

The Coordination Complexity Index: A Parameter-Free Checkpoint Diagnostic for Silent Coordination Failure in Cooperative Multi-Agent Reinforcement Learning

Chatchitsanu Pothisakha¹, Pudsadee Boonrawd^{2*}, Siranee Nuchitprasitchai³

Data Science for Innovation Program-Faculty of Information Technology and Digital Innovation,
King Mongkut's University of Technology North Bangkok, Bangkok, Thailand¹

Department of Information Technology-Faculty of Information Technology and Digital Innovation,
King Mongkut's University of Technology North Bangkok, Bangkok, Thailand^{1,2,3}

Abstract—Cooperative multi-agent reinforcement learning (MARL) can fail silently while standard reward and loss curves remain inconclusive. This study introduces the Coordination Complexity Index (CCI), a checkpoint statistic with no learned parameters that summarizes dispersion among per-agent greedy Q-values. A frozen checkpoint cannot by itself establish a temporal failure; it can reveal Q-value patterns that motivate further inspection. On Level-Based Foraging (LBF), rules calibrated on the VDN/QMIX/QTRAN triad assign operational high-occupancy, dual-occupancy, and near-zero-dispersion labels, respectively. These labels are not independent proof of stagnation, oscillation, or behavioral roles. The pooled collapsed cohort differs from the non-collapsed LBF cohort by more than four orders of magnitude at the cohort geometric mean, while QMIX attains higher return with lower coordination variance than VDN. Held-out QPLEX falsifies the proposed class-complete-IGM prediction: return is zero, yet C does not approach the collapse cut-off. At 10% of the canonical LBF budget, a retrospective $\bar{C} < 10^{-3}$ rule flags all retained QTRAN seeds without flagging retained QMIX seeds; on RWARE it produces at least one environment-induced false positive. CCI is therefore a candidate domain-specific checkpoint screen, not a universal early-stop rule or a coordination-quality score. With the cut-off frozen before evaluation, 5/5 additional QTRAN training seeds remain below it at the retained 10%-budget checkpoint and 5/5 remain below it near 10M steps; this is a positive-class repeatability check, not independent calibration or a new specificity control.

Keywords—Cooperative Multi-Agent Reinforcement Learning (MARL); value decomposition; coordination signatures; coordination collapse; Coordination Complexity Index (CCI); checkpoint screening; Centralized Training with Decentralized Execution (CTDE)

I. INTRODUCTION

A cooperative multi-agent reinforcement learning (MARL) experiment can consume days of training compute yet fail silently: even without outright collapse, the loss may keep falling and team return may keep rising after the joint policy has already settled into a degenerate, task-irrelevant regime. Because a cooperative team maximizes a shared return under partial observability and decentralized execution, neither the

reward curve nor any single agent's local view necessarily exposes that failure.

Value-decomposition (VD) methods, a leading family for cooperative training, combine per-agent Q-values through centralized mixers with progressively relaxed constraints, from VDN and QMIX to QTRAN and QPLEX [1]–[4]. This progression suggests an implicit assumption: greater mixer expressiveness should yield at least equal team performance. That assumption is tested directly with a concrete counterexample. On the canonical Level-Based Foraging (LBF) 9×9 -5p grid, QTRAN collapses to zero return (0.000), whereas QMIX reaches 0.732. Thus, greater expressiveness alone does not prevent coordination failure. In single-agent deep learning, a network can quietly lose the ability to learn even as gradient-descent training continues [5]. Cooperative MARL similarly lacks a low-cost checkpoint observable for inspecting silent degeneration. Table I positions the proposed scalar against related diagnostics, but the comparison is qualitative rather than an empirical head-to-head. Reward curves can obscure whether disagreement is useful, and return is not a substitute for the proposed coordination proxy C : QMIX earns a higher return than VDN with lower coordination variance, consistent with the externally reported sample-efficiency gain under monotonicity [6].

This study makes three contributions.

- A checkpoint diagnostic with no learned parameters. C , the Coordination Complexity Index [CCI, Eq. (2)], is a per-timestep scalar computed from the variance of per-agent Q-values without an auxiliary loss. Its classifier uses thresholds calibrated on the LBF VDN/QMIX/QTRAN triad. The diagnostic can be extracted from a frozen checkpoint and is provided as the `maccci` package described in the Data and Code Availability statement.
- A three-pattern operational taxonomy. On LBF, VDN, QMIX, and QTRAN exhibit high occupancy, dual occupancy, and near-zero dispersion under the calibrated rules. Only the QTRAN pattern transfers unchanged to RWARE. The conventional labels STAGNATION,

*Corresponding author.

BI-REGIME, and COLLAPSE are retained for reproducibility, but they describe checkpoint statistics rather than independently verified temporal or behavioral mechanisms. On seed-matched cohorts from QFix, QTypeMix, and DuelMIX, C also separates successful dual-occupancy seeds from failed one-band seeds; this is partial rather than universal cross-architecture evidence.

- A held-out evaluation that bounds where C applies. The held-out prediction is that a *class-complete* Individual-Global-Max (IGM) architecture reaches the $\bar{C} < 10^{-3}$ collapse cut-off by the fixed 10M-step checkpoint. On five held-out QPLEX seeds that prediction fails: $\bar{C}$ remains heterogeneous rather than collapsing. The accompanying return collapse (5/5 seeds at 0.000) is reported separately. The result indicates that C tracks coordination-variance collapse rather than every form of policy failure.

II. RELATED WORK

Value-decomposition methods organize cooperative training around a centralized mixing function that recombines per-agent Q-values, and the family is usually read as a ladder of increasing expressiveness. VDN sums the per-agent values, satisfying the Individual-Global-Max (IGM) condition trivially but representing no inter-agent interaction [1]. QMIX replaces the sum with a state-conditioned monotonic mixing network whose non-negative weights preserve IGM while admitting nonlinear interactions [2]. QTRAN and QPLEX climb further, the former relaxing monotonicity through auxiliary losses and the latter through a duplex dueling advantage formulation, so as to cover a strictly larger IGM function class [3], [4]. The organizing premise of this ladder is that broader function-class coverage yields at-least-equal team performance.

The 2024–2026 literature continues to climb the same ladder rather than questioning the premise. DuelMIX, QTypeMix, and QFix extend factorization for heterogeneous teams or incomplete base mixers [7]–[9]; related fixed mixers reshape the mixing algebra itself (including QVF, which integrates quantile regression into the IGM condition) [10]–[15]. Each mixer above is fixed at design time and unchanged during training, and none poses the complementary question of what to measure when added expressiveness does not translate into performance—the question taken up here.

A second thread supplies the contrary empirical evidence that motivates a diagnostic. Rashid et al. established monotonic mixing as the practical default on cooperative StarCraft (SMAC) [2], and Hu et al. later showed that a carefully tuned QMIX matches the SMAC state of the art, with the monotonicity constraint improving sample efficiency on purely cooperative tasks [6]. This inverts the ladder’s premise, since the constraint that the more expressive methods discard may itself be doing useful work. Consistent with that reading, value decomposition provably cannot represent the global value function on symmetric multi-modal reward landscapes, for every IGM-consistent mixer and whether or not the local networks are shared [16]. Linear and monotonic factorizations, in turn, cannot guarantee consistency between per-agent greedy actions and the team-optimal joint value [17]. Empirically, Papadopoulos et al. report that QPLEX fails with near-zero return

on the hardest LBF tasks [18]—a silent collapse that registers in neither the training loss nor a cursory return summary until the compute has already been spent. Standardized cooperative-evaluation protocols [19], [20] improve reporting but still do not contribute a single observable that diagnoses coordination structure from a checkpoint.

The third thread concerns direct measurement of inter-agent coordination, which remains scarce relative to mixer design. Parameter-free probes such as Role Diversity [21], System Neural Diversity [22], CASEC [23], the Specialization Index [24], LVMQ [25], entropy and update probes [26], neuron statistics [27]–[29], greedy-action churn [30], and action-value gaps [31] characterize related but different objects. This study does not implement the published Role Diversity, policy-churn, or neural-diversity diagnostics under a shared protocol; Table I records their stated scope and instrumentation requirements.

Closest in spirit is Yamaguchi [32], who identifies coordinated-stable, fragile, and jammed/disordered regimes using cooperative success rate and TD-error variance. That method uses two training-time axes, whereas CCI reads Q-value dispersion from a frozen checkpoint. The present study targets value decomposition and evaluates LBF and RWARE, but the observed cross-domain split is associative rather than causal. Concurrent work reports that similar return curves can conceal distinct coordination mechanisms [33]. Table I is a qualitative positioning table, not evidence that CCI outperforms Role Diversity, phase-regime analysis, entropy, churn, or Q-gap diagnostics. Joint TD-error and final return are the matched main-protocol alternatives; a post-hoc retained-checkpoint comparison with entropy, Q-gap, and two behavioral summaries is reported separately in Section V-A. Cross-domain support is marked Partial because only the QTRAN near-zero-dispersion pattern transfers unchanged to RWARE. The all-Yes entries for CCI are therefore scoped to the properties listed and do not constitute an exhaustive priority claim.

TABLE I. C (CCI) VS. PRIOR COORDINATION DIAGNOSTICS

Diagnostic	Param-free	No instr.	VD-collapse	Cross-domain	Held-out
Role Diversity [21]	Yes	Yes	No	Yes	No
Phase regimes [32]	Yes	No	No	No	No
Joint TD-error (Q_{tot}) [34]	Yes	No	No	–	No
Final return [19]	Yes	Yes	No	–	No
C (CCI, this work)	Yes	Yes	Yes	Partial	Yes

III. METHODOLOGY

The coordination diagnostic C has no learned parameters and is extractable from a frozen checkpoint. It is designed to discriminate the VDN/QMIX/QTRAN calibration triad on $\bar{C}$, with a return-based check retained for class-complete IGM mixers such as QPLEX. Its 10%-budget use is evaluated retrospectively on retained LBF QMIX/QTRAN checkpoints and is treated as a candidate screen rather than a validated stopping policy.

A. Setting and the Value-Decomposition Family

Cooperative MARL is cast as a Dec-POMDP under Centralized Training with Decentralized Execution (CTDE) [1],

[2], [35]: joint value decomposition is trained with global state and histories available, while each agent executes from its own history τ_i via $Q_i(\tau_i, u_i)$. An action-value decomposition satisfies the Individual-Global-Max (IGM) condition [2], [3] when:

$$\arg \max_{\mathbf{u}} Q_{\text{tot}}(\boldsymbol{\tau}, \mathbf{u}, s) = \left(\arg \max_{u_i} Q_i(\tau_i, u_i) \right)_{i=1}^n. \quad (1)$$

Eq. (1) licenses decentralized greedy execution. The value-decomposition family realizes it with decreasing mixing restrictiveness: VDN [1] sums per-agent utilities; QMIX [2] uses a state-conditioned mixer with non-negative weights (monotonicity $\partial Q_{\text{tot}}/\partial Q_i \geq 0$); QTRAN [3] drops monotonicity and approximates the full IGM-consistent class via auxiliary losses; QPLEX [4] attains the *complete* IGM function class with a duplex-dueling architecture. This VDN→QMIX→QTRAN→QPLEX axis is the structural ladder along which coordination behavior is measured below.

B. CCI Formula and Components

Given the IGM-consistent value-decomposition family above, a frozen checkpoint permits inspection of whether per-agent greedy Q-values remain numerically heterogeneous or become nearly identical. CCI reads that numerical property through the variance of per-agent greedy-action Q-values, passed through a tanh wrapper. It does not by itself identify behavioral roles, coordination success, or a causal failure mechanism. The **Coordination Complexity Index (CCI)**, denoted C , is:

$$C = \tanh(\text{Var}(q_1, \dots, q_n)) \quad (2)$$

where, $\mathbf{q} = (q_1, \dots, q_n)$ is the vector of per-agent greedy-action Q-values at a given timestep and Var is the unbiased sample variance over the n agents (dividing by $n-1$; ddof=1). That is the estimator which produced every C value reported in this study, and reproducing the reported numbers requires it. The `macci` implementation uses the same estimator, so the package and the reported values agree by construction. The population variance (ddof=0) differs only by the constant factor $n/(n-1)$ on the pre-tanh argument, but substituting it under the same fixed thresholds is not label-neutral: two of the twenty-five LBF seeds cross the STAGNATION boundary (VDN 23/25 → 21/25), while the cohort-level signature of each algorithm is unchanged.

High C indicates dispersed per-agent utilities; it does not by itself imply role differentiation, adaptation, or success. Likewise, $C \approx 0$ denotes nearly identical greedy utilities under the realized scale, not every possible form of policy collapse. For illustration, $\mathbf{q} = (2.0, 2.0, 2.0)$ gives $C = 0$, whereas $\mathbf{q} = (0.5, 2.0, 3.5)$ gives $C = \tanh(1.5) \approx 0.91$. A common additive shift $q'_i = q_i + b$ leaves C unchanged, but a common positive rescaling gives $C' = \tanh(a^2 \text{Var}(\mathbf{q}))$ and can change every finite threshold decision. Agent-specific action-independent offsets can also preserve each agent's greedy policy while changing inter-agent variance. Thus the tanh wrapper bounds $C \in [0, 1)$ but does not make it policy- or scale-invariant.

The headline scalar C above is one component of a richer per-timestep (I, C, T) triple logged by the artifact. Here $I = \sigma(\text{mean}_{i \neq j} |(q_i - \bar{q})(q_j - \bar{q})|)$ is an interdependence proxy computed from pairwise products of centered selected-action values; it is not softmax entropy. C is the Eq. (2) scalar, and T is an L2 trajectory-change term. T is zero in the post-hoc single-checkpoint extraction because no preceding trajectory is supplied. I and C are non-orthogonal functionals of the same selected-action values, so the component analysis below is descriptive rather than a comparison of independent diagnostics.

C. Signature Extraction Pipeline

The signature-extraction routine records C at every timestep of greedy evaluation episodes ($\epsilon = 0$). A timestep is LOW if $C < \theta_{\text{lo}}$ and HIGH if $C > \theta_{\text{hi}}$; the neutral band is unmarked. A regime transition is counted when consecutive timesteps lie on opposite sides of θ_{lo} , and the reported rate is the episode mean. The operational classifier is applied in order: COLLAPSE if $\bar{C} < 10^{-3}$; STAGNATION if $\max(\text{High}\%, \text{Low}\%) > 80\%$; and BI-REGIME otherwise. These names preserve the registered analysis vocabulary. Occupancy ignores temporal order, so STAGNATION means one-band occupancy and BI-REGIME means dual-band occupancy; neither label alone establishes halted adaptation or persistent oscillation. Transitions are reported separately and do not enter classification. The constants were calibrated on the LBF VDN/QMIX/QTRAN triad and were not re-tuned per environment. “No learned parameters” refers only to C , which has no fitted weights or auxiliary loss.

1) *Threshold robustness*: The cohort label is unchanged across five $(\theta_{\text{lo}}, \theta_{\text{hi}})$ settings: $(0.03, 0.07)$, $(0.04, 0.08)$, $(0.05, 0.10)$, $(0.08, 0.15)$, and $(0.10, 0.20)$. QTRAN is COLLAPSE in all 25/25 seed-by-threshold cells because occupancy thresholds do not enter that decision. A sensitivity check of collapse cut-offs $\{10^{-4}, 5 \times 10^{-4}, 10^{-3}, 2 \times 10^{-3}, 10^{-2}\}$ also leaves every canonical QTRAN seed below and every VDN/QMIX seed above. This analysis varies numerical cut-offs on the same calibration triad; it does not validate the rule form on an independent calibration set. VDN is STAGNATION in 23/25 cells and QMIX is BI-REGIME in 17/25, so the non-collapse boundary is cohort-level rather than seed-universal. The retained sweep enables the same sensitivity check on new checkpoints. A separate post-hoc rescaling audit applies $q' = aq$ exactly through the retained pre-tanh variances. At $a = 0.1$, three of five QMIX seeds cross the collapse cut-off even though greedy actions are unchanged; at $a = 10$, all five QMIX seeds become one-band high-occupancy cases. QTRAN remains below the collapse cut-off for all five seeds over $a \in [0.1, 10]$ because its observed variance is much smaller. This finite-range robustness of the reported QTRAN cohort does not restore policy invariance to the statistic or its thresholds.

D. Dual Evaluation Protocol

The tanh normalization in Eq. (2) assumes $|Q| \lesssim 1$, which holds for environments with bounded episodic return and short horizons. On reward-unbounded environments where $|Q|$ can exceed unity by an order of magnitude or more, C saturates to one regardless of inter-agent coordination structure. Two layers

of diagnostic are therefore defined. Layer 1 is the canonical C of Eq. (2). Layer 2 is its un-bounded counterpart:

$$\text{RawVar}(Q) = \text{Var}_i(\max_{u_i} Q_i(\tau_i, u_i)), \quad (3)$$

and is reported together with the per-state $\max_i Q_i - \min_i Q_i$ range; applying the tanh wrapper to Eq. (3) recovers Layer 1, so the two layers differ only in whether the dispersion is compressed. Each layer targets the same numerical property—dispersion among per-agent selected-action values—in the regime where its observable remains informative. All headline tables use Layer 1 on the bounded-reward LBF and RWARE cohorts; Layer 2 is defined for reward-unbounded regimes where tanh saturates, and is not reported as an empirical probe in this study. Signature extraction runs a fresh-evaluation protocol on the frozen checkpoint under evaluation: 100 greedy episodes ($\epsilon = 0$), reduced to 50 for some early-checkpoint extractions (Table VII), with evaluation seeds independent of the training seeds. It therefore depends on no quantity logged during training.

E. Algorithms and Experimental Setup

This work studies three primary value-decomposition algorithms: VDN, QMIX and QTRAN. QPLEX, held out from the design of C , is *class-complete*: it realizes the complete IGM function class as a hard architectural constraint rather than through auxiliary penalties. The three primary algorithms train from one shared hyperparameter configuration on the LBF grids at a learning rate of 5×10^{-4} . Two things are not shared: the optimiser is Adam for VDN and QMIX and RMSprop for QTRAN [36], [37], and the QTRAN learner does not consume every key of that configuration (it ignores `td_lambda`, the leniency keys, `q_value_clip` and `standardise_rewards`). The mixer is therefore the primary intended per-algorithm difference, not the only realized one. The RWARE cohorts run to 2×10^7 steps. Every setting is reported in full, as standardised evaluation requires [19]; the shared table is Table V (batch size / replay-buffer 128 / 5000 episodes; agent network Feed-forward MLP, hidden dim 64). One setting is *not* shared: the optimiser. Rather than leave that optimiser asymmetry as a confound, a matched QTRAN-with-Adam control ($n=5$ seeds) was run to separate the mixer from the optimiser; without it, QTRAN would have been both the sole COLLAPSE class and the sole RMSprop run, and the two would not have been separable. The shared hyperparameters follow Hu et al.'s implementation discipline [6], using code-level tricks that are not specific to QMIX. As executed they are not generic across the family either: the QTRAN learner ignores them.

1) *Held-out prediction H1*: Prediction H1 is an exploratory, falsifiable conjecture tied to the training budget. Let $\mathcal{A}$ realize the complete IGM function class through a hard architectural constraint. Under assumptions (A1)–(A3), H1 predicts that every seed reaches $\bar{C} < 10^{-3}$ by the fixed 10M-step checkpoint on canonical LBF. The assumptions are not identified from the reported data or derived in general, so H1 is neither a theorem nor a convergence claim. VDN, QMIX, and QTRAN supplied the taxonomy and thresholds; QPLEX is only a held-out falsification test of that calibration.

The working assumptions are:

- (A1) parameter-shared agent networks with a symmetric team reward $r(s, \mathbf{u})$;
- (A2) the TD fixed point under parameter-shared training admits the additive-plus-shared decomposition $\max_{u_i} Q_i(\tau_i, u_i) = f(\tau_i) + c(s)$ for a shared function f and a state-level residual $c(s)$ (a structural assumption, not derived here); and
- (A3) a gradient-descent training-limit regime under which parameter-shared updates pull the shared function f of (A2) toward a common attractor on the visited history distribution (a regularity assumption, not derived here, in the spirit of the gradient-flow analysis of T. Hu et al. [38]). The agent networks receive a one-hot agent identifier (Table V), which breaks exact exchangeability of the local histories by construction; (A3) is therefore a statement about the learned representation in the training limit, not about the raw histories. In the configuration studied here the agent network is feed-forward (`use_rnn=False`; Table V), so the realized per-agent function is $Q_i(z_i, u_i)$ over the current observation; τ_i is retained for consistency with the CTDE formalism.

The motivation is heuristic. Under (A1), shared TD targets [2], [3] and (A2) yield $\max_{u_i} Q_i = f(\tau_i) + c(s)$, hence $\text{Var}_i(\max_{u_i} Q_i) = \text{Var}_i(f(\tau_i))$. Only if representation sharing under (A3) dominates should this variance contract. These assumptions are neither identified empirically nor sufficient for a general proof. Consequently, H1 is evaluated solely as a budget-anchored conjecture: one QPLEX seed with $\bar{C} \geq 10^{-3}$ at 10M steps rejects it.

2) *Environments and seed counts*: The primary benchmark is Level-Based Foraging (LBF) [34], [39]. Five primary 5-agent grids (8×8 through 12×12) plus a 6-agent 9×9 robustness probe are tested, all with three food items, `force_coop=True`, and sight radius two. Those settings are denoted by the compact labels 9×9 -5p-3f-coop and 9×9 -6p; the 9×9 -5p grid on which C , its thresholds, and the held-out test are defined is hereafter the canonical grid. The seed count of every reported cohort is stated explicitly, and the per-seed records are available from the corresponding author on reasonable request. On LBF, VDN/QMIX/QTRAN each contribute 5 matched seeds per grid, so the headline taxonomy rests on the 25 QTRAN seeds across the five primary grids. VDN and QMIX were run with more seeds, but the shared 5 form the matched reported cohort, consistent with the 5-vs-5 effect sizes of Table III. The 6-agent 9×9 -6p variant likewise reports 5 seeds per algorithm, and the held-out QPLEX test uses 5 seeds on the canonical grid. Cross-domain validation on RWARE [40] (the four-agent `rware-small-4ag` layout) uses 5 seeds per algorithm and 20M environment steps under the parallel runner. The cohort is matched on seed count and budget but not on the leniency schedule or batch size (32 for VDN and QTRAN against 128 for QMIX). A second matched $n=5$ cohort is retrained under the episode runner, which collects one full episode at a time. Reported returns are undiscounted, environment-specific episodic metrics [40]: on LBF the fraction of food items collected per episode, on RWARE the number of successful deliveries per episode. LBF

TABLE II. THREE COORDINATION SIGNATURES ON LBF 9×9-5p

Alg.	Signature	$\bar{C}$	Tr/ep	Lo%	Hi%	Return	95% CI
VDN	Stagnation	0.483	1.2	3.9	91.8	0.632	[0.55, 0.71]
QMIX	Bi-regime	0.172	2.4	22.8	57.2	0.732	[0.70, 0.76]
QTRAN	Collapse	3.3×10^{-6}	0.0	100.0	0.0	0.000	[0.000, 0.000]

and RWARE return magnitudes are therefore not on a common scale.

3) *Statistical protocol*: Effect sizes use Cohen's d [41] with the Hedges' g small-sample correction [42]

$$J(df) = 1 - \frac{3}{4(n_1 + n_2) - 9}. \quad (4)$$

Confidence intervals are percentile bootstrap intervals over the per-seed values, computed directly in NumPy with a fixed seed: 10,000 resamples for effect-size intervals and 50,000 for return intervals and $P(X > Y)$; the summaries reported—the interquartile mean (IQM), optimality gap, and $P(X > Y)$ —follow the reporting practice argued for by Agarwal et al. [43], but the stratified-bootstrap procedure of the `rliable` library itself was not used. Where $n \leq 3$ makes $J(df)$ unreliable, the row is footnoted. Wide bootstrap CIs from near-zero within-group variance (collapsed conditions) are reported unadjusted. All seed-level statistics (CIs, effect sizes) are computed over the per-seed $\bar{C}$ values—each seed's $\bar{C}$ being the mean of C over all timesteps pooled across its evaluation episodes—so the reported spread reflects across-seed reproducibility.

IV. RESULTS

Results cover the canonical signatures, a frozen-cut-off extension on five additional QTRAN training seeds, cross-setting robustness, an optimiser control, generalization to unseen architectures, and a held-out evaluation on QPLEX.

A. Three Coordination Signatures

The CCI extraction routine uses 100 greedy episodes per seed on canonical LBF. VDN receives the operational STAGNATION label because HIGH occupancy is approximately 92%; this indicates a stable one-band Q-value profile, not verified behavioral stagnation or fixed roles. QMIX receives BI-REGIME because trajectories occupy both bands, with approximately 2.4 threshold crossings per episode; the statistic does not establish what agents do on either side of the threshold. QTRAN receives COLLAPSE because greedy Q-values are nearly identical and LOW occupancy is 100%. Its seed-mean $\bar{C}$ is nearly five orders of magnitude below QMIX (Table II). Returns and bootstrap intervals use the matched five-seed cohort. Return and C are aligned by training seed but come from training logs and separate greedy checkpoint evaluations, respectively; they are not paired episode-level observations. Canonical QMIX uses last-saved rather than strictly end-of-budget checkpoints, as disclosed in Section V-C.

These operational labels are cohort-level summaries rather than artifacts of one threshold setting. Per-seed stability is exact for COLLAPSE and majority-stable for the other labels: VDN is STAGNATION in 23/25 cells and QMIX is BI-REGIME in 17/25 across the five threshold pairs. Attribution

to the mixer rather than the optimiser rests on Section IV-D. Table X adds a descriptive seed-0 cross-checkpoint trace. VDN transitions fall from 6.60 at 2.5M steps to 0.14 at 5.0M and 7.5M; QMIX remains between 2.38 and 4.66 across six checkpoints; QTRAN remains at zero across six checkpoints. This limited trace preserves temporal ordering for one retained seed but is not a cohort-level temporal validation.

Neither final return nor joint TD-error identifies the observed Q-value-dispersion pattern as directly as C . At the 10% canonical-LBF checkpoint, the retrospective candidate rule $\bar{C} < 10^{-3}$ flags all five retained QTRAN seeds and none of the five retained QMIX seeds. VDN checkpoints were not retained there, so the evidence covers one grid and two algorithms. The flagged LBF seeds did not recover by the retained final checkpoint. On RWARE, however, two of three retained early QMIX checkpoints are flagged, including at least one seed that later succeeds. The same cut-off therefore has an environment-induced false positive and is not a validated universal stop rule.

1) *Frozen-cut-off seed extension*: After fixing the rule and evaluation protocol, five additional canonical-LBF QTRAN training seeds (5–9) were run under the same 10M-step parallel protocol. Under the same 100-episode greedy harness, 5/5 seeds fall below 10^{-3} at the nearest retained 10%-budget checkpoint (cohort mean $\bar{C} = 3.93 \times 10^{-6}$), and 5/5 do so near 10M steps (mean $\bar{C} = 5.27 \times 10^{-6}$). Fresh-evaluation mean returns are 0.001 and 0.006, respectively. Table XII reports every seed. Because this extension adds only QTRAN, it tests within-architecture repeatability of the positive class; it neither recalibrates the cut-off nor supplies new negative-class specificity evidence.

B. Statistical Separation

The three architecture cohorts on the canonical grid separate statistically on seed-level $\bar{C}$. Table III reports pairwise effect sizes on seed-level $\bar{C}$ (Cohen's d with the Hedges' g correction of Eq. (4); bootstrap CIs on d from the same unstratified percentile bootstrap, 10,000 resamples with a fixed seed). VDN-versus-QTRAN separates at $|d| = 13.46$ ($g = 12.15$), a magnitude that reflects QTRAN's near-zero within-group variance under collapse. Two of the three pairs exceed $|d| = 2.0$, with QMIX-versus-QTRAN ($d = 1.73$) just below. This holds at the point estimate; at $n = 5$ the QMIX-versus-VDN interval extends to $|d| = 1.74$, so the $|d| = 2.0$ boundary is unresolved for that pair, as it is for QMIX-versus-QTRAN, whose interval runs from 1.36 to 10.18. All three intervals exclude zero, so the direction of every pair is resolved, but each spans a factor of six or more in $|d|$, so at $n = 5$ this sample fixes the sign of each separation without fixing its size. The widest CI, [9.69, 61.44] on VDN-versus-QTRAN, reflects QTRAN's near-zero within-group variance under bootstrap rather than point-estimate instability, and is reported transparently rather than narrowed via a non-standard estimator. On the same matched $n = 5$ cohort, final return separates QMIX from VDN at $d = +1.60$ and joint TD-error at $d = -1.31$; Discussion interprets these margins against C .

These pairwise effects generalize to a population-level summary. Comparing non-collapsed $\bar{C}$ on the canonical 9×9-5p cohort (ten VDN/QMIX seed-mean values; geometric mean

≈ 0.252) against collapsed $\bar{C}$ pooled across QTRAN on the canonical 9×9 -5p grid and QTRAN on cross-domain RWARE small-4ag (geometric mean $\approx 5.59 \times 10^{-6}$) defines the pooled contrast. The geometric-mean separation is approximately 4.5×10^4 , or about 4.66 orders of magnitude relative to that non-collapsed foraging geometric mean. Only QTRAN's RWARE values enter this collapsed pool; the VDN/QMIX RWARE cohort stays non-collapsed (below). The worst-case two-point extremes ratio—the smallest non-collapsed $\bar{C}$ (0.0613, QMIX seed 3) over the largest collapsed one (1.92×10^{-4} , RWARE QTRAN seed 2)—is 3.2×10^2 , about 2.50 orders of magnitude. It is a two-point statistic; leaving out either extreme moves it only to 4.0×10^2 or 4.3×10^2 . The retained per-seed checkpoint records used here are available from the corresponding author on reasonable request.

TABLE III. PAIRWISE EFFECT SIZES ON SEED-LEVEL $\bar{C}$ (LBF 9×9 -5P)

Pair	Cohen's d	Hedges' g	95% CI on d	n seeds
QMIX vs QTRAN	1.73	1.56	[1.36, 10.18]	5 vs 5
VDN vs QTRAN	13.46	12.15	[9.69, 61.44]	5 vs 5
QMIX vs VDN	-2.93	-2.65	[-16.42, -1.74]	5 vs 5

Across the five 5-agent grids the per-component breakdown of I , C [Eq. (2)], and T attributes 93–98% of the attributable within-cohort variance to C , versus 2–7% to I . T is structurally zero under single-checkpoint extraction—temporal drift needs a previous trajectory—so it cannot explain variance here; that is why C is reported as the single headline diagnostic. On the canonical 9×9 -5p grid the VDN/QMIX/QTRAN triad alone spans 0.483 on C (comparable to the other grids); including the held-out QPLEX raises the four-architecture C spread to 0.541 versus 0.163 on I (a $3.3 \times C/I$ ratio that mixes the calibration triad with the held-out test). Because I and C are non-orthogonal functionals of the same Q-vectors, this dominant share is a descriptive corroboration of the C/I ratio rather than an orthogonal variance analysis.

C. Cross-Setting Robustness

The canonical-grid finding replicates across five LBF grid sizes and a six-agent variant (per-grid QMIX/VDN $\bar{C}$ and returns in Table VIII). VDN's mean $\bar{C}$ stays above 0.47 on all five 5-agent grids and on the 6-agent variant; QMIX's ranges from 0.083 at 8×8 to 0.208 at 10×10 across the five 5-agent grids, rising to 0.220 on the 6-agent variant; QTRAN's remains below 10^{-3} on all five 5-agent grids and on the 6-agent variant. On all six configurations, C identifies a near-zero-dispersion QTRAN pattern that return alone does not distinguish, since QTRAN's mean return ranges from exactly zero on the canonical grid to 0.029 (0.000–0.029) rather than being uniformly zero (the six-agent case is detailed below). The QMIX-versus-VDN effect size on $\bar{C}$ is non-monotonic with an intermediate-grid-size minimum: $|d_C|$ drops from 10.10 at 8×8 to 1.47 at 10×10 , then rebounds to 12.78 at 12×12 . The intermediate minimum is driven by a single QMIX seed at $\bar{C} = 0.660$, which raises the QMIX across-seed standard deviation at 10×10 to 0.25 against 0.05 for VDN; at the two extremes both cohorts are tight (across-seed standard deviations between 0.022 and 0.056), so the pooled standard deviation is small and $|d_C|$ is correspondingly large. The 9×9 -5p effect size is $d = -2.93$ ($g = -2.65$) at

$n = 5$ versus 5. On returns, the QMIX-vs-VDN gap is 9.1–16.9 pp across grids (non-monotonic, peaking at 10×10), with $P(\text{QMIX} > \text{VDN}) \geq 0.91$ on every grid (the smallest, 0.92 on 10×10 , is 0.9186 before rounding). All reported cross-grid effect sizes and returns use $n = 5$ seeds per algorithm on every grid, the six-agent 9×9 -6p variant included, with two exceptions: the 9×9 $P(X > Y)$ uses six QMIX seeds, and on the 6p variant the $\bar{C}$ and return columns do not share the same five seeds.

Across the five 5-agent grids, every tested QTRAN seed remains below the collapse cut-off with narrow within-grid variance. This replication covers the five grids and seeds tested; it does not establish invariance outside them. The 6-agent variant on 9×9 preserves VDN's stagnation and QMIX's bi-regime under increased agent count, with QTRAN in full coordination-variance collapse ($\bar{C} < 10^{-3}$) even though its return (0.029) is marginally above zero.

One of the three canonical-grid signatures replicates cross-domain on the RWARE robotic warehouse benchmark [40], a structurally distinct cooperative environment with corridor navigation and sparse reward ($n = 5$ matched training seeds per algorithm). The QTRAN near-zero-dispersion classification recurs under the same numerical rule: seed-mean $\bar{C} \approx 9 \times 10^{-5}$, mean return 0.000 across all five RWARE seeds, and QTRAN is the only algorithm below the $\bar{C} < 10^{-3}$ collapse threshold. The two non-collapsed algorithms stay above it: QMIX learns the task (mean return 0.86) yet its coordination-variance is compressed to $\bar{C} \approx 3 \times 10^{-3}$ (only about $3 \times$ above the collapse cut-off, unlike its LBF bi-regime), and VDN shows LOW-occupancy STAGNATION ($\bar{C} \approx 0.029$, Low% $\approx 93\%$), differing from its LBF HIGH-occupancy stagnation. Both non-collapsed algorithms therefore sit at low, near-uniform $\bar{C}$ on RWARE, QMIX falling from its LBF $\bar{C} \approx 0.17$. The LBF QMIX-vs-VDN separation therefore shrinks substantially here (Cohen's $d = -1.11$, still a large effect but well below the LBF value), while the collapsed QTRAN class stays sharply separated from QMIX ($d = -2.77$). QTRAN's separation from VDN ($d = -1.23$) is comparable in magnitude to the QMIX-vs-VDN effect, because RWARE compresses VDN's own $\bar{C}$ toward the threshold. Fig. 1 contrasts the recurring QTRAN near-zero-dispersion pattern—below the 10^{-3} threshold in both domains—with the domain-dependent VDN/QMIX change. This contrast is associative, not a causal attribution to architecture. A matched-seed episode-runner replication of this four-agent cohort ($n=5$) shows the same near-zero-dispersion classification, so the observation is not limited to the parallel-runner evaluation. Under a common 100-episode greedy evaluation harness, QTRAN's geometric-mean $\bar{C} \approx 1 \times 10^{-8}$ with return 0.000 on every seed, whereas VDN does not collapse (geometric-mean $\bar{C} \approx 0.44$, mean return 3.23). Returns are not directly comparable with the parallel-runner figures above: that cohort used the episode runner under a 50-step training time limit rather than the parallel cohort's 500-step limit. Non-collapsed $\bar{C}$ magnitudes are harness-sensitive; only the within-harness QTRAN-versus-VDN contrast is claimed here.

D. Mixer–Optimiser Control

QTRAN was retrained from scratch on the canonical 9×9 -5p grid with the *same* optimiser the non-collapsed baselines use—Adam at learning rate 5×10^{-4} . It was evaluated under

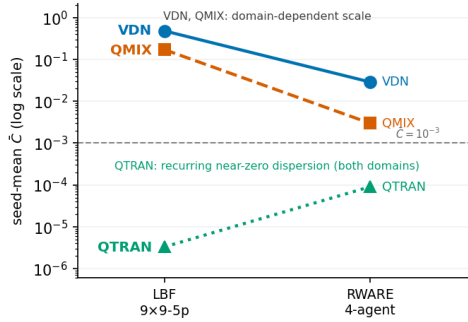


Fig. 1. Seed-mean $\bar{C}$ on LBF 9x9-5p vs. RWARE (log axis; 10^{-3} threshold).

the same 100-episode greedy harness, holding every other setting identical (runner, seeds 0–4, and the 10M-step budget). The optimiser is fixed by the learner implementation rather than by the configuration, so this matched control tests it directly. It does not isolate the mixer: the QTRAN learner also ignores four further keys that the VDN/QMIX learner consumes, and those differences are present here too.

The cohort-mean near-zero-dispersion classification persists. At 10M steps the QTRAN-with-Adam cohort reaches seed-mean $\bar{C} = 6.9 \times 10^{-4}$, below the $\bar{C} < 10^{-3}$ cut-off and far below both non-collapsed baselines on the same grid: 2.8 orders of magnitude below VDN (0.483) and 2.4 below QMIX (0.172). The regime signature is unchanged from the RMSprop cohort on every seed: Low% = 100, High% = 0, and zero regime transitions per episode. At the 10M-step budget, giving QTRAN the baselines' optimiser does not give it their coordination structure.

The optimiser is nevertheless not inert. The swap raises seed-mean $\bar{C}$ from 3.3×10^{-6} to 6.9×10^{-4} , a factor of about 209, with complete separation between the two cohorts: the lowest Adam seed (2.4×10^{-4}) exceeds the highest RMSprop seed (7.8×10^{-6}) by more than an order of magnitude. One Adam seed reaches 1.6×10^{-3} and thus sits above the cut-off; a checkpoint sweep over the final quarter of training shows two near-cut seeds alternating across the cut-off rather than settling on either side, so their class assignment reflects where training stopped rather than a distinct behavior. The cohort-level verdict is stable under that ambiguity: recomputed at each of the twelve final checkpoints, the seed-mean stays below the cut-off in every case.

The near-zero-dispersion pattern is a *learned* network state, not present in the untrained network: at the earliest logged checkpoint (200 environment steps, effectively untrained) every seed sits above the cut-off ($\bar{C} \approx 2 \times 10^{-3}$ – 1.4×10^{-2}), and training drives $\bar{C}$ down by three to four orders of magnitude before the partial recovery just described. The near-zero $\bar{C}$ of this cohort therefore reports a learned numerical state rather than an initialization artifact. One boundary is stated explicitly: the claim is anchored at 10M steps—unlike the RMSprop cohort, whose $\bar{C}$ is flat, the Adam cohort's $\bar{C}$ is still rising at the end of training on all five seeds, so the result bounds behavior at this budget rather than at convergence.

E. Generalization to Unseen Architectures

The taxonomy only partially generalizes beyond the four studied architectures. On three 2025 value-decomposition methods (QFix [9], QTypeMix [8], DuelMIX [7]; five matched seeds each, seeds 0–4, at the same 10M-step budget; Table IX), $\bar{C}$ separates all 12 successfully-trained (B1-REGIME) seeds from all 3 failed (STAGNATION) seeds, with no seed falling below the $\bar{C} < 10^{-3}$ collapse cut-off. The three failed seeds are the informative case: their zero return is indistinguishable from QTRAN's on the reward curve, yet $\bar{C}$ separates their numerical patterns: a uniform-high- $\bar{C}$ one-band case ($\bar{C} \in [0.84, 0.99]$) rather than QTRAN-like near-zero dispersion. This difference does not identify the underlying behavioral or causal mechanism. Only two of the three signatures manifest.

F. Held-Out Evaluation on QPLEX

QPLEX is a duplex-dueling architecture not used to design $\bar{C}$, choose its thresholds, or fit any free parameter of the diagnostic [4]. Across five QPLEX training seeds on the canonical 9x9-5p-3f-coop LBF grid under the canonical protocol (time_limit=50, $t_{\max} = 10M$), test return is 0.000 for every seed at the 10M-step checkpoint (Table IV); that 5/5 return-zero statement is protocol-tied (footnote). The corresponding $\bar{C}$ values are heterogeneous: three exceed 0.7, one is approximately 0.10, and one is 0.036; none falls below the 10^{-3} collapse cut-off. Return collapse and coordination-variance collapse therefore dissociate: QPLEX has zero return on 5/5 seeds without reproducing QTRAN's low- $\bar{C}$ signature. The QPLEX pattern is treated as an architecture-specific heterogeneity mode that may refine the taxonomy, not as evidence for a confirmed fourth signature.*

TABLE IV. HELD-OUT QPLEX PER-SEED $\bar{C}$ AND RETURN (LBF 9x9-5P)

Seed	$\bar{C}$	Return
0	0.956	0.000
1	0.705	0.000
2	0.102	0.000
3	0.036	0.000
4	0.906	0.000

All five seeds fail Prediction H1 under the assumptions stated in the Methodology. H1's sole prediction, operationalized as per-seed $\bar{C} < 10^{-3}$ at 10M steps, does not hold. As the dissociation above shows, the QTRAN-calibrated threshold does not detect every form of policy failure. Per-seed values for the other reported cohorts are available from the corresponding author on reasonable request. The cohort-wide return-collapse at 5/5 is an outcome H1 does not predict and is reported as an observation, not as partial support for H1.

*Per-seed origin: the QPLEX registry contains canonical (time_limit=50, $t_{\max} = 10M$) runs for all five seeds plus non-canonical (time_limit=100, $t_{\max} = 5M$) runs for seeds 0–4; on the non-canonical sub-cohort 1/5 seeds reached return-collapse and 4/5 retained mid-range returns (0.35–0.59). The 5/5 return-collapse statement therefore applies only to the canonical 10M-step cohort.

V. DISCUSSION

A. Interpretation and Checkpoint Reading

The three signatures support a scope-controlled checkpoint reading (Table VI). The only candidate action evaluated retrospectively is COLLAPSE→consider stopping on canonical LBF. The other mappings are interpretive prompts, not tested interventions. Even the collapse action requires domain-specific validation because it produces a false positive on RWARE and misses QPLEX's return-collapse mode. The constraint-performance ordering is consistent with prior evidence on monotonicity [6], but the present observations do not establish a causal mechanism.

A practitioner should not abort solely because $\bar{C} < 10^{-3}$. On canonical LBF, that one-million-step rule would have saved nine million additional steps for the retained QTRAN seeds without flagging retained QMIX seeds. Before operational use it requires validation on the target domain, algorithm family, Q-value scale, and checkpoint protocol. RWARE false positives, missing VDN early checkpoints, and the absence of an early QPLEX extraction delimit the evidence.

Joint TD-error and final return are evaluated under the matched main protocol. A further post-hoc retained-checkpoint evaluation uses the same 100-episode harness for C , normalized softmax action entropy, the top-1–top-2 Q-gap, pairwise greedy-action disagreement, and within-episode action switching (Table XI). The retained checkpoints are not horizon-matched: VDN is at approximately 7.5M steps, QMIX at 7.5–9.5M, and QTRAN at approximately 10M. Entropy lies near its ceiling for all three cohorts (0.9925–0.9999995), while the Q-gap, disagreement, and switch summaries also distinguish QTRAN from VDN/QMIX. For QTRAN versus QMIX, the absolute seed-level Cohen's d values are 1.86 for C , 2.38 for entropy, 3.71 for Q-gap, 4.47 for disagreement, and 22.35 for switching; at $n = 5$ with unequal checkpoint horizons, these are descriptive point estimates. These associations supply behavioral context but do not show that C measures roles or causes success. Against the two matched main-protocol scalars, C separates canonical QMIX from VDN most strongly at $n = 5$ ($d = -2.93$, versus $+1.60$ for final return and -1.31 for TD-error; Table III). The post-hoc additions are not the published Role Diversity or policy-churn diagnostics, and no superiority over those methods, phase-regime analysis, or neural diversity is claimed.

B. Scope and Transfer

Two descriptive cross-domain observations emerge. QTRAN's near-zero-dispersion pattern appears on LBF, on RWARE under two runners, and under the optimiser control. In contrast, VDN and QMIX magnitudes and occupancy patterns change between LBF and RWARE. These associations are consistent with an architecture-related QTRAN pattern and environment-dependent non-collapse patterns, but the study does not identify the causal contribution of architecture, environment geometry, representation sharing, or reward scale. The LBF/RWARE comparison therefore motivates, rather than proves, a structural explanation and requires domain-specific calibration.

C. Limitations and Reproducibility

The QTRAN near-zero-dispersion claim rests on 25 LBF training seeds across the five 5-agent grids; the underlying per-seed records are available from the corresponding author on reasonable request. The pooled geometric-mean separation of more than four orders of magnitude, on the pool defined above, is the headline; the worst-case two-point extremes ratio reported there is about two orders of magnitude smaller. Layer-1 C targets bounded-reward regimes; Layer-2 $\text{RawVar}(Q)$ is intended where $\tanh$ saturates. A further bound is scale identifiability: under monotonic mixing, intermediate cross-architecture $\bar{C}$ magnitudes are read on the realized training scale rather than as scale-invariant distances, while exact zero variance remains zero under finite common rescaling. The finite rule $\bar{C} < 10^{-3}$ is not invariant: the rescaling audit above changes QMIX labels without changing greedy policies.

The canonical QMIX $\bar{C}$ arm is not strictly horizon-matched: its five seeds use last-saved checkpoints (9.50M, 9.50M, 10.40M, 9.75M, and 9.75M), seed 2 interrupted at 10.40M. Omitting it moves $\bar{C}$ from 0.172 to 0.196 and $|d_C|$ versus VDN from 2.93 to 2.71; BI-REGIME and $|d| > 2$ hold either way; no horizon-matched 9–10M checkpoint exists on disk for seed 2, so the mismatch is disclosed rather than re-extracted. H1's literal $\bar{C} \rightarrow 0$ prediction does not hold per-seed; the unseen-architecture check is likewise partial; SMACv2 and robotic manipulation were not tested.

Of the 417 training-run attempts in the run index, 349 (84%) ran to at least 95% of their planned timesteps and 36 (9%) stopped before the 10% checkpoint. Because the last group stopped before the 10% checkpoint, no early-checkpoint $\bar{C}$ exists for them, so their exclusion rests on the absence of that checkpoint rather than on a statistical test; the full accounting is recomputed from the retained run index. The Adam control is budget-anchored at 10M steps with cohort-mean reporting. Three of the five seeds sit within a factor of two of the cut-off, and two of them alternate across it between adjacent checkpoints. The QTRAN near-zero-dispersion pattern recurs on all five tested LBF grids, on RWARE, and under the matched Adam control.

Five bounds follow from the design. First, C uses Q-value magnitudes and is not policy- or scale-invariant; $\tanh$ bounds the range but does not remove this dependence. Second, occupancy discards order. The separately reported transition rate and seed-0 cross-checkpoint trace in Table X add temporal description, but one retained seed does not validate temporal labels across the cohort. Third, thresholds and the rule form were calibrated on the same VDN/QMIX/QTRAN triad. Sweeping threshold values tests sensitivity, not independent calibration. Fourth, C is not a coordination-quality score: return and dispersion are not monotone, and QPLEX demonstrates return failure without near-zero dispersion. Fifth, joint TD-error and final return are the matched main-protocol comparisons. The post-hoc retained-checkpoint comparison adds softmax action entropy, top-two Q-gap, greedy-action disagreement, and within-episode switching, but its horizons differ and it does not implement Role Diversity or policy churn. All effect sizes use five seeds per cohort and have wide intervals. The additional QTRAN seeds 5–9 were evaluated only after the cut-off was frozen and therefore strengthen same-architecture repeatability, but they are not an independent cal-

ibration set and include no additional VDN/QMIX specificity controls. Artifacts are listed in the Data and Code Availability statement.

VI. CONCLUSION

This study presented the Coordination Complexity Index (CCI), a checkpoint statistic with no learned parameters that summarizes per-agent greedy-Q dispersion. It separates the reported non-collapsed LBF cohort from the pooled collapsed cohort by more than four orders of magnitude at the geometric mean. On canonical LBF, VDN, QMIX, and QTRAN receive operational one-band, dual-band, and near-zero-dispersion labels. These labels do not by themselves prove stagnation, oscillation, or behavioral roles. QTRAN's pattern recurs on RWARE, while other patterns change across domains. QFix, QTypeMix, and DuelMIX provide partial unseen-architecture evidence. Held-out QPLEX rejects H1 and shows that return collapse need not imply near-zero CCI. The optimiser control shows that the reported QTRAN pattern is learned under the tested setup rather than present at initialization.

Layer-1 C can saturate in reward-unbounded regimes, where Layer-2 $\text{RawVar}(Q)$ is the intended probe. The $\bar{C} < 10^{-3}$ rule is only a retrospectively supported candidate on retained canonical-LBF checkpoints: it has incomplete early-checkpoint coverage and produces a false positive on RWARE. It should therefore prompt targeted inspection or a domain-validated stopping policy, not automatic termination. With the threshold frozen, the additional QTRAN seed extension retains 5/5 positives at 10% and 5/5 near completion, but this one-class repeatability result does not remove the need for independent calibration and negative controls.

ACKNOWLEDGMENTS

This research was conducted at the Faculty of Information Technology and Digital Innovation, King Mongkut's University of Technology North Bangkok. The authors thank the developers of the EPyMARL framework [34], the LBF environment [39], and the RWARE environment [40].

FUNDING

The authors received no specific external grant from public, commercial, or non-profit funding agencies; all experiments were run on the lead author's own workstation.

CONFLICTS OF INTEREST

The authors declare no competing interests.

DATA AND CODE AVAILABILITY

The analysis code, run index, matched seed matrix, threshold and scale-sensitivity sweeps, the held-out QTRAN per-seed evaluation records, and the raw Sacred metrics and configurations used for the reported results are held by the authors and are available from the corresponding author on reasonable request. A permanent public repository and archival DOI will be added after deposit; no inactive URL is asserted here. Reported table values of C and the `macci` implementation both use the unbiased variance estimator ($\text{ddof}=1$), as defined with Eq. (2).

DECLARATION ON GENERATIVE AI TOOLS

During the preparation of this work the authors used Claude (Anthropic) in the writing process, to draft and revise the text, and in the research process, where it assisted in writing the analysis scripts that recompute the reported statistics and generate the figures from logged run output. The learning algorithms are the framework's own implementations, the `macci` implementation is the authors' own work, and the study design, the experiments, and the claims and conclusions are the authors' own; no reported quantity was produced by the model. Every number in this article is derived from the training runs' own logged output, retained checkpoint evaluations, or the run index built over them; the analysis scripts and supporting records are available as stated in the Data and Code Availability statement. The authors reviewed and edited every AI-assisted output and take full responsibility for the content of the published article.

REFERENCES

- [1] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, and T. Graepel, "Value-decomposition networks for cooperative multi-agent learning based on team reward," in *Proceedings of the 17th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2018, preprint arXiv:1706.05296. [Online]. Available: <https://dl.acm.org/doi/10.5555/3237383.3238080>
- [2] T. Rashid, M. Samvelyan, C. Schroeder de Witt, G. Farquhar, J. N. Foerster, and S. Whiteson, "QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.
- [3] K. Son, D. Kim, W. J. Kang, D. E. Hostallero, and Y. Yi, "QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.
- [4] J. Wang, Z. Ren, T. Liu, Y. Yu, and C. Zhang, "QPLEX: Duplex dueling multi-agent Q-learning," in *International Conference on Learning Representations (ICLR)*, 2021.
- [5] S. Dohare, J. F. Hernandez-Garcia, Q. Lan, P. Rahman, A. R. Mahmood, and R. S. Sutton, "Loss of plasticity in deep continual learning," *Nature*, vol. 632, no. 8026, pp. 768–774, 2024.
- [6] J. Hu, S. Jiang, S. A. Harding, H. Wu, and S.-w. Liao, "Rethinking the implementation tricks and monotonicity constraint in cooperative multi-agent reinforcement learning," *arXiv preprint arXiv:2102.03479*, 2023, ICLR 2023 Blog Post.
- [7] E. Marchesini, A. Baisero, R. Bhati, and C. Amato, "On stateful value factorization in multi-agent reinforcement learning," in *Proceedings of the 24th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2025, preprint arXiv:2408.15381. [Online]. Available: <https://dl.acm.org/doi/10.5555/3709347.3743778>
- [8] S. Fu, S. Zhao, T. Li, and Y. Yan, "QTypeMix: Enhancing multi-agent cooperative strategies through heterogeneous and homogeneous value decomposition," *Neural Networks*, vol. 184, p. 107093, 2025.
- [9] A. Baisero, R. Bhati, S. Liu, A. Pillai, and C. Amato, "Fixing incomplete value function decomposition for multi-agent reinforcement learning," *arXiv preprint arXiv:2505.10484*, 2025.
- [10] H. Li, H. Zhou, Y. Zou, D. Yu, and T. Lan, "ConcaveQ: Non-monotonic value function factorization via concave representations in deep multi-agent reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, 2024.
- [11] K. Liu, T. Zhang, X. Xu, and Y. Zhao, "Counterfactual value decomposition for cooperative multi-agent reinforcement learning," *Neural Networks*, 2025.
- [12] D. Yao, C. Sun, K. Li, K. Zhou, H. Li, R. Wang, and L. Liu, "CMIX: Causal value decomposition for cooperative multi-agent reinforcement learning," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2024.

- [13] A. Huang, Y. Wang, R. Liu, H. Zou, and X. Zhou, "QVF: Incorporating quantile value function factorization into cooperative multi-agent reinforcement learning," *Pattern Recognition*, 2025.
- [14] Z. Liu, L. Wan, S. Sun, X. Sui, X. Chen, X. Lan, and N. Zheng, "Enhancing value decomposition with target transformation in cooperative multi-agent reinforcement learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.
- [15] S. Wang, R. Zhang, Y. Zhou, J. Shao, and Y. Cheng, "QuadQ: Quadratic-based value decomposition for cooperative policy optimization in multi-agent reinforcement learning," *IEEE/CAA Journal of Automatica Sinica*, 2026.
- [16] W. Fu, C. Yu, Z. Xu, J. Yang, and Y. Wu, "Revisiting some common practices in cooperative multi-agent reinforcement learning," in *Proceedings of the 39th International Conference on Machine Learning (ICML)*, vol. 162, 2022, pp. 6863–6877.
- [17] L. Wan, Z. Liu, X. Chen, X. Lan, and N. Zheng, "Greedy based value representation for optimal coordination in multi-agent reinforcement learning," in *Proceedings of the 39th International Conference on Machine Learning (ICML)*, vol. 162, 2022, pp. 22 512–22 535.
- [18] G. Papadopoulos, A. Kontogiannis, F. Papadopoulou, C. Pouliaou, I. Kountentis, and G. Vouras, "An extended benchmarking of multi-agent reinforcement learning algorithms in complex fully cooperative tasks," in *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2025, pp. 1613–1622.
- [19] R. Gorsane, O. Mahjoub, R. J. de Kock, R. Dubb, S. Singh, and A. Pretorius, "Towards a standardised performance evaluation protocol for cooperative MARL," in *Advances in Neural Information Processing Systems*, vol. 35, 2022.
- [20] B. Ellis, J. Cook, S. Moalla, M. Samvelyan, M. Sun, A. Mahajan, J. N. Foerster, and S. Whiteson, "SMACv2: An improved benchmark for cooperative multi-agent reinforcement learning," in *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track*, 2023.
- [21] S. Hu, C. Xie, X. Liang, and X. Chang, "Policy diagnosis via measuring role diversity in cooperative multi-agent RL," in *Proceedings of the 39th International Conference on Machine Learning (ICML)*, ser. PMLR, vol. 162, 2022.
- [22] M. Bettini, A. Shankar, and A. Prorok, "System neural diversity: Measuring behavioral heterogeneity in multi-agent learning," *Journal of Machine Learning Research*, vol. 26, no. 163, pp. 1–27, 2025.
- [23] T. Wang, L. Zeng, W. Dong, Q. Yang, Y. Yu, and C. Zhang, "Context-aware sparse deep coordination graphs," in *International Conference on Learning Representations (ICLR)*, 2022, arXiv:2106.02886.
- [24] E. Mieczkowski, R. Mon-Williams, N. Bramley, C. G. Lucas, N. Vélez, and T. L. Griffiths, "Predicting multi-agent specialization via task parallelizability," *arXiv preprint arXiv:2503.15703*, 2025.
- [25] T. Verma and P. Varakantham, "Learning individual policies in large multi-agent systems through local variance minimization," *arXiv preprint arXiv:2212.13379*, 2022.
- [26] W. Khelifi, S. Singh, O. Mahjoub, R. de Kock, A. Vall, R. Gorsane, and A. Pretorius, "On diagnostics for understanding agent training behaviour in cooperative MARL," in *XAI4DRL Workshop at the AAAI Conference on Artificial Intelligence*, 2023, arXiv:2312.08468.
- [27] G. Sokar, R. Agarwal, P. S. Castro, and U. Evci, "The dormant neuron phenomenon in deep reinforcement learning," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, vol. 202, 2023, pp. 32 145–32 168.
- [28] H. Qin, C. Ma, M. Deng, Z. Liu, S. Mei, X. Liu, C. Wang, and S. Shen, "The dormant neuron phenomenon in multi-agent reinforcement learning value factorization," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 35 727–35 759.
- [29] Z. Liu, Z. Gao, H. Qin, J. Hu, J. Wu, M. Zhu, H. Zhang, C. Ma, S. Shen, and C. Wang, "Stagnant neuron: Towards understanding the plasticity loss in multi-agent reinforcement learning value factorization methods," *arXiv preprint arXiv:2606.25335*, 2026.
- [30] T. Schaul, A. Barreto, J. Quan, and G. Ostrovski, "The phenomenon of policy churn," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 2537–2549.
- [31] Y.-N. Weng and H.-W. Lee, "How exploration breaks cooperation in shared-policy multi-agent reinforcement learning," *arXiv preprint arXiv:2601.05509*, 2026.
- [32] A. Yamaguchi, "Emergent coordination and phase structure in independent multi-agent reinforcement learning," *arXiv preprint arXiv:2511.23315*, 2025.
- [33] M. A. Cardei, M. Landers, and A. Doryab, "Coordination matters: Evaluation of cooperative multi-agent reinforcement learning," *arXiv preprint arXiv:2605.06557*, 2026.
- [34] G. Papoudakis, F. Christianos, L. Schäfer, and S. V. Albrecht, "Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks," in *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.
- [35] D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein, "The complexity of decentralized control of Markov decision processes," *Mathematics of Operations Research*, vol. 27, no. 4, pp. 819–840, 2002.
- [36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations (ICLR)*, 2015.
- [37] T. Tieleman and G. Hinton, "Lecture 6.5—RMSProp: Divide the gradient by a running average of its recent magnitude," 2012. [Online]. Available: https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf
- [38] T. Hu, Y. Cui, R. Tang, B. Luo, and K. Li, "Beyond monotonicity: Revisiting factorization principles in multi-agent Q-learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 26, 2026, pp. 21 876–21 884.
- [39] S. V. Albrecht and S. Ramamoorthy, "A game-theoretic model and best-response learning method for ad hoc coordination in multiagent systems," in *Proceedings of the 12th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2013.
- [40] F. Christianos, L. Schäfer, and S. V. Albrecht, "Shared experience actor-critic for multi-agent reinforcement learning," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [41] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.
- [42] L. V. Hedges, "Distribution theory for Glass's estimator of effect size and related estimators," *Journal of Educational Statistics*, vol. 6, no. 2, pp. 107–128, 1981.
- [43] R. Agarwal, M. Schwarzer, P. S. Castro, A. C. Courville, and M. G. Bellemare, "Deep reinforcement learning at the edge of the statistical precipice," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

APPENDIX

APPENDIX A

TABLES REFERENCED IN THE TEXT

This appendix collects the configuration and per-seed tables the preceding sections refer to, so that every value cited in the text can be read from this article alone.

Table V lists the shared training hyperparameters. They are applied identically to VDN/QMIX/QTRAN on the LBF grids except the optimiser (Adam for VDN/QMIX, RMSprop for QTRAN; learning rate 5×10^{-4} throughout), so that the mixer is the primary intended per-algorithm difference; the RWARE cohorts run to 2×10^7 steps.

TABLE V. SHARED TRAINING HYPERPARAMETERS

Hyperparameter	Value
Learning rate	5×10^{-4} (all)
Optimiser	Adam for VDN and QMIX; RMSprop ($\alpha=0.99$) for QTRAN ^b
Gradient-norm clip	10
Discount γ	0.99
TD(λ)	0.5
Double-Q	True
Batch size	128 episodes
Replay-buffer size	5000 episodes
Target-update interval	200 (hard)
ϵ -greedy schedule	$1.0 \rightarrow 0.05$ over 5×10^5 steps
Agent network	Feed-forward MLP, hidden dim 64 (use_rnn=False)
Agent input	observation + one-hot agent id (obs_agent_id=True)
Runner / parallel envs	parallel, batch_size_run=4 ^a
Training horizon $t_{\max}$	1×10^7
Mixer (intended difference)	vdn / qmix / qtran_base

^aThe static configuration files differ on the runner key (QTRAN declares parallel=batch_size_run=4 while QMIX and VDN declare episode/1); the launch scripts override all three to parallel, batch_size_run=4 on the command line, so the realized runs are matched, as confirmed by the run inventory. One mixer-internal width is not numerically identical (mixing_embed_dim: QMIX 32, QTRAN 64, VDN n/a); this is treated as part of the mixing architecture under test rather than as a shared hyperparameter.

^bThe optimiser is hardcoded per learner rather than exposed as a configuration key, so it could not be matched by the shared-configuration discipline: q_learner (VDN, QMIX) constructs Adam at $\text{lr}=5 \times 10^{-4}$, and qtran_learner constructs RMSprop at the same learning rate with $\alpha=0.99$, $\epsilon=10^{-5}$. Those two RMSprop-specific constants are never read on the Adam path. This is the difference the QTRAN-with-Adam control in the main text isolates.

Table VI states the scope-controlled reading of CCI. Its criteria reproduce the classifier of Section III-C and are evaluated in order. Signature names are operational labels; transitions are descriptive and do not enter classification.

TABLE VI. SCOPE-CONTROLLED CCI INTERPRETATION

Operational label	Criterion	Interpretation / next check
Collapse (first)	$\bar{C} < 10^{-3}$	Consider stopping only after target-domain validation; otherwise inspect return, scale, and later checkpoints.
Uni-regime (STAGNATION)	$\max(\text{High}\%, \text{Low}\%) > 80\%$	Inspect additional checkpoints; one-band occupancy alone does not establish halted adaptation.
Dual-occupancy (BI-REGIME)	otherwise	Inspect transition traces or additional checkpoints; occupancy alone does not establish oscillation.

Table VII reports the per-seed seed-mean $\bar{C}$ at the retained early-budget checkpoint. Values are extracted from frozen checkpoints under 50 or 100 greedy episodes: 50 for the five QTRAN seeds and for QMIX seeds 0 and 1, 100 for the remaining eight cells. The collapse threshold 10^{-3} is the one used in the signature-classification rule of main study Section III-C.

TABLE VII. PER-SEED SEED-MEAN $\bar{C}$ AT THE RETAINED EARLY-BUDGET CHECKPOINT (LBF 9 $\times$ 9-5P-3F-COOP-V3)

Seed	QMIX (10%)	QTRAN (10%)	QPplex (100%)
$s = 0$	0.045	2.4×10^{-6}	0.956
$s = 1$	0.163	2.3×10^{-6}	0.705
$s = 2$	0.043	2.4×10^{-6}	0.102
$s = 3$	0.034	1.9×10^{-6}	0.036
$s = 4$	0.497	1.8×10^{-6}	0.906
Range	[0.034, 0.497]	$\sim 10^{-6}$	[0.036, 0.956]
Return-collapse?	NO	YES (5/5)	YES (5/5, return = 0)
C-collapse? (vs 10^{-3})	NO	YES (5/5)	NO (0/5)

Table IX extends the signature taxonomy to 2025 value-decomposition contemporaries on LBF 9 $\times$ 9-5p ($n = 5$ each, agent-only C extraction). “Working” = mean return > 0 ; “Failed” = 0. The DuelMIX working subset reproduces the prior $\bar{C} = \{0.122, 0.144, 0.063\}$ exactly; no contemporary seed exhibits QTRAN-style COLLAPSE ($\bar{C} < 10^{-3}$).

TABLE VIII. PER-GRID $\bar{C}$ AND RETURNS, QMIX VS. VDN, ACROSS LBF GRID SIZES AND A SIX-AGENT VARIANT

Grid	V- $\bar{C}$	Q- $\bar{C}$	d_C/g_C	VDN Ret.	QMIX Ret.	P
8×8	0.510	0.083	−10.10/ − 9.12	0.678 [0.61, 0.74]	0.775 [0.71, 0.84]	0.99
9×9	0.483	0.172	−2.93/ − 2.65	0.632 [0.55, 0.71]	0.732 [0.70, 0.76]	0.97
10×10	0.477	0.208	−1.47/ − 1.33	0.590 [0.54, 0.65]	0.759 [0.73, 0.80]	0.92
11×11	0.475	0.099	−10.08/ − 9.10	0.567 [0.49, 0.66]	0.657 [0.59, 0.70]	1.00
12×12	0.487	0.090	−12.78/ − 11.54	0.504 [0.47, 0.54]	0.631 [0.58, 0.67]	1.00
9×9-6p	0.493	0.220	−1.55/ − 1.40	0.737 [0.67, 0.78]	0.843 [0.83, 0.86]	1.00

TABLE IX. SIGNATURE TAXONOMY EXTENDED TO 2025 VALUE-DECOMPOSITION CONTEMPORARIES ON LBF 9 × 9-5P

Method (2025)	Working	$\bar{C}$ (working)	Failed	$\bar{C}$ (failed)
QFix (Baisero 2025)	4 BI-REGIME	0.04–0.12	1 STAGNATION	0.99
QTypeMix (Fu 2025)	5 BI-REGIME	0.03–0.09	0	—
DuelMIX (Marchesini 2025)	3 BI-REGIME	0.06–0.14	2 STAGNATION	0.84–0.99

Table X gives a descriptive seed-0 cross-checkpoint trace from retained canonical-LBF checkpoints (50 greedy episodes per checkpoint). The trace preserves checkpoint order but is not a cohort-level temporal classifier or proof of recovery, stagnation, or continued adaptation.

TABLE X. DESCRIPTIVE SEED-0 CROSS-CHECKPOINT TRACE FROM RETAINED CANONICAL-LBF CHECKPOINTS

Algorithm	Retained steps (M)	$\bar{C}$ sequence	Transitions/episode	Permitted inference
VDN	0.0002, 2.50, 5.00, 7.50 1.00, 2.00, 4.00,	0.0049, 0.450, 0.565, 0.531 0.045, 0.039, 0.065,	0, 6.60, 0.14, 0.14 4.24, 4.66, 3.18,	Single seed: late one-band persistence.
QMIX	6.00, 8.00, 9.50 1.00, 2.00, 4.00,	0.084, 0.094, 0.094	2.98, 2.38, 2.68	Single seed: repeated within-episode crossings.
QTRAN	6.00, 8.00, 9.95	3.4×10^{-7} – 7.8×10^{-6}	0 at all six checkpoints	Single seed: persistent near-zero dispersion.

Table XI compares checkpoint diagnostics post hoc on retained canonical-LBF runs (five seeds per algorithm, 100 greedy episodes per seed). Entropy is $H(\text{softmax}(Q))/\log|A|$ at temperature 1; Q-gap is the per-agent top-1 minus top-2 action-value gap; disagreement is the fraction of agent pairs choosing different greedy actions; switch is the within-episode fraction of agents whose greedy action changes between consecutive timesteps. These two behavioral summaries are not Role Diversity or policy churn.

TABLE XI. POST-HOC COMPARISON OF CHECKPOINT DIAGNOSTICS ON RETAINED CANONICAL-LBF RUNS

Algorithm	$\bar{C}$	Entropy	Q-gap	Disagreement	Switch
VDN	0.480	0.9925	0.1058	0.770	0.522
QMIX	0.181	0.9991	0.0369	0.741	0.494
QTRAN	3.31×10^{-6}	0.9999995	0.00123	0.472	0.0092

Table XII reports the frozen-cut-off extension on five additional canonical-LBF QTRAN training seeds. Each checkpoint is evaluated for 100 greedy episodes using the pre-existing $\bar{C} < 10^{-3}$ cut-off. This is a positive-class repeatability check, not an independent calibration or specificity set.

TABLE XII. FROZEN-CUT-OFF EXTENSION ON FIVE ADDITIONAL CANONICAL-LBF QTRAN TRAINING SEEDS

Seed	Early step	Early $\bar{C}$	Early return	Final step	Final $\bar{C}$	Final return
5	1,000,200	1.44×10^{-6}	0.000	10,000,790	9.99×10^{-6}	0.007
6	1,000,200	1.35×10^{-5}	0.003	10,000,977	5.74×10^{-6}	0.000
7	1,000,200	1.67×10^{-6}	0.003	10,000,600	2.65×10^{-6}	0.010
8	1,000,200	1.98×10^{-6}	0.000	10,000,200	2.21×10^{-8}	0.000
9	1,000,200	1.01×10^{-6}	0.000	10,000,200	7.93×10^{-6}	0.013