underlined.

Method	Team value
Validation-selected BC/agreement GRU	$.60452 \pm .00263$
IQL	$.58885 \pm .03037$
VDN	$.60229 \pm .01743$
QMIX	$.59629 \pm .00534$
Validation-selected IQL/VDN/QMIX	$\underline{.61051 \pm .01395}$
STCP	$\mathbf{.74890 \pm .00147}$

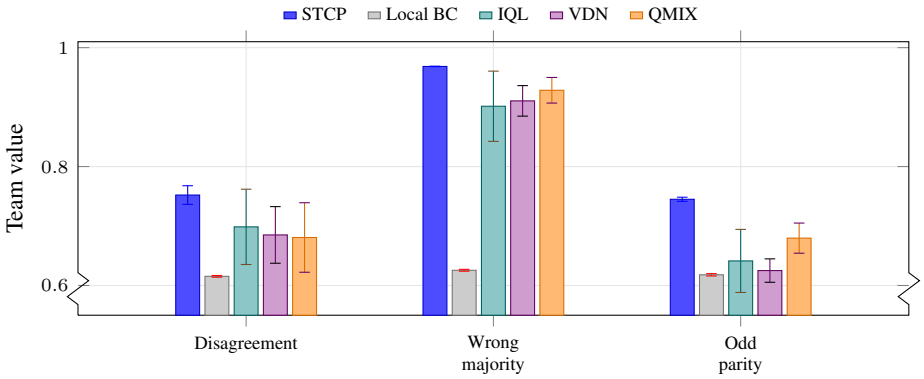

Figure 2: Action-dependent team value over five random seeds. STCP denotes the expanded actor-and-value portfolio; bars show means and sample standard deviations. The broken axis marks the truncated display range.

agreement selection, and the global single-branch objective recover the same shared rule in this intervention, while the regime-conditioned joint-marginal objective selects a constant rule. Team-value selection recovers the utility-conditioned switch from identical demonstrations.

We also extend the paired intervention to ten-step target sequences and shared GRU policies. The target follows a Markov chain, but actions do not affect its transition; consequently IQL, VDN, and QMIX use the correct $\gamma = 0$ contextual-bandit target. All three train on identical paired random-exploration trajectories, and QMIX's centralized mixer observes the hidden target and utility regime during training. The utility-agnostic neural control in Table 2 is selected by validation among local BC and agreement-regularized GRUs with coefficients 0.1, 0.5, and 1.

The paired STCP advantage over the validation-selected value-factorization ensemble is $.13838 \pm .01499$. The result supports the predicted utility-conditioned switch under function approximation in this controlled sequential setting.

5.4 Persistence under action-dependent dynamics

To extend the second research question, we make joint actions affect future state. Four agents act for eight steps under three transition mechanisms. The next hidden binary target flips after any disagreement, after a strict wrong majority, or after odd action parity. The centralized teacher always outputs the true target, obtains unit reward under both utilities, and generates identical paired demonstrations. Shared GRU actors use policy gradients, while BC-initialized IQL, VDN, and QMIX use online temporal-difference learning and independent validation checkpoint selection. All reported methods use fixed training budgets and five random seeds.

Figure 2 reports the expanded STCP portfolio. Independent validation selects one candidate from the fixed actor-and-value portfolio in each run. Applying the same validation rule within the IQL/VDN/QMIX family provides an additional portfolio comparator. STCP retains the higher mean on all three transitions and positive paired 95% intervals on wrong-majority and odd-parity transitions. On disagreement transitions, selection among value learners and actors obtains

Table 3: COGNAC robustness over five random seeds. Best in bold and second best underlined.

Method	Return	Success	Mean steps
Anonymous IQL	$\underline{6.88390 \pm 1.45217}$	$\underline{.97436 \pm .04424}$	$\underline{12.5456 \pm 6.5442}$
Two-phase STCP	$\mathbf{9.49584 \pm .00741}$	$\mathbf{1.00000 \pm .00000}$	$\mathbf{2.7652 \pm .0181}$

$.75213 \pm .01567$ with test-oracle regret $.00202 \pm .00272$, supporting optimizer complementarity. Appendix Table 8 reports all intervals.

5.5 Transfer to independent graph consensus

To answer the third research question, we test structured protocol transfer beyond our generators in the externally maintained COGNAC Binary Consensus environment (Sintes & Bušić, 2025). Ten agents occupy a ring and observe their own and two neighboring binary states. We retain the official transition, default factored reward, and 100-step horizon. Each decentralized candidate is a shared Boolean rule over $(\text{left}, \text{self}, \text{right})$; there are $2^{2^3} = 256$ stationary rules. The expanded STCP portfolio additionally allows one validation-selected shared-clock switch between two such rules.

We independently sample 1,024 training initial states to shortlist six stationary rules, 1,024 validation initial states to select the rule pair and switch time, and 5,000 test initial states for the reported score. Each split follows the official reset distribution, which samples the number of zeros uniformly and their positions uniformly conditional on that count. The three splits use independent draws. The validation search contains $6 \times 6 \times 256 = 9{,}216$ two-phase candidates. A designated representative selection chooses rule 242 for one step and rule 183 thereafter. We freeze this protocol for the exhaustive certificate and ring-size analysis below. The five-run robustness result independently repeats the same fixed search procedure. Every selected candidate is executable because each action depends only on the agent's local ring neighborhood and the shared clock.

We also enumerate all $2^{10} - 2 = 1{,}022$ non-consensus states after selection. The chosen protocol succeeds from every state, takes at most five steps, and has minimum return 7.8. A separate 256-trajectory check matches the official PettingZoo implementation exactly in return, success, and episode length. To match the executable information structure, anonymous IQL receives the same canonical $(\text{left}, \text{self}, \text{right})$ ordering, a shared recurrent agent without agent identifiers, and a shared previous-action input. It trains for 100,000 environment steps with validation-only checkpoint selection.

Across five random seeds, independently selected STCP protocols have higher mean return and success and lower mean completion time than anonymous IQL. Exhaustive evaluation of the frozen representative protocol certifies consensus from all non-consensus ring-10 states. The component ablation in Appendix Table 6 shows that temporal composition is necessary for this certificate.

Freezing the ring-10 protocol across sizes reveals the periodic structure in Appendix Figure 3. The proved period-three and period-four obstruction families account for every observed non-consensus case through size twenty. The protocol is never below either local baseline through size sixteen and is strictly better in every remaining comparison. Complete per-size results appear in Appendix Table 10.

5.6 Resolving the synchronization obstruction

On CASEC Hallway, exact and sampled STCP recover the worst-case optimal time-indexed protocol from every initial state in all five runs, while MIFQ does so in four. The five-agent instance is also solved exhaustively; Appendix Table 5 gives the complete comparison.

6 Conclusion

STCP turns the team calibration gap into executable private-evidence, convention, and temporal protocols selected by held-out team value. Controlled interventions and finite-state certificates connect the theory to learned team behavior.

AI USE STATEMENT

Generative AI tools assisted with research development, technical implementation, analysis, and manuscript preparation. The authors reviewed and verified all AI-assisted work and take responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

The project artifact contains the environment adapters, candidate-generation and selection code, exact evaluation scripts, structured JSON outputs, and a claim-to-evidence matrix linking numerical claims to their source records. Hallway and the final COGNAC certificate enumerate their finite state spaces. All reported learned-policy results use five random seeds and documented validation-based checkpoint selection. Appendix D records the full protocol details and the boundary between official code, local adapters, and controlled generators.

REFERENCES

Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. H-consistency bounds for surrogate loss minimizers. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1117–1174. PMLR, 2022. URL https://proceedings.mlr.press/v162/awasthi22c.html.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006. doi: 10.1198/016214505000000907.

Daniel S. Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of markov decision processes. *Mathematics of Operations Research*, 27(4): 819–840, 2002. doi: 10.1287/moor.27.4.819.297.

The Viet Bui, Tien Mai, and Thanh Hong Nguyen. Inverse factorized soft Q-learning for cooperative multi-agent imitation learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2fbeed1dd7162f91804e7b9246e0c1a8-Abstract-Conference.html.

The Viet Bui, Tien Mai, and Thanh Hong Nguyen. MisoDICE: Multi-agent imitation from mixed-quality demonstrations. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/efe51d056f63d4546e73aa8a97dcb6de-Abstract-Conference.html.

Moula Cherikh and Paul B. Kantor. Counterexamples in distributed detection. *IEEE Transactions on Information Theory*, 38(1):162–165, 1992. doi: 10.1109/18.108263. URL https://doi.org/10.1109/18.108263.

Scott Emmons, Caspar Oesterheld, Andrew Critch, Vincent Conitzer, and Stuart Russell. For learning in symmetric teams, local optima are global Nash equilibria. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5924–5943. PMLR, 2022. URL https://proceedings.mlr.press/v162/emmons22a.html.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11794. URL https://ojs.aaai.org/index.php/AAAI/article/view/11794.

Tim Franzmeyer, Edith Elkind, Philip Torr, Jakob Foerster, and João F. Henriques. Select to perfect: Imitating desired behavior from large multi-agent data. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=L6crLU7MIE.

Divyansh Garg, Shuvam Chakraborty, Chris Cundy, Jiaming Song, and Stefano Ermon. IQ-Learn: Inverse soft-q learning for imitation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4028–4039, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/210f760a89db30aa72ca258a3483cc7f-Abstract.html.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/cc7e2b878868cbae992d1fb743995d8f-Abstract.html.

Hoang M. Le, Yisong Yue, Peter Carr, and Patrick Lucey. Coordinated multi-agent imitation learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1995–2003. PMLR, 2017. URL https://proceedings.mlr.press/v70/le17a.html.

Yi-Chen Li, Zhongxiang Ling, Tao Jiang, Fuxiang Zhang, Pengyuan Wang, Lei Yuan, Zongzhang Zhang, and Yang Yu. Multi-agent imitation by learning and sampling from factorized soft Q-function. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/13a6f537e943a598b8719b03181a8e2a-Abstract-Conference.html.

Alex Tong Lin, Mark Debord, Katia Estabridis, Gary Hewer, Guido Montufar, and Stanley Osher. Decentralized multi-agents by imitation of a centralized controller. In *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145 of *Proceedings of Machine Learning Research*, pp. 619–651. PMLR, 2022. URL https://proceedings.mlr.press/v145/lin22a.html.

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*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/68a9750337a418a86fe06c1991a1d64c-Abstract.html.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Multi-label learning with stronger consistency guarantees. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/04ad66d02234541aac5143de7876e880-Abstract-Conference.html.

Anton Osokin, Francis Bach, and Simon Lacoste-Julien. On structured prediction theory with calibrated convex surrogate losses. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/38db3aed920cf82ab059bfccbd02be6a-Abstract.html.

Harish G. Ramaswamy and Shivani Agarwal. Convex calibration dimension for multiclass loss matrices. *Journal of Machine Learning Research*, 17(14):1–45, 2016. URL https://jmlr.org/papers/v17/14-316.html.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4295–4304. PMLR, 2018. URL https://proceedings.mlr.press/v80/rashid18a.html.

Tabish Rashid, Gregory Farquhar, Bei Peng, and Shimon Whiteson. Weighted QMIX: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/73a427badebe0e32caa2e1fc7530b7f3-Abstract.html.

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

Sina Sanjari, Naci Saldi, and Serdar Yüksel. Optimality of independently randomized symmetric policies for exchangeable stochastic teams with infinitely many decision makers. *Mathematics of Operations Research*, 48(3):1254–1285, 2023. doi: 10.1287/moor.2022.1296. URL https://pubsonline.informs.org/doi/10.1287/moor.2022.1296.

Jules Sintes and Ana Bušić. COGNAC: Cooperative graph-based networked agent challenges for multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/089ae994919488e764ed271a132124f1-Abstract-Datasets_and_Benchmarks_Track.html.

Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5887–5896. PMLR, 2019. URL https://proceedings.mlr.press/v97/son19a.html.

Jiaming Song, Hongyu Ren, Dorsa Sadigh, and Stefano Ermon. Multi-agent generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/240c945bb72980130446fc2b40fbb8e0-Abstract.html.

Sagar Sudhakara and Ashutosh Nayyar. Optimal symmetric strategies in multi-agent systems with decentralized information. In *2023 IEEE 62nd Conference on Decision and Control*, pp. 2784–2789, 2023. doi: 10.1109/CDC49753.2023.10384231. URL https://doi.org/10.1109/CDC49753.2023.10384231.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2085–2087, 2018. URL https://www.ifaamas.org/Proceedings/aamas2018/pdfs/p2085.pdf.

Jingwu Tang, Gokul Swamy, Fei Fang, and Zhiwei Steven Wu. Multi-agent imitation learning: Value is easy, regret is hard. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3103b25853719847502559bf67eb4037-Abstract-Conference.html.

Ambuj Tewari and Peter L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8(36):1007–1025, 2007. URL https://jmlr.org/papers/v8/tewari07a.html.

John N. Tsitsiklis. Decentralized detection by a large number of sensors. *Mathematics of Control, Signals, and Systems*, 1(2):167–182, 1988. doi: 10.1007/BF02551407.

John N. Tsitsiklis. Decentralized detection. In H. Vincent Poor and John B. Thomas (eds.), *Advances in Statistical Signal Processing*, volume 2, pp. 297–344. JAI Press, 1993. URL https://www.mit.edu/~jnt/Papers/B-93-detection.pdf.

Willem Waegeman, Krzysztof Dembczyński, Arkadiusz Jachnik, Weiwei Cheng, and Eyke Hüllermeier. On the bayes-optimality of f-measure maximizers. *Journal of Machine Learning Research*, 15(103):3513–3568, 2014. URL https://jmlr.org/papers/v15/waegeman14a.html.

Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. QPLEX: Duplex dueling multi-agent Q-learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Rcmk0xxIQV.

Tonghan Wang, Liang Zeng, Weijun Dong, Qianlan Yang, Yang Yu, and Chongjie Zhang. Context-aware sparse deep coordination graphs. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=wQfgfb8VKTn.

Guoqiang Wu and Jun Zhu. Multi-label classification: Do Hamming loss and subset accuracy really conflict with each other? In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/20479c788fb27378c2c99eadcf207e7f-Abstract.html.

Serdar Yüksel and Tamer Başar. *Stochastic Teams, Games, and Control under Information Constraints*. Systems and Control: Foundations and Applications. Birkhäuser Cham, 2024. doi: 10.1007/978-3-031-54071-4. URL https://link.springer.com/book/10.1007/978-3-031-54071-4.

A Benchmark Datasets and Evaluation Protocol

A.1 Controlled utility interventions

These paper-specific environments isolate executable team calibration in finite binary teams. The static study uses eight utility and signal-accuracy cases; the paired study uses four, six, and eight agents; the sequential study uses six agents over ten steps; and the action-dependent study uses four agents over eight steps under disagreement, wrong-majority, and odd-parity transitions. Each agent observes a noisy private binary signal and chooses a binary action. The centralized state records the latent target, public utility regime, and joint history, while deployment exposes only the local history and public regime. Static rewards are monotone functions of the number of correct agents. Paired tasks use all-correct and any-correct team utilities. Evaluation reports held-out team value, selected branch, oracle regret, and transition diagnostics. Learned configurations use five random seeds with independent training, validation, and test samples; finite shared-policy classes are additionally enumerated exactly. These are controlled diagnostic tasks whose exact branch structure makes the theoretical prediction directly testable.

A.2 COGNAC binary consensus

COGNAC is an independently maintained graph-based MARL suite (Sintes & Bušić, 2025). We use its Binary Consensus environment with ten agents on a ring and retain the official transition, FactoredRewardModel, and one-hundred-step horizon. Each agent observes its own binary state and those of its two neighbors, then chooses one binary action; the centralized state is the ten-bit ring configuration. The reward gives local disagreement penalties, a time-weighted consensus reward, and a truncation penalty. We report mean return, consensus success, and completion steps.

For each of five random seeds, 1,024 training states rank all stationary Boolean rules and retain six, 1,024 independently sampled validation states select the two-rule protocol and shared switch time, and 5,000 independent test states provide the reported estimates. The switch-time search is fixed to one, two, four, eight, sixteen, and twenty-four steps. We additionally enumerate all 1,022 non-consensus ring states and freeze the selected protocol for transfer across ring sizes five through twenty. The ring instance is finite but nontrivial, with an exponentially growing state space and no benchmark-defined easy, hard, and super-hard tiers.

A.3 Hallway synchronization

Hallway is the synchronization task from the MACO benchmark introduced with CASEC (Wang et al., 2022). Our main instance uses one group of three agents with corridor length four, and the larger confirmation uses five agents with length six. Each agent observes its private position and active flag; the centralized state concatenates all local observations. The three discrete actions wait, move toward the boundary, and move away from it. The group receives reward ten only when every member reaches the boundary together; arrival by a strict subset terminates that group without success.

We report greedy all-state success and completion steps. Training and validation use the budgets specified in Appendix D, and every learned configuration uses five random seeds. Test evaluation enumerates all sixty-four initial states in the main instance and all 7,776 states in the larger instance. The main case is deliberately small enough for an exact certificate, while the larger case tests the same hidden-timing obstruction on a substantially larger finite state space.

Table 4: Held-out COGNAC performance. Best mean in bold and second best underlined.

Policy	Return	Success	Mean steps
Fixed convention	-5.9989 ± 12.1624	.3410	66.4816
Local majority	-12.3715 ± 12.6517	.1872	81.5540
Best stationary local rule	$\underline{4.5431 \pm 20.6373}$	$\underline{.9374}$	$\underline{8.8176}$
Two-phase STCP	$\mathbf{9.4881 \pm .3864}$	**1.0000**	**2.7844**

Table 5: Hallway performance. Best greedy success in bold, second best underlined.

Method	Greedy success	Mean steps
Reactive BC	0.0625	1.5625
History BC	0.4375	3.0156
Hand boundary sweep	**1.0000**	7.0000
Exact STCP	**1.0000**	4.0000
Sampled STCP	**1.0000**	4.0000
MIFQ official-core adapter	$\underline{0.9500}$	4.0000

B Per-Environment Numerical Results

Tables 1, 2, 4, 3, and 5, together with Figures 1, 2, and 3, report the retained held-out comparisons. Table 10 additionally gives the frozen-policy baseline comparison for ring sizes five through sixteen. Tables 7 and 9 give the exact values behind the two corresponding figures. Table 6 isolates the contribution of the validation-selected temporal composition.

C Theorem Proofs

C.1 Proof of Proposition 1

Proof. For every f,

$$S_P(f) = \sum_i \mathbb{E}_{P_{X_i,Y_i}} s_i(f_i(X_i), Y_i) = \sum_i \mathbb{E}_{Q_{X_i,Y_i}} s_i(f_i(X_i), Y_i) = S_Q(f).$$

Any common surrogate-regret minimizing sequence would, under simultaneous calibration, have linked team regret converging to zero under both distributions, contradicting the uniform positive gap. For an explicit witness, let local observations be constant and let exact-match utility act on $Y \in \{0,1\}^2$. The two distributions are

$$P = (0.3, 0.3, 0.4, 0), \qquad Q = (0.6, 0, 0.1, 0.3)$$

over $(00, 01, 10, 11)$. Both local marginals satisfy $\Pr(Y_1 = 1) = 0.4$ and $\Pr(Y_2 = 1) = 0.3$, but the unique joint modes are 10 and 00. The minimum over decisions of the maximum regret across the pair is 0.1. □

C.2 Proof of Corollary 1

Proof. Because the demonstration distribution and every input available to the surrogate and link are unchanged, any fixed surrogate-regret minimizing sequence produces the same linked executable decisions under U and U'. For every member of that sequence, the uniform incompatibility condition makes at least one of the two team regrets at least η. The two regrets therefore cannot both converge to zero, contradicting simultaneous calibration. If $\mathcal{D}_{\mathrm{dec}}$ is finite and its two optimal sets are disjoint, each decision has positive regret for at least one utility; the minimum over the finite class supplies such an η. □

C.3 Proof of Theorem 1

Proof. The four maps from one binary signal to one binary action are identity, flip, and the two constants. Conditional independence gives values $B_g(p)$, $B_g(1-p)$, and C_g for the two constants.

Table 6: COGNAC protocol ablation. Best in bold and second best underlined.

Variant	Return	Success	Mean steps
Local majority	$-12.07977 \pm .17199$	$.19764 \pm .00745$	$80.5214 \pm .7365$
First rule only	$\underline{4.92796 \pm .28993}$	$\underline{.94300 \pm .00401}$	$\underline{8.2662 \pm .3912}$
Second rule only	$-12.30295 \pm .44857$	$.62300 \pm .00651$	$39.0718 \pm .6336$
Validation-best stationary	$4.84357 \pm .30319$	$.94168 \pm .00434$	$8.3917 \pm .4231$
Full two-phase STCP	$\mathbf{9.49584 \pm .00741}$	$\mathbf{1.00000 \pm .00000}$	$\mathbf{2.7652 \pm .0181}$

Table 7: Action-dependent team value underlying Figure 2. Selected VF denotes validation selection among IQL, VDN, and QMIX; best is bold and second best is underlined.

Transition	STCP	Local BC	IQL	VDN	QMIX	Selected VF
Disagreement	$\mathbf{.75213 \pm .01567}$	$.61541 \pm .00144$	$.69862 \pm .06331$	$.68503 \pm .04762$	$.68072 \pm .05845$	$\underline{.72480 \pm .04542}$
Wrong majority	$\mathbf{.96841 \pm .00038}$	$.62559 \pm .00158$	$.90146 \pm .05907$	$.91049 \pm .02566$	$.92834 \pm .02155$	$\underline{.92853 \pm .02165}$
Odd parity	$\mathbf{.74500 \pm .00347}$	$.61798 \pm .00196$	$.64132 \pm .05303$	$.62508 \pm .01970$	$.67964 \pm .02535$	$\underline{.68484 \pm .03170}$

For monotone nonconstant g, the standard derivative identity for Bernstein polynomials gives

$$B_g'(p) = n \sum_{k=0}^{n-1} \Delta g(k) \binom{n-1}{k} p^k (1-p)^{n-1-k} > 0 \qquad (0 < p < 1),$$

because every summand is nonnegative, at least one finite difference is positive, and every Bernstein basis term is positive in the open interval. Thus $B_g(p) \geq B_g(1-p)$ when $p \geq 1/2$. Moreover, $B_g(0) = g(0) < C_g < B_g(1) = g(n)$, so continuity and strict monotonicity give a unique root $p_c \in (0,1)$ of $B_g(p_c) = C_g$. Comparing the remaining identity and constant branches yields the claimed shared-class maximum and decision boundary. When $p_c < 1/2$, the convention region lies outside the assumed $p \geq 1/2$ domain, so identity is optimal throughout that domain.

Now allow agent-specific deterministic rules and assume g is discrete convex. For an agent-specific uniform variable U_i, couple the correctness indicator of its identity rule as $\mathbf{1}\{U_i \leq p\}$ and that of its flip rule as $\mathbf{1}\{U_i \leq 1-p\}$. Replacing a flip by identity weakly increases the number correct, so monotonicity eliminates all flip rules.

Suppose a agents use identity and the remaining $m = n - a$ agents are constant. If b constants output zero and $m - b$ output one, with $S_a \sim \mathrm{Binomial}(a, p)$ the value is

$$\frac{1}{2}\mathbb{E}[g(S_a + b) + g(S_a + m - b)].$$

The sequence $h_a(t) = \mathbb{E}g(S_a + t)$ is discrete convex. Therefore $t \mapsto h_a(t) + h_a(m-t)$ is convex and is maximized at $t \in \{0, m\}$. Thus all constants can be aligned. Their common action is immaterial by state symmetry, and the value becomes

$$F(a) = \frac{1}{2}\mathbb{E}[g(S_a) + g(n - a + S_a)].$$

For $0 \leq a \leq n-2$, direct differencing gives

$$\Delta^2 F(a) = \frac{1}{2}\left[p^2 \mathbb{E}\Delta^2 g(S_a) + (1-p)^2 \mathbb{E}\Delta^2 g(n - a - 2 + S_a)\right] \geq 0.$$

Therefore F is discrete convex and attains its maximum at an endpoint. The endpoints are $F(0) = C_g$ and $F(n) = B_g(p)$, proving convex closure over the full heterogeneous deterministic class.

For sharpness, take $n = 8$, $p = 7/10$, and quorum utility $g(k) = \mathbf{1}\{k \geq 4\}$. The best shared rule has value $18840647/20000000$, while assigning four agents constant zero and four agents constant one gives value one. Thus discrete convexity marks a sharp boundary for the heterogeneous closure claim. □

C.4 Proof of Corollary 2

Proof. If $g(k) = a + bk$ with $b > 0$, then $B_g(p) = a + bnp$ and $C_g = a + bn/2$, hence $p_c = 1/2$. For a strictly discrete-convex sequence, every interior point lies strictly below the chord joining $(0, g(0))$

Table 8: Paired uncertainty for principal comparisons. Positive differences favor STCP; intervals are two-sided 95% Student-t intervals across five runs.

Comparison	Mean difference	95% interval	Wins
Static, 4 agents vs. local BC	.13048	$[.12959, .13138]$	5/5
Static, 6 agents vs. local BC	.19015	$[.18885, .19145]$	5/5
Static, 8 agents vs. local BC	.20060	$[.19874, .20246]$	5/5
Sequential vs. selected IQL/VDN/QMIX	.13838	$[.11977, .15700]$	5/5
Disagreement vs. selected IQL/VDN/QMIX	.02733	$[-.01942, .07407]$	3/5
Wrong majority vs. selected IQL/VDN/QMIX	.03988	$[.01305, .06670]$	5/5
Odd parity vs. selected IQL/VDN/QMIX	.06016	$[.02350, .09682]$	5/5

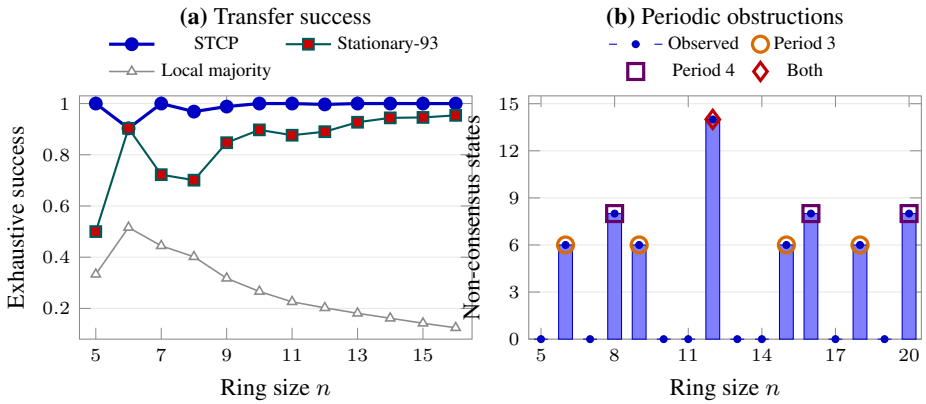

Figure 3: Frozen COGNAC transfer. The panels show exhaustive success against local baselines and agreement between observed non-consensus cases and the proved periodic obstruction families.

and $(n, g(n))$. A Binomial$(n, 1/2)$ variable assigns positive mass to the interior, so $B_g(1/2) < C_g$ and strict monotonicity implies $p_c > 1/2$. Reversing the chord inequality proves the concave case.

For $g_n(k) = (k/n)^r$ with fixed $r > 0$, let $K_n \sim \text{Binomial}(n, p)$. Uniform continuity of u^r on $[0, 1]$, together with Hoeffding concentration for K_n/n, gives $\sup_{p\in[0,1]} |\mathbb{E}[(K_n/n)^r] - p^r| \to 0$. Monotonicity and bracketing of the unique roots therefore give $p_c(n) \to 2^{-1/r}$. For all-match utility, $B_g(p) = p^n$ and $C_g = 1/2$, yielding $p_c = 2^{-1/n}$. Finally, for fixed $0 < \rho < 1$ and $g(k) = \mathbf{1}\{k \geq \lceil \rho n \rceil\}$, binomial concentration sends $B_g(p)$ to zero for every $p < \rho$ and to one for every $p > \rho$. The unique half-probability root is consequently trapped in every neighborhood of ρ, proving $p_c(n) \to \rho$. □

C.5 Proof of Theorem 2

Proof. The displayed formula for J follows by conditioning on Z. Both r_0 and r_1 are affine in (q_0, q_1). If B_g is convex, then J is convex on the unit square. Writing any point as a convex combination of the four vertices shows its value is no larger than the largest vertex value, so a deterministic maximizer exists.

If B_g is concave, Jensen's inequality gives

$$J(q_0, q_1) \leq B_g\left(\frac{r_0 + r_1}{2}\right).$$

Since $(r_0 + r_1)/2 = [1 + (2p - 1)(q_1 - q_0)]/2 \leq p$ for $p \geq 1/2$, and B_g is nondecreasing, $J(q_0, q_1) \leq B_g(p)$, the identity value. For strict improvement in the mixed-curvature regime, use the rational witness in Appendix C.8. Direct substitution gives

$$\max_{(q_0,q_1)\in\{0,1\}^2} J(q_0, q_1) = \frac{33087231}{64000000}, \qquad J\left(\frac{17}{25}, 1\right) = \frac{554346834479969}{953674316406250},$$

Table 9: Periodic obstruction structure underlying Figure 3.

Ring sizes	Non-consensus cases per size	Exhaustive success pattern
$5, 7, 10, 11, 13, 14, 17, 19$	0	1.0
$6, 9, 15, 18$	6	$1 - 6/(2^n - 2)$
$8, 16, 20$	8	$1 - 8/(2^n - 2)$
12	14	$1 - 14/(2^{12} - 2)$

and their difference is $125560160032554637/1953125000000000000 > 0$. Thus a randomized shared rule strictly exceeds every deterministic shared rule. For agent-specific randomized rules, hold all conditional action probabilities except one fixed. Expected team value is affine in the remaining probability, so one endpoint is weakly better. Repeating this coordinate by coordinate produces a deterministic heterogeneous rule with no smaller value. □

C.6 Proof of Theorem 3

Proof. For candidate π, write $J_{\bar{Q}}(\pi) = \mathbb{E}_{\xi\sim\nu} V_{\bar{Q}}(\pi;\xi)$. Hoeffding's inequality and a union bound imply that, simultaneously for all M candidates, with probability at least $1-\delta$,

$$|\widehat{J}_{\bar{Q}}(\pi) - J_{\bar{Q}}(\pi)| \le c_m := \sqrt{\frac{\log(2M/\delta)}{2m}}.$$

Let $\pi_{\mathcal{C}} \in \arg\max_{\pi\in\mathcal{C}} J(\pi)$. Empirical optimality of $\widehat{\pi}$ and the simultaneous event give

$$J_{\bar{Q}}(\pi_{\mathcal{C}}) - J_{\bar{Q}}(\widehat{\pi}) \le 2c_m.$$

The uniform value error changes each candidate value by at most ϵ_Q, so replacing both clipped-critic values by true values costs at most $2\epsilon_Q$. This proves the displayed bound. Adding $J(\pi^*) - J(\pi_{\mathcal{C}}) = \epsilon_{\text{app}}$ gives the comparison with the full executable class. The argument requires the candidate set to be fixed without using these m validation samples; training data may be used arbitrarily to construct it. □

C.7 Proofs of Propositions 2 and 3

Proof. For the stationary impossibility, consider the all-one state. A rule that waits forever at position one cannot finish, while a rule that moves away and later revisits one repeats the same stationary action and cannot cross the boundary. Universal success therefore requires the arrival action at one. Now consider an initial state containing at least one agent at distance one and another at distance two. The former arrives immediately while the latter cannot move more than one unit in that step, so a strict subset arrives and the episode fails. This contradicts universal success.

Under the time-aware rule, an agent initially at distance x waits for $L - x$ steps and then makes exactly x unit moves, arriving at time L. Hence all agents arrive simultaneously from every initial state. From the all-L state no agent can arrive in fewer than L unit moves, so no policy has smaller worst-case completion time. □

Proposition 4 (Periodic ring obstructions)**.** *For the protocol that applies rule 242 once and rule 183 thereafter, if* $3 \mid n$ *then the six states obtained by translating* $(100)^{n/3}$ *and* $(011)^{n/3}$ *never reach consensus. If* $4 \mid n$*, the eight states obtained by translating* $(1000)^{n/4}$ *and* $(0111)^{n/4}$ *never reach consensus.*

Proof. The cellular update is translation equivariant, so it suffices to evaluate one representative of each periodic word and its complement. Write only one period of each ring state. For period three, rule 242 maps $100 \mapsto 011$ and $011 \mapsto 100$. Rule 183 maps $100 \mapsto 011$ and fixes 011. Consequently either initial word reaches the non-consensus fixed word 011 after at most one rule-183 step. Their three translations give six states.

For period four, rule 242 maps $1000 \mapsto 0010$ and $0111 \mapsto 1000$. Under rule 183 the four non-consensus words form the cycle

$$0010 \mapsto 1101 \mapsto 1000 \mapsto 0111 \mapsto 0010.$$

Table 10: Frozen-policy success through ring size sixteen. Best in bold, second best underlined.

n	Two-phase STCP	Stationary-93	Local majority
5	**1.000000**	.500000	.333333
6	**.903226**	**.903226**	.516129
7	**1.000000**	.722222	.444444
8	**.968504**	.700787	.401575
9	**.988235**	.847059	.317647
10	**1.000000**	.897260	.266145
11	**1.000000**	.876344	.225806
12	**.996580**	.890083	.202247
13	**1.000000**	.926984	.180952
14	**1.000000**	.944024	.161641
15	**.999817**	.945797	.142221
16	**.999878**	.953917	.124363

Thus both initial words enter the same non-consensus four-cycle. Their four translations give eight states. Repeating a verified period around any ring whose length is divisible by that period preserves these local updates, proving the two claims. □

C.8 Randomized-policy counterexample

For $n = 8$, $p = 0.5$, and

$$g = (0, .404044, .422707, .429694, .454418, .550707, .802767, .886448, 1),$$

the best deterministic shared rule has value 0.51698798. The explicit randomized rule $(q_0, q_1) = (0.68, 1)$ has value 0.58127479 and therefore strictly dominates every deterministic shared rule. All displayed inputs are finite decimals, so clearing denominators turns this substitution into an exact rational inequality rather than a numerical optimality assumption.

D Implementation Details

D.1 Selection procedure

D.2 Our method's hyperparameters and executed configurations

All neural STCP candidates use Adam with coefficients $(0.9, 0.999)$, numerical constant 10^{-8}, and zero weight decay. Unless stated otherwise, training is full batch, the final iterate is evaluated on validation data, and no replay buffer, target network, target update, warm-up, or early stopping is used. Candidate generation uses training data, candidate selection uses only validation value, and test data are accessed once after selection. Every learned configuration uses five random seeds.

The static phase experiment uses a shared two-hidden-layer MLP of width 32, learning rate 0.02, and 300 epochs. Each of the eight utility cases uses 3,000 training, 3,000 validation, and 50,000 test samples. Its fixed portfolio contains local imitation, pure team value, an imitation-weighted candidate with coefficient 0.1, and both constant conventions. Validation team value selects the final member. The paired utility intervention evaluates all 16 shared deterministic policies. It uses 2,000 training samples to select each objective-specific candidate; 4,000 validation and 100,000 test samples are used. The fixed agreement-control search is 0.1, 0.5, and 1, and the selected STCP policy is the exact validation-best team-value policy.

The sequential utility intervention uses a shared one-layer GRU of width 32 and a linear two-action head. Adam uses learning rate 0.01 for 180 epochs with gradient-norm clipping at 10 on 1,000 training sequences; validation and test contain 2,000 and 20,000 sequences. The fixed portfolio contains local imitation, pure team value, and imitation weights 0.01 and 0.1. STCP uses direct team-value

Algorithm 1 Structured Team-Calibrated Projection

1: Declare atom generators, public routing variables, and switch grid.
2: Generate or train every decentralized atom using training data.
3: Apply the declared routing and switch operators to form the portfolio.
4: Score every executable candidate on independent validation data.
5: Deploy the candidate with maximum lower-confidence team value.

Table 11: Fixed instantiations of the STCP protocol grammar. Every atom generator and operator is specified before validation.

Study	Decentralized atoms	Grammar component	Validated set
Static phase	BC, team/fidelity, conventions	atom coverage	5 candidates
Paired utility	four binary rules per public regime	public-context routing	16 policies
Sequential utility	BC, team/fidelity GRUs	public-context routing	4 candidates
Action-dependent	six GRU actors, IQL, VDN, QMIX	public-context routing	9 candidates
COGNAC	256 Boolean rules, six training-shortlisted first rules	one shared-clock switch	9,216 protocols
Hallway	position-based fidelity candidates	shared time index	4 objectives and checkpoints

optimization without a bootstrapped target. The corresponding value-learning comparison uses discount zero, and validation team value selects the final candidate. In each action-dependent environment, the same GRU is initialized by 120 imitation epochs at learning rate 0.01 and optimized for 180 policy-gradient updates at learning rate 0.003. Each update uses two regime-balanced half-batches of 512, discount 0.95, entropy coefficient 0.005, and gradient-norm clipping at 10. Training uses 1,000 expert sequences; each regime contributes 2,000 validation and 10,000 test sequences. The actor pool contains local imitation, agreement coefficients 0.5 and 1, pure team value, and imitation weights 0.01 and 0.1. The expanded pool adds IQL, VDN, and QMIX, for nine candidates in total; validation return selects the final member.

For exact Hallway, tabular logits are optimized for 300 epochs at learning rate 0.1 over fidelity weights 0, 0.01, 0.1, and 1. Sixteen states define the validation subset, and all 64 initial states are exhaustively evaluated. The Monte Carlo variant uses 600 updates, batch size 2,048, learning rate 0.03, entropy coefficient 0.001, gradient-norm clipping at 10, and at most 64 validation states. Validation is evaluated every 25 updates and at the final update; the best validation checkpoint is retained. The fidelity weight is selected by validation return, with completion steps as the tie breaker. The five-agent, length-six configuration changes only the state space and exhaustive test set.

COGNAC uses no learned network, optimizer, or gradient update in the final STCP candidate. In each run, 1,024 training states rank all 256 stationary Boolean rules and retain the top six. Independent validation searches first rules, switch times 1, 2, 4, 8, 16, and 24, and all 256 second rules, giving 9,216 two-phase candidates. A designated representative selection returns rule 242 followed by rule 183 with switch time one. This representative is frozen for exhaustive enumeration of all 1,022 non-consensus states and for the ring-size audit over sizes five through twenty. Five independent executions of the same search procedure provide the robustness summary, each followed by a test on 5,000 independently sampled states.

D.3 Baseline implementation summary

Reactive and history behavior cloning, IQL, VDN (Sunehag et al., 2018), and QMIX (Rashid et al., 2018) use local adapters with the same observations, training data or interaction budget, validation split, and held-out evaluation as STCP. VDN and QMIX follow the cited additive and monotone mixing architectures; their checkpoints are chosen by validation return. Hallway MIFQ uses the official inverse-Q objective, recurrent Q network, and monotone mixer (Bui et al., 2024) inside a local CASEC Hallway adapter (Wang et al., 2022). Its supplementary archive has SHA256 `774b8269a33fadadde864237b3dc9f5cf18e1400a0405ded8ec86bf0dc80c06b`. The COGNAC comparison uses the official environment at commit `dd357cc2e7c46f707cca2e766810b19de3017f73` and a local anonymous-IQL adapter with the same interaction, validation, and held-out test-state protocol. These provenance labels distinguish official components from local adapters while keeping the comparison protocol explicit.

E COMPUTATIONAL RESOURCES

Experiments were executed on a Linux server with NVIDIA A100-SXM4 80 GB GPUs. The software environment uses Python 3.12, NumPy 1.26.4, PyTorch 2.2.0, and CUDA 12.1.

F LIMITATIONS

The analysis centers on symmetric information structures that permit exact calibration certificates, while the experiments span controlled interventions, sequential synchronization, and independently maintained graph consensus. STCP shifts work to training-time portfolio construction, while deployment remains decentralized under the same local information structure as the compared policies.