

Anchor and Perturb: Lazy Agent Remediation by Exploration Injection

Chengxi Zhong, Yongzhe Chang

September 2026

1 Abstract

Anchor and Perturb (AnP) is a lightweight framework that resolves multi-agent coordination failures by decoupling exploratory variance injection from recurrent manifold stability. Existing remediation strategies predominantly alter mixing network architectures or enforce simultaneous exploration across the collective, which inevitably precipitates severe temporal-difference penalties in non-monotonic reward spaces. Specifically, AnP isolates underperforming lazy agents and injects an asymmetric exploratory pulse into targeted coordinates whilst anchoring converged teammates to nominal greedy exploitation. Empirical telemetry benchmarks demonstrate that AnP successfully rescues collapsed joint policies (recovering from a 5% evaluation win rate nadir back to 85%) and facilitates escape from suboptimal coordination plateaus, sustaining peak win rates of 90% without requiring structural network modifications.

2 Introduction

2.1 Current Limitations

In cooperative Multi-Agent Reinforcement Learning (MARL) under the Centralised Training with Decentralised Execution (CTDE) framework, value factorisation methods, such as monotonic networks like QMIX or weighted and monotonic variants like WQMIX and QPLEX, frequently suffer from asymmetric convergence. In complex coordination environments, a subset of agents successfully discovers and exploits near-optimal cooperative policies, while remaining agents continue to exploit local actions deemed suboptimal that yield marginal returns. Because joint rewards reflect the collective outcome, the under-performing actions of suboptimal-exploiting agents severely limit the total team return, preventing the entire system from achieving global optimality.

2.2 Proposal and Optimisation

Here, we propose Anchor-and-Perturb (AnP), a novel framework that selectively perturbs M out of N agents via targeted ϵ -exploration, enabling suboptimally performing agents to discover actions that yield a higher return while anchoring the remaining $N - M$ converged teammates to nominal exploitation. Regular ϵ -increment was proposed as a solution for other agents to explore an optimised reward coordination. However, this failure cannot be resolved by simply increasing exploratory drive under standard paradigms. Conventional exploration in MARL is symmetric ($M = N$), meaning that all N agents independently and simultaneously perturb their actions. In this regime, while an under-performing agent may attempt an exploratory action that has the potential to return a higher overall reward, its teammates, already converged on optimal cooperative policies, are simultaneously forced to perform exploration. Because multi-agent coordination landscapes are inherently non-monotonic, these uncoordinated teammate re-exploration frequently land the system in globally suboptimal penalty states. Due to the non-negative monotonicity constraint enforced by modern mixing networks ($\frac{\partial Q_{tot}}{\partial Q_i} \geq 0$), the resulting catastrophic drop in joint reward generates a severe negative temporal-difference error that artificially reduces the individual utility estimates Q_i of all participating agents. Consequently, setting $N = M$ actively penalises the very exploratory actions that could unlock global optimality, entrenching the system in relative overgeneralization.

3 Related Work

3.1 Value Factorisation and Representational Limits

Under the Centralised Training with Decentralised Execution (CTDE) framework, standard value factorisation methods rely on the Individual-Global-Max (IGM) principle to ensure scalable decentralised execution. QMIX enforces this consistency by constraining the mixing network weights to non-negative values, guaranteeing monotonicity ($\frac{\partial Q_{tot}}{\partial Q_i} \geq 0$). However, this strict monotonic projection inherently induces relative overgeneralisation in non-monotonic coordination landscapes. When a subset of agents converges onto a suboptimal policy coordinate, their depressed utility estimates drag down the aggregate value representation. The resulting negative temporal-difference errors artificially penalise viable cooperative trajectories, trapping under-performing agents in a persistent “lazy” state [4]. Subsequent frameworks have attempted to alleviate these representational limitations by directly altering the underlying learning infrastructure. Weighted QMIX (WQMIX) introduces a weighted projection that applies an asymmetric weighting scheme to assign higher importance to joint actions that outperform the current monotonic bound, effectively establishing a soft floor against severe negative updates. Similarly, QPLEX overhauls the architectural formulation by introducing a duplex dueling mixing network that factorises the joint action-value into state-dependent values and advantage

functions, mathematically recovering the complete IGM function class. Crucially, while both WQMIX and QPLEX expand the representational capacity of the mixing network during the optimisation phase, neither resolves the underlying behavioral pathology of lazy agents during experience collection. Because exploration remains symmetric across all coordinates, any attempt by an underperforming agent to re-explore uncoordinated actions simultaneously perturbs its converged teammates, precipitating exploratory collisions and joint value collapse. Rather than introducing structural architectural overhead or modifying the underlying value factorisation mechanics, Anchor-and-Perturb (AnP) resolves this pathology orthogonally during the data generation phase. By decoupling exploratory variance from nominal exploitation, AnP remediates lazy agents directly across standard QMIX, WQMIX, and QPLEX backbones without requiring alterations to their underlying networks.

3.2 Multi-Agent Exploration and Relative Overgeneralisation

Recent approaches have attempted to diversify multi-agent exploration by explicitly penalising behavioural overlap. For instance, Multi-Agent Divergence Policy Optimisation (MADPO) [5] addresses heterogeneous exploration through a sequential updating scheme that maximises mutual policy divergence. By utilising conditional Cauchy-Schwarz divergence, MADPO enforces exploration incentives both across training episodes and between individual agents. While MADPO successfully drives agents toward distinct policies, it does so entirely via optimisation-level regularisation. Because agents still inject simultaneous variance into the environment, the mixing network remains vulnerable to the catastrophic temporal-difference penalties inherent to non-monotonic reward landscapes. Anchor and Perturb sidesteps this structural bottleneck by enforcing behavioural isolation strictly during the data generation phase, clamping non-targeted teammates to a stationary background rather than attempting to disentangle exploratory collisions mathematically during the gradient update.

3.3 Credit Assignment and Asymmetric Convergence

Recent studies have directly targeted the lazy agent pathology in cooperative MARL. Counterfactual credit assignment frameworks such as COMA [1] laid the theoretical foundation for measuring individual marginal contributions against joint returns. Notably, LAIES [3] constructs causal interaction graphs to identify lazy agents and introduces counterfactual intrinsic rewards to encourage agents to influence external environment states. Similarly, meta-exploration frameworks like MAVEN condition agent policies on hierarchical latent variables to achieve temporally extended joint exploration beyond monotonic constraints, while CDS [2] balances parameter sharing with trajectory diversity objectives.

However, these approaches address lazy or suboptimal convergence by introducing structural model overhead—such as auxiliary state transition models, variational discriminators, or dense intrinsic reward shaping. Furthermore, they

continue to explore across the entire agent collective simultaneously. In contrast, Anchor-and-Perturb requires neither causal transition models nor auxiliary loss terms. By decoupling exploratory variance at the environment-sampling interface and anchoring non-targeted agents to nominal exploitation, AnP provides an orthogonal, plug-and-play remediation schedule that directly stabilises credit assignment across unaugmented value factorisation backbones.

4 Methodology

4.1 Anchor

4.1.1 Definition

We formally define **Anchor** as the localised act of identifying a lazy agent at timestep t_0 , when we isolate an underperforming agent coordinate while simultaneously clamping the operational manifold of its remaining teammates. Crucially, anchoring serves a dual theoretical role: it designates the target subspace requiring policy rehabilitation, and it establishes an exploitation barrier across the non-targeted coordinates $\mathcal{A}_{anch} = \{1, \dots, N\} \setminus \mathcal{A}_{pert}$. By freezing exploratory entropy across converged agents, the anchor operation halts the propagation of asymmetric policy drift and constructs a stationary empirical background. This ensures that any subsequent credit assignment remains strictly decoupled from teammate non-stationarity. To dynamically partition the agent cohort into anchored and perturbed sets at t_0 , we identify candidate suboptimal agents through several diagnostic indicators, listed below.

4.1.2 Infrastructure Preservation

Crucially, clamping $\mathcal{A}_{anch}$ to nominal greedy exploitation does not degenerate the environment into an isolated single-agent Markov Decision Process (MDP), nor does it violate the core premises of Centralised Training with Decentralised Execution (CTDE). The joint transition dynamics $\mathcal{P}(s' \mid s, \mathbf{u})$ and environmental rewards $r(s, \mathbf{u})$ remain strictly governed by the full joint action vector $\mathbf{u} = (\mathbf{u}^{pert}, \mathbf{u}^{anch})$. More importantly, anchoring is enforced strictly at the behavioural rollout interface; network weights across both $\mathcal{A}_{pert}$ and $\mathcal{A}_{anch}$ remain entirely dynamic and participate simultaneously in the centralised gradient backward pass. As the perturbed coordinate discovers viable cooperative trajectories, the utility estimates Q_j of anchored teammates continuously adapt to accommodate the shift in the collective manifold. Consequently, AnP preserves the full multi-agent credit assignment dynamic while eliminating the catastrophic temporal-difference penalties induced by symmetric exploratory collisions.

4.1.3 Human-Based Observation and Quantitative Diagnostic Rules

Under the human-in-the-loop diagnostic regime, identifying candidate agents for $\mathcal{A}_{pert}$ relies on expert oversight of environment telemetry and policy trajectory visualisations. In practice, this identification translates qualitative failure modes into three trackable diagnostic indicators:

- **Exploratory and Policy Atrophy:** The action distribution entropy of agent i , defined over a trailing window W as $\bar{\mathcal{H}}_i(t) = \frac{1}{|W|} \sum_{\tau=t-W}^t \mathcal{H}(\pi_i(\cdot | \tau))$, collapses prematurely towards zero ($\bar{\mathcal{H}}_i(t) < \delta_{\mathcal{H}}$), indicating that the agent has frozen onto trivial primitives (e.g. constant holding or loitering).
- **Marginal Contribution Asymmetry:** Real-time environmental metrics (such as target damage, task interactions, or localized goal transitions) show extreme skewness. Let $c_i(t)$ denote the rolling marginal contribution of agent i ; a candidate satisfies $c_i(t) \ll \frac{1}{N} \sum_{j=1}^N c_j(t)$.
- **Joint Return Stagnation:** The moving average of team evaluation return $\bar{R}(t)$ exhibits a persistent plateau ($\Delta \bar{R}(t) \approx 0$) below near-optimal task baselines.

To consolidate these telemetry signals into a concrete diagnostic criterion, we define an empirical *Laziness Disparity Index* $\mathcal{L}_i(t)$:

$$\mathcal{L}_i(t) = \left(1 - \frac{c_i(t)}{\max_j c_j(t) + \epsilon_0}\right) \cdot \exp(-\beta \bar{\mathcal{H}}_i(t)) \quad (1)$$

where $\epsilon_0 > 0$ is a numerical smoothing constant and $\beta > 0$ balances contribution against distributional collapse. At intervention epoch t_0 , the operator designates the coordinate exhibiting the maximal disparity as the single perturbed agent:

$$\mathcal{A}_{pert} = \left\{ \arg \max_i \mathcal{L}_i(t_0) \right\} \quad (2)$$

Formalising the operator’s diagnosis into this explicit criterion clarifies why a specific coordinate is isolated and provides a rigorous mathematical bridge towards fully automated triggering.

4.1.4 Automated Lazy-Agent Diagnostic Trigger

While human-in-the-loop intervention offers an empirical baseline for validating localised perturbation, it remains inherently inefficient and susceptible to stochastic variance. In multi-agent environments, manual inspection typically requires repeated evaluation sweeps to locate the exact anchor coordinate. Furthermore, cross-seed variations can render previously recorded behavioural patterns invalid across distinct runs, limiting manual intervention to post-hoc remediation.

To overcome these bottlenecks, we propose an automated lazy-agent trigger designed to dynamically identify both the underperforming agent coordinate and

the critical intervention timestep t_0 within a single forward training trajectory, completely bypassing manual oversight. Specifically, we are investigating adapting causal and mutual divergence diagnostics into the Anchor-and-Perturb infrastructure to trigger perturbation autonomously. Looking towards long-term deployment, this automated pipeline can be orchestrated via an LLM meta-supervisor. By monitoring trailing behavioural telemetry and reward stability, the meta-supervisor performs discrete diagnostic passes strictly bounded to at most $K + 1$ inference calls per run, thereby preserving computational resources whilst maintaining adaptive, zero-human-overhead coordination recovery.

4.2 Perturbation

We formally define **Perturbation** as an asymmetric, localised exploration-and-annealing intervention applied over an active subset of agent coordinates. Formally, let $\mathcal{A}_{pert} \subset \{1, \dots, N\}$ denote the index set of perturbed agents with cardinality $|\mathcal{A}_{pert}| = M$, and let $\mathcal{A}_{anch} = \{1, \dots, N\} \setminus \mathcal{A}_{pert}$ denote the remaining $N - M$ anchored teammates.

Unlike standard symmetric exploration where all agents are set to explore concurrently after a specific point where ϵ is increased simultaneously over all agents, AnP restricts exploratory variance exclusively to $\mathcal{A}_{pert}$ while enforcing strict greedy exploitation across $\mathcal{A}_{anch}$:

$$u_t^j = \arg \max_{u^j} Q_j(\tau_t^j, u^j), \quad \forall j \in \mathcal{A}_{anch} \quad (3)$$

For agents $i \in \mathcal{A}_{pert}$, exploration is driven by a localised cyclic schedule. At timestep t_0 , we initiate a perturbation state, where the local exploration rate $\epsilon_t^{(i)}$ is reset to an elevated exploration rate ϵ_{boost} and subsequently annealed toward a nominal baseline ϵ_{min} over an annealing window of duration T_{anneal} :

$$\epsilon_t^{(i)} = \epsilon_{min} + (\epsilon_{boost} - \epsilon_{min}) \cdot \max \left(0, 1 - \frac{t - t_0}{T_{anneal}} \right) \quad (4)$$

This schedule ensures that the perturbed agent initiates a sufficient exploratory drive to discover high-utility cooperative actions, followed by a consolidation phase where the newly discovered actions are exploited and reinforced via temporal-difference updates against a stationary teammate background.

This perturbation intervention can be applied iteratively across K successive exploration cycles. In this work, we establish the foundational proof of concept by restricting our study to the minimal rank-1 single-cycle setting:

$$M = 1, \quad K = 1 \quad (5)$$

Under this setting, exactly one agent coordinate ($M = 1$) undergoes a single localised perturbation-and-annealing cycle ($K = 1$), guaranteeing that the probability of multi-agent collision is identically zero ($P(M \geq 2) \equiv 0$).

5 Experiments

5.1 Experimental Setup

To evaluate the empirical viability of Anchor-and-Perturb (AnP), we benchmark our framework against standard monotonic value factorisation under the Centralised Training with Decentralised Execution (CTDE) paradigm. All experiments are conducted over a training budget of 3,000 episodes, logging evaluation metrics across 25-episode intervals. The primary evaluation metrics tracked via telemetry are evaluation win rate ($\text{eval_win_rate} \in [0, 1]$), evaluation team return (eval_reward), and individual training returns (train_return).

We evaluate three operational regimes:

- **Nominal Baseline (Unperturbed QMIX):** Standard decentralized ϵ -greedy exploration annealed symmetrically across all N agents toward a terminal baseline $\epsilon_{min} = 0.05$.
- **AnP Remediation Regime:** An illustrative run exhibiting severe policy degradation, wherein a human operator detects early stagnation and triggers a single rank-1 perturbation cycle ($M = 1, K = 1$) to evaluate policy restoration.
- **AnP Enhancement Regime:** A representative run initialized under standard monotonic training where the AnP intervention is applied at a nominal coordination plateau to test escape from sub-optimal local equilibria.

5.2 Telemetry Proof-of-Concept

To validate the core premise of AnP prior to automating heuristic triggers, we conduct an empirical smoke test of the underlying intervention mechanics. Specifically, we examine whether restricting exploratory variance exclusively to a single designated coordinate ($\mathcal{A}_{pert}$, with $|\mathcal{A}_{pert}| = 1$) while clamping remaining teammates ($\mathcal{A}_{anch}$) enables recovery without inducing team-wide coordination collapse.

5.2.1 Remediation of Severe Policy Degradation

Under the standard QMIX paradigm, asymmetric convergence often degenerates into relative overgeneralization. As captured in our remediation telemetry, the agent team initially attains a viable policy near episode 1,275 (85% win rate, evaluation reward 18.17), before suffering a severe policy collapse by episodes 1,650–1,675, where evaluation win rates bottom out at 5% and rewards drop to 7.48.

At episode $t_0 = 1,925$, the operator designates a single under-performing coordinate into $\mathcal{A}_{pert}$, elevating its local exploration rate to $\epsilon_{boost} \approx 0.292$ and annealing over $T_{anneal} = 450$ episodes. During the active pulse window (episodes 1,925–2,350), the system tolerates controlled exploratory disruption while the

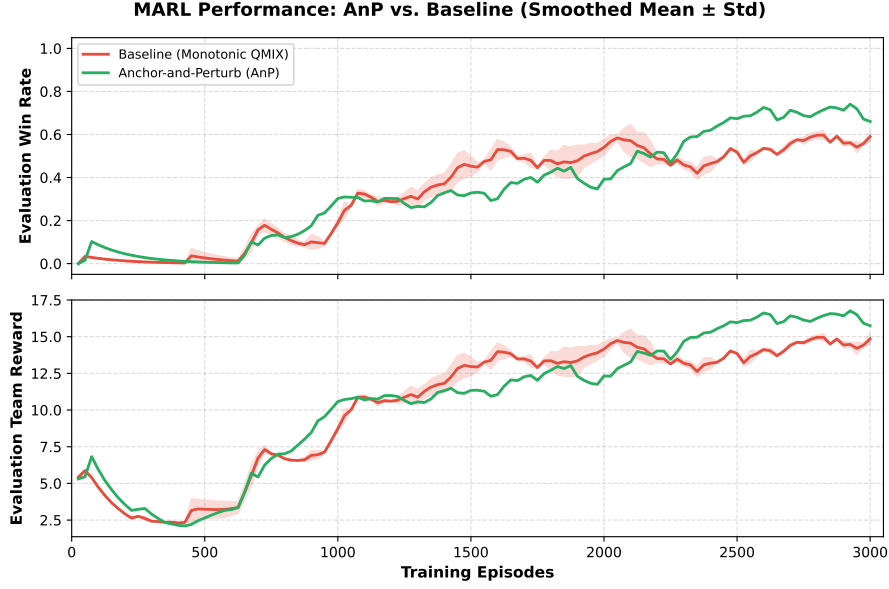


Figure 1: Training and evaluation telemetry trajectories across 3,000 episodes, contrasting standard symmetric QMIX against the AnP enhancement regime ($M = 1, K = 1$, pulse window $t \in [1,875, 2,300]$).

perturbed coordinate reorganizes its policy against the stationary background of $\mathcal{A}_{anch}$. Immediately upon completing the annealing cycle at episode 2,375, the joint policy recovers instantly to an 85% evaluation win rate (18.01 reward), subsequently sustaining average win rates between 40% and 50% through episode 3,000. This confirms that asymmetric perturbation can actively rescue stalled policies from catastrophic stagnation.

5.2.2 Escape from Sub-Optimal Plateaus

To evaluate AnP in non-degraded but sub-optimally saturated regimes, we assess an intervention triggered during steady-state coordination. In the nominal unperturbed baseline, performance exhibits recurring instability, with the evaluation win rate repeatedly dropping to floors of 10%–25% in late training.

In contrast, applying AnP at $t_0 = 1,875$ ($\epsilon_{boost} \approx 0.287$, $M = 1$) yields substantial policy consolidation. Following a transient exploratory dip, the system achieves peak win rates of 90% (reward 18.82) during the pulse tail. In the post-perturbation consolidation window (episodes 2,325–3,000), the anchored-and-perturbed architecture maintains an evaluation win rate $\geq 60\%$ across 78.6% of checkpoints, consistently sustaining evaluation rewards between 15.5 and 17.5.

5.3 Discussion

These telemetry results establish two critical empirical validations:

1. **Exploration Isolation:** Injecting localized exploratory variance into a single coordinate ($M = 1$) avoids the catastrophic reward collapse inherent to symmetric $M = N$ re-exploration. The non-targeted coordinates $\mathcal{A}_{anch}$ successfully buffer the team’s operational manifold.
2. **Decoupled Credit Assignment:** The stationary background constructed by clamping $\mathcal{A}_{anch}$ allows newly discovered high-utility actions from $\mathcal{A}_{pert}$ to be reinforced cleanly through the mixing network, verifying the operational foundation required for automated heuristic-based anchoring.

6 Future Work

While this work establishes the foundational proof-of-concept for Anchor-and-Perturb via minimal rank-1 ($M = 1, K = 1$) interventions on representative trajectories, several critical avenues remain for advancing towards a fully autonomous and comprehensively benchmarked framework:

6.1 Multi-Seed Statistical Validation

Cooperative MARL inherently suffers from significant environment and exploration stochasticity. To rigorously confirm that policy rehabilitation is not an artifact of favourable random seeding, future evaluations will benchmark AnP across a minimum of 5 independent random seeds. Telemetry trajectories, evaluation returns, and convergence win rates will be reported with comprehensive statistical aggregations (mean $\pm$ standard deviation and 95% confidence intervals), systematically establishing the variance-reduction and stabilization properties of AnP.

6.2 Exploration Mechanism Ablation Study

To decouple the distinct theoretical mechanics underpinning AnP, we plan two dedicated ablation baselines:

- **Symmetric Re-exploration Baseline ($M = N$):** Evaluating team-wide exploration injection, wherein all N agents simultaneously elevate and anneal their exploration rates (ϵ_{boost}) at timestep t_0 . This will explicitly verify whether recovery stems merely from supplementary exploratory entropy or strictly from anchoring the converged teammates ($\mathcal{A}_{anch}$) to prevent catastrophic temporal-difference collapse.
- **Random Coordinate Perturbation ($i \sim \mathcal{U}(\{1, \dots, N\})$):** Injecting the localized pulse into an arbitrarily chosen agent rather than the diagnosed lazy coordinate. This will validate the necessity and precision

of the lazy-agent identification criterion ($\mathcal{L}_i$), confirming that arbitrarily perturbing an already-converged cooperative agent actively destabilizes the joint policy manifold.

6.3 Autonomous Triggers and LLM Meta-Supervision

Beyond experimental ablations, future architectural extensions will transition AnP from human-in-the-loop diagnostics to fully autonomous operation:

- **Autonomous Heuristic Triggers:** Implementing an online, event-driven detection loop based on the streaming Laziness Disparity Index $\mathcal{L}_i(t)$ to self-determine t_0 and $\mathcal{A}_{pert}$ within a single forward pass.
- **Bounded LLM Meta-Supervision:** Deploying an asynchronous LLM supervisor constrained to at most $K + 1$ inference passes per run, dynamically adapting the annealing window T_{anneal} and perturbation magnitude ϵ_{boost} upon detected stagnation while maintaining zero runtime overhead during standard gradient updates.
- **Higher-Rank Coordination ($M \geq 2, K > 1$):** Extending localized perturbation to multi-agent subsets with mutual exploratory orthogonality constraints to remediate multiple lagging teammates concurrently.
- **Causal and Architectural Integration:** Benchmarking AnP across advanced non-monotonic backbones by combining asymmetric exploration schedules with causal credit assignment frameworks (such as LAIES) to verify whether extrinsic causal influence can directly inform the anchoring boundary.

References

- [1] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nemanja Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [2] Runji Li et al. Celebrating diversity in shared multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- [3] Boyin Liu, Zhiqiang Pu, Yi Pan, Jianqiang Yi, Yanyan Liang, and Du Zhang. Lazy agents: A new perspective on solving sparse reward problem in multi-agent reinforcement learning. In *International Conference on Machine Learning (ICML)*, pages 22097–22115, 2023.
- [4] Arilton Silva et al. Learning independently from causality in multi-agent environments. *arXiv preprint arXiv:2311.02741*, 2023.

- [5] Tianpei Yang, Zhaopeng Du, Jianye Ren, Zhendong Wang, Jianye Hao, and Matthew E. Taylor. Measuring mutual policy divergence for multi-agent sequential exploration. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023)*, pages 155–163, 2023.