# STABILIZING INDEPENDENT LEARNING IN GENERAL-SUM SELF-PLAY VIA COUPLING-ADAPTIVE REGULARIZATION

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

## ABSTRACT

Independent policy updates in general-sum self-play offer computational efficiency but introduce severe non-stationarity. To guarantee joint objective improvement, existing theory bounds the induced independent-to-joint objective gap and constrains policy updates. The dependence of these bounds on the total agent count can make permissible steps vanish in large systems. We derive a bound on this gap in which each agent's update is weighted by its state-dependent coupling with the others. Under interaction sparsity, the coupling coefficient for each agent remains $\mathcal{O}(1)$ as the population grows. Building on this analysis, we introduce Coupling-Adaptive Regularized Policy Optimization (CARPO). CARPO uses a vector critic to estimate local coupling and adapts each agent's trust region, tightening updates in highly coupled states and relaxing them in loosely coupled ones. Empirical evaluations range from a tabular game to large-scale autonomous driving. By restricting updates during critical interactions, CARPO induces a "safe hesitation" phase that prevents premature convergence in a coordination trap. Its state-dependent adaptation outperforms the tested static clipping choices across coordination levels. In autonomous driving, CARPO maintains the rapid initial learning of independent updates while surpassing the final performance of the joint-update baseline.

## 1 INTRODUCTION

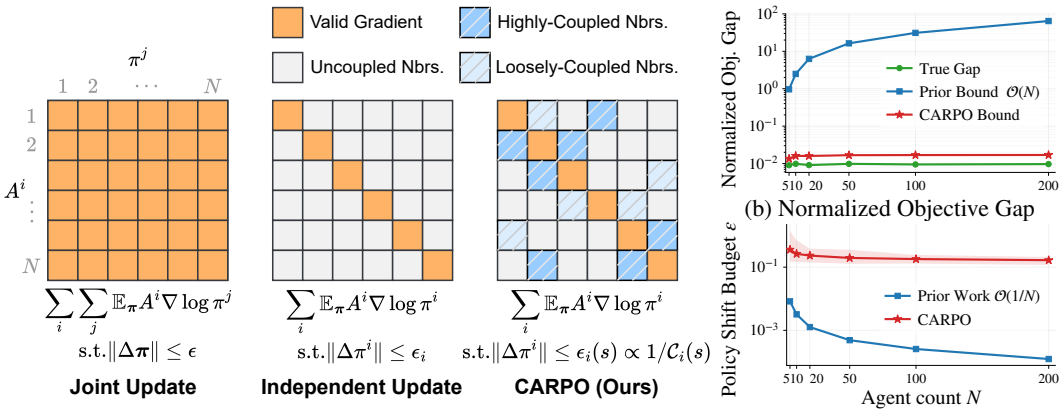
As self-play reinforcement learning (RL) scales to massive systems such as traffic networks, the exponential growth of joint action spaces makes exact optimization intractable (Lanctot et al., 2017; Dinneweth et al., 2022; Ma et al., 2024). While directly optimizing the joint policy objective can offer theoretical monotonic improvement guarantees, estimating the gradient of this objective suffers from prohibitive variance. To bypass this, recent approaches employ independent policy updates, enabling empirical applications of self-play in massive domains such as autonomous driving (Cornelisse & Vinitsky, 2024; Cusumano-Towner et al., 2025; Chang et al., 2025). In this paradigm, each agent computes its local surrogate objective by treating the policies of other agents as fixed during its update. By permitting simultaneous updates, independent learning bypasses the computational bottleneck of joint updates, making it appealing for large-scale systems.

The critical limitation of independent learning is the non-stationarity introduced by concurrent updates. This issue plagues both fully decentralized algorithms (e.g., IPPO (De Witt et al., 2020)) and centralized-critic methods (e.g., MAPPO (Yu et al., 2022)), as both rely on independent actor optimization. From the perspective of an individual agent, other agents' concurrent updates make its local environment non-stationary, violating the stationary Markov assumption of its local learning process (Claus & Boutilier, 1998; Papoudakis et al., 2019). The resulting misalignment between its independent surrogate objective and the joint policy objective is quantified by the independent-to-joint objective gap. While sequential update schemes can bypass this non-stationarity to restore monotonic improvement guarantees (Kuba et al., 2021; Zhong et al., 2024), their sequential nature forfeits the simultaneous updates that make independent learning computationally efficient.

To preserve simultaneous updates, theoretically grounded methods bound the independent-to-joint objective gap to constrain the induced non-stationarity (Sun et al., 2022; Su & Lu, 2024). However, these analyses rely on the total agent count $N$ to constrain each individual agent's update (Fig. 1b).



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




(a) Policy Gradient Structures and Cross-agent Couplings

(b) Normalized Objective Gap

(c) Permissible Trust Region $\epsilon$

Figure 1: **Independent-to-joint objective gap and bounds.** The objective gap quantifies non-stationarity as the difference between joint and independent objectives. (a) Independent updates ignore cross-agent dependencies. CARPO adapts trust regions based on local coupling. (b) In the CACC example, the prior penalty increases with its $\mathcal{O}(N)$ factor, while our coupling term follows the nearly stable normalized gap. (c) Giving each update the same allowance in the gap bound yields an $\mathcal{O}(1/N)$ range with the prior coefficient and a nearly stable range with coupling. Appendix G specifies the plotted quantities.

If the contribution of each update to the gap bound is held fixed, this dependence on $N$ makes the corresponding update range shrink as $\mathcal{O}(1/N)$ (Fig. 1c, blue line), restricting learning as the system scales. This scaling pushes practical implementations toward fixed clipping constants, which cannot respond to changes in interaction strength.

In many large-scale systems, however, interactions exhibit sparsity: agents rarely interact with the entire population, but couple most strongly with local neighborhoods. Building on this property, we identify the state-dependent coupling strength between agents as the key quantity in the objective gap. When the influence of distant agents on advantages decays, the coupling coefficient for each agent remains $\mathcal{O}(1)$ as the population grows. This local structure motivates coupling-adaptive trust regions (Fig. 1c, red line).

We propose **C**oupling-**A**daptive **R**egularized **P**olicy **O**ptimization (**CARPO**). By formulating an advantage centering identity, we bound the independent-to-joint objective gap using the magnitude of each policy update and its state-dependent coupling with other agents. In practice, CARPO estimates this coupling with a vector critic and dynamically modulates each agent's PPO clipping range. It restricts updates in highly coupled states to suppress the effects of non-stationarity while permitting larger learning steps in loosely coupled states.

Our contributions are summarized as follows:

1. We derive a bound on the independent-to-joint objective gap using advantage centering. The term for each agent's policy update uses the state-dependent coupling coefficient $\mathcal{C}_i(s)$ in place of a coefficient proportional to $N$. Under the interaction sparsity condition in Appendix B.1, this coefficient remains $\mathcal{O}(1)$ for each agent as the population grows.

2. We propose CARPO, an independent learning algorithm for general-sum self-play. Its auxiliary vector critic estimates how local action stochasticity affects the advantages of other agents. By modulating the clipping range inversely to estimated coupling, CARPO tightens updates in highly coupled states and relaxes them in loosely coupled ones while retaining simultaneous updates.

3. We validate CARPO from tabular games to large-scale autonomous driving. First, its coupling-adaptive updates induce a "safe hesitation" phase that prevents premature convergence in the Stochastic Climbing Game. Second, state-level adaptation outperforms the tested static clipping choices across coordination levels in LBF. Finally, in autonomous

driving, CARPO maintains rapid initial learning while surpassing the final performance of the joint-update baseline.

## 2 PRELIMINARIES

### 2.1 NETWORKED GENERAL-SUM MARKOV GAMES

We formulate the large-scale self-play process as a general-sum Markov game (Littman, 1994), defined by the tuple $\mathcal{G} = \langle \mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \mathcal{P}, \{r_i\}_{i\in\mathcal{N}}, \gamma \rangle$, where $\mathcal{N} = \{1, \dots, N\}$ is the agent set and $\mathcal{S}$ is the global state space. Each instance $i$ selects an action $a_i \in \mathcal{A}_i$ simultaneously based on the shared self-play policy, forming the joint action vector $\boldsymbol{a} = (a_1, \dots, a_N) \in \boldsymbol{\mathcal{A}}$. The environment transitions to a new state $s'$ according to the dynamics $\mathcal{P}(s', \boldsymbol{a})$. Because agents in large-scale physical systems such as traffic networks often interact locally, we focus on the strength of these interactions. Appendix B.1 states a sufficient condition for local coupling to remain bounded as the population grows.

### 2.2 POLICY OPTIMIZATION OBJECTIVES

Let $\pi^i(\cdot|s)$ denote the stochastic policy of agent $i$. In our self-play setting, all agents share the identical policy network during execution, but perform simultaneous independent policy updates. The joint policy is given by $\boldsymbol{\pi}(\boldsymbol{a}|s) = \prod_{i\in\mathcal{N}} \pi^i(a^i|s)$. Let $V^i_{\boldsymbol{\pi}}(s)$, $Q^i_{\boldsymbol{\pi}}(s, \boldsymbol{a})$, and $A^i_{\boldsymbol{\pi}}(s, \boldsymbol{a})$ denote the standard value, action-value, and advantage functions for agent $i$ under a joint policy $\boldsymbol{\pi}$. Let $J_i(\boldsymbol{\pi}) = \mathbb{E}_{s_0\sim\rho_0}[V^i_{\boldsymbol{\pi}}(s_0)]$ denote the expected return for agent $i$. To quantify the collective performance of the system, we define the joint return, a standard metric for evaluating overall efficiency in general-sum games (Leibo et al., 2017), as $J(\boldsymbol{\pi}) \triangleq \sum_{i\in\mathcal{N}} J_i(\boldsymbol{\pi})$. Our goal is to improve this global objective while retaining simultaneous policy updates.

**The Joint Surrogate Objective.** The performance difference lemma (Kakade & Langford, 2002) establishes an exact identity for the joint return $J(\boldsymbol{\pi})$. By approximating the state distribution of the new policy via that of the old one, trust region methods (Schulman et al., 2015a) derive a tractable surrogate for improvement analysis. Under simultaneous updates, this *joint surrogate objective* is

$$\mathcal{L}_{\text{joint}}(\boldsymbol{\pi}_{\text{new}}) = \mathbb{E}_{s\sim d_{\boldsymbol{\pi}_{\text{old}}}, \boldsymbol{a}\sim\boldsymbol{\pi}_{\text{new}}} \left[ \sum_{i\in\mathcal{N}} A^i_{\boldsymbol{\pi}_{\text{old}}}(s, \boldsymbol{a}) \right], \tag{1}$$

where $d_{\boldsymbol{\pi}_{\text{old}}}$ is the discounted state distribution, defined by $d_{\boldsymbol{\pi}_{\text{old}}} = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t P(s_t = s \mid \boldsymbol{\pi}_{\text{old}})$. However, directly maximizing $\mathcal{L}_{\text{joint}}$ is intractable because the joint action space $\boldsymbol{\mathcal{A}}$ scales exponentially with $N$. While sequential update schemes (Kuba et al., 2021; Zhong et al., 2024) can bypass this complexity, they forfeit parallel efficiency and restrict their guarantees to fully cooperative settings. This makes them inapplicable to the large-scale, general-sum paradigm we focus on.

**The Independent Surrogate Objective.** To overcome these scalability issues, independent learning methods employ a simplified approximation. Each agent $i$ optimizes its policy assuming other agents remain fixed at joint policy $\boldsymbol{\pi}^{-i}_{\text{old}}$. This yields the *independent surrogate objective*:

$$\mathcal{L}^i_{\text{ind}}(\pi^i_{\text{new}}) = \mathbb{E}_{s\sim d_{\boldsymbol{\pi}_{\text{old}}}, a_i\sim\pi^i_{\text{new}}, \boldsymbol{a}_{-i}\sim\boldsymbol{\pi}^{-i}_{\text{old}}} \left[ A^i_{\boldsymbol{\pi}_{\text{old}}}(s, a_i, \boldsymbol{a}_{-i}) \right]. \tag{2}$$

This objective enables independent policy updates but introduces a critical discrepancy: from agent $i$'s perspective, the environment changes as other agents concurrently update their own policies. This non-stationarity violates the stationary Markov assumption underlying each agent's local update, so standard policy improvement guarantees no longer directly apply to independent updates.

**The Independent-to-Joint Objective Gap.** Independent updates inherently ignore the concurrent policy shifts of other agents. To formally quantify this induced non-stationarity, we define the *independent-to-joint objective gap* as the absolute difference between the joint objective and the summed independent objectives:

$$\Delta\mathcal{L} = \left| \mathcal{L}_{\text{joint}}(\boldsymbol{\pi}_{\text{new}}) - \sum_{i\in\mathcal{N}} \mathcal{L}^i_{\text{ind}}(\pi^i_{\text{new}}) \right|. \tag{3}$$

Detailed policy gradient formulations for these objectives are provided in Appendix C. In Section 3, we will derive an upper bound for this gap, linking it to the local coupling coefficients among agents and the magnitude of their policy updates.

# 3 COUPLING-ADAPTIVE REGULARIZED POLICY OPTIMIZATION

In this section, we quantify the surrogate approximation error introduced by independent updates in general-sum games. Specifically, we derive a bound revealing how the independent-to-joint objective gap depends on coupling strength among agents. Based on this analysis, we propose CARPO, which adapts each agent's trust region using a coupling estimate from a vector critic.

## 3.1 A COUPLING-DEPENDENT BOUND ON THE INDEPENDENT-TO-JOINT OBJECTIVE GAP

To establish a permissible trust region for independent agents, we first quantify the interaction strength between them. The *coupling coefficient* captures the aggregate sensitivity of other agents' advantages to the stochasticity of agent $i$'s policy.

**Definition 3.1** (Coupling Coefficient)**.** Let $\bar{A}^{(j,i)}_{\boldsymbol{\pi}}(s, a_j, a_i) \triangleq \mathbb{E}_{\boldsymbol{a}_{-ji}\sim\boldsymbol{\pi}^{-ji}(\cdot|s)}\left[A^j_{\boldsymbol{\pi}}(s, a_j, a_i, \boldsymbol{a}_{-ji})\right]$ denote the pairwise marginalized advantage of agent $j$ given the joint action of $(j, i)$. The coupling coefficient $\mathcal{C}_i(s)$ for agent $i$ is defined as the sum of standard deviations of this marginalized advantage with respect to agent $i$'s policy, aggregated over all affected agents $j$:

$$\mathcal{C}_i(s) \triangleq \sum_{j\neq i} \mathbb{E}_{a_j\sim\pi^j_{\text{old}}}\left[\sqrt{\text{Var}_{a_i\sim\pi^i_{\text{old}}}\left(\bar{A}^{(j,i)}_{\boldsymbol{\pi}_{\text{old}}}(s, a_j, a_i)\right)}\right]. \tag{4}$$

Equipped with this metric, we derive a bound on the objective gap. We use Pearson $\chi^2$-divergence, $\chi^2_i(s) \triangleq \mathbb{E}_{a_i\sim\pi^i_{\text{old}}}\left[\left(\pi^i_{\text{new}}(a_i|s)/\pi^i_{\text{old}}(a_i|s) - 1\right)^2\right]$, to quantify the magnitude of the policy update.

**Theorem 3.1** (Bound on Independent-to-Joint Objective Gap)**.** *Consider a general-sum Markov game $\mathcal{G}$ with $\gamma < 1$ and per-agent rewards $|r_i(s, \boldsymbol{a})| \leq R_{\max}$ uniformly over agents and system sizes. Assume $\pi^i_{\text{new}} \ll \pi^i_{\text{old}}$. The independent-to-joint objective gap $\Delta\mathcal{L}$ is bounded by:*

$$\Delta\mathcal{L} \leq \sum_{i\in\mathcal{N}} \mathbb{E}_{s\sim d_{\boldsymbol{\pi}_{\text{old}}}}\left[\mathcal{C}_i(s)\cdot\sqrt{\chi^2_i(s)}\right] + R_N(\epsilon_\pi), \tag{5}$$

*where $\epsilon_\pi \triangleq \max_i \sup_s \sqrt{\chi^2_i(s)} < \infty$ and $R_N(\epsilon) \triangleq \frac{2NR_{\max}}{1-\gamma}\sum_{k=2}^{N}\binom{N}{k}\epsilon^k$. The residual is second order in the policy shift for fixed $N$, with an explicit dependence on $N$.*

*Proof Sketch.* The core derivation relies on an *advantage centering identity*: exploiting the zero-sum property of policy probability shifts (i.e., $\sum_{a_i}(\pi^i_{\text{new}} - \pi^i_{\text{old}}) = 0$), we subtract the baseline expected advantage $\mathbb{E}_{a_i}[\bar{A}^{(j,i)}]$. Applying the Cauchy-Schwarz inequality to this zero-mean formulation bounds the first-order error by the standard deviation $\sqrt{\text{Var}_{a_i}(\bar{A}^{(j,i)})}$. Bounding each remaining product of at least two policy shifts yields the explicit residual $R_N$. The complete proof is deferred to Appendix A. □

**Remark (Scalability in Large-Scale Self-Play).** Prior theoretical bounds (Sun et al., 2022; Su & Lu, 2024) for the independent-to-joint objective gap $\Delta\mathcal{L}$ scale linearly with the global agent count $N$:

$$\Delta\mathcal{L} \leq \sum_{i\in\mathcal{N}} \mathbb{E}_{s\sim d_{\boldsymbol{\pi}_{\text{old}}}}\left[(N-1)A_{\max}\cdot 2D_{\text{TV}}(\pi^i_{\text{new}}||\pi^i_{\text{old}})[s]\right], \tag{6}$$

where $A_{\text{max}} = \max_i \max_{s,\mathbf{a}} |A^i_\pi(s, \mathbf{a})|$. This global dependence forces the trust region prescribed by the bound to vanish at a rate of $\mathcal{O}(1/N)$ as the system scales.

In contrast, Theorem 3.1 replaces the global $(N-1)A_{\max}$ factor in the term linear in policy changes with state-dependent coupling $\mathcal{C}_i(s)$. Under the interaction sparsity condition in Appendix B.1, $\mathcal{C}_i(s)$ remains $\mathcal{O}(1)$ as the population grows. A telescoping argument under the same condition also bounds

the full objective gap for simultaneous updates with an $\mathcal{O}(1)$ coefficient on each agent's update (Appendix B). The corresponding per-agent trust region therefore need not vanish merely because more distant agents are added. The appendix also gives a conditional result for joint return.

### 3.2 Tractable Implementation: Vector Critic and Adaptive Clipping

Direct optimization of the $\chi^2$ penalty is difficult in deep RL. As a practical alternative, we use PPO clipping (Schulman et al., 2017) to control update size. Theorem 3.1 motivates a smaller clipping range when coupling $\mathcal{C}_i(s)$ is larger. If the realized probability ratio satisfies $|r_i(s, a_i) - 1| \leq \epsilon_i(s)$ for every action, then $\sqrt{\chi^2_i(s)} \leq \epsilon_i(s)$. PPO objective clipping encourages smaller updates, although it does not impose this ratio constraint by itself.

However, directly computing $\mathcal{C}_i(s)$ is impractical in large action spaces because the required counterfactual evaluations incur $\mathcal{O}(|A_i| \times |A_j|)$ cost. To retain computational efficiency, we interchange the expectation and variance operators to obtain a tractable surrogate:

$$\tilde{\mathcal{C}}_i(s) = \sum_{j \neq i} \sqrt{\mathrm{Var}_{a_i \sim \pi^i_{\mathrm{old}}} \left( \mathbb{E}_{a_j \sim \pi^j_{\mathrm{old}}} \left[ \bar{A}^{(j,i)}(s, a_j, a_i) \right] \right)}. \tag{7}$$

Minkowski's integral inequality gives $\tilde{\mathcal{C}}_i(s) \leq \mathcal{C}_i(s)$ (Appendix D). Taking the expectation over the evaluated agent's action before Cauchy–Schwarz also shows that $\tilde{\mathcal{C}}_i(s)$ directly bounds the first-order term in the gap under the old policy (Appendix A).

By reformulating the coupling as a variance of a marginal expectation, this surrogate avoids nested action sampling. We estimate it with a *vector critic*, parameterized by $\psi$. While Theorem 3.1 defines $\mathcal{C}_i(s)$ over all $N$ agents, practical estimation over distant agents can obscure dominant local coupling signals. We therefore restrict the critic to the $M$ most strongly interacting neighbors $\mathcal{N}_i$. To handle varying neighbor counts and preserve permutation equivariance, we use cross-attention (Iqbal & Sha, 2019) (Appendix E). Each neighbor feature $s_j$ ($j \in \mathcal{N}_i$) queries a shared representation of $(s, a_i)$, producing $A^{(j,i)}_\psi(s, s_j, a_i)$ as a prediction of agent $j$'s marginalized advantage. The vector critic is trained by regressing to Generalized Advantage Estimation (GAE) (Schulman et al., 2015b) targets $\hat{A}^j_{\mathrm{GAE}}$ computed during routine policy updates:

$$\mathcal{L}_{\mathrm{VC}}(\psi) = \mathbb{E}_{(s,\boldsymbol{a}) \sim \mathcal{B}} \left[ \frac{1}{N} \sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{N}_i} \left( A^{(j,i)}_\psi(s, s_j, a_i) - \hat{A}^j_{\mathrm{GAE}} \right)^2 \right]. \tag{8}$$

Since the critic does not observe the other agents' actions $\boldsymbol{a}_{-i}$, MSE regression estimates the conditional expectation of the GAE target given its inputs. This approximates the marginalized advantage required in Eq. equation 7.

We approximate the variance in $\tilde{\mathcal{C}}_i(s)$ by querying the vector critic using $K$ action samples drawn from $\pi^i_{\mathrm{old}}$. To establish a stable normalization baseline, we track an Exponential Moving Average (EMA) of these estimates across optimization steps: $\bar{\tilde{\mathcal{C}}} \leftarrow (1 - \alpha)\bar{\tilde{\mathcal{C}}} + \alpha \tilde{\mathcal{C}}^{(t)}$, where $\tilde{\mathcal{C}}^{(t)}$ is the empirical batch average at iteration $t$. Finally, the adaptive clipping for CARPO is defined as:

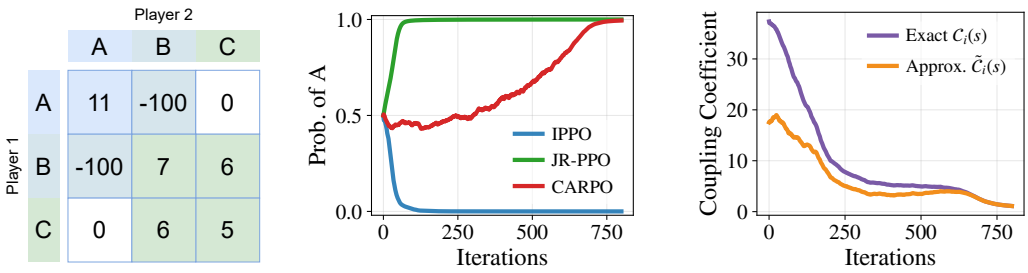
$$\epsilon_i(s) = \mathrm{clip} \left( \frac{\bar{\tilde{\mathcal{C}}}}{\tilde{\mathcal{C}}_i(s)} \cdot \epsilon_{\mathrm{base}}, \epsilon_{\min}, \epsilon_{\max} \right). \tag{9}$$

**Remark (Robustness to Lower-Bound Approximation).** Using the inverse of an underestimated coupling value directly could make the clipping range too wide. Our ratio-based design in Eq. equation 9 mitigates a common multiplicative bias. If $\tilde{\mathcal{C}}_i(s) \approx \alpha \mathcal{C}_i(s)$ across states, the EMA baseline absorbs the same factor $\alpha$. The clipping limits $[\epsilon_{\min}, \epsilon_{\max}]$ also restrict unusually wide updates. This argument does not cover errors that vary across states.

The training pseudocode is provided in Appendix F.

## 4 Mechanism Validation

This section examines how the state-dependent coupling coefficient shapes learning dynamics. We first deploy the analytical bound in a tabular game to show how coupling-adaptive regularization




(a) Payoff Matrix (b) Policy Trajectory $P(A)$ (c) Coupling Coefficient

Figure 2: **Mechanism analysis in the Stochastic Climbing Game.** (a) Payoff matrix. (b) IPPO prematurely converges to the suboptimal equilibrium ($P(A) = 0$). CARPO exhibits a "safe hesitation" phase to resist early penalties, reaching the same global optimum as the JR-PPO oracle. (c) The tractable approximation $\tilde{\mathcal{C}}_i(s)$ effectively tracks the exact coupling coefficient $\mathcal{C}_i(s)$, validating its use as a practical surrogate.

prevents premature convergence. We then evaluate CARPO in deep RL across coordination levels to show how adaptive clipping addresses the trade-off faced by a single static value.

### 4.1 Safe Hesitation in Coordination Traps

We analyze CARPO in the Stochastic Climbing Game (Claus & Boutilier, 1998). Fig. 2a shows the payoff matrix: a strict global optimum at $(A, A)$ $(+11)$ is surrounded by a severe miscoordination penalty $(-100)$ at $(A, B)$ and $(B, A)$, creating a risk-dominant suboptimal equilibrium at $(B, B)$ $(+7)$. A safe but highly suboptimal third action, $C$, further expands the exploration space. In this symmetric game, agents are initialized with an identical policy ($P(A) = 0.5, P(B) = P(C) = 0.25$). We optimize the analytical bound in this tabular setting to isolate the mechanism from function approximation errors.

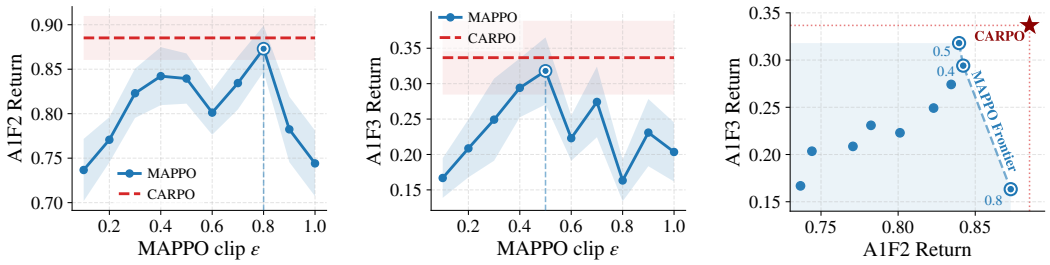
**Policy Dynamics and Safe Hesitation.** Fig. 2b plots the policy dynamics to explain how CARPO avoids this coordination trap. During early training, random opponent sampling frequently triggers the $-100$ penalty, producing steep empirical gradients. While the joint-ratio oracle (JR-PPO) maximizes the joint objective and identifies the optimal path, IPPO prematurely converges to the suboptimum ($P(A) \to 0$). CARPO, however, regulates its trust region via the coupling coefficient $\mathcal{C}_i(s)$. In this phase, high initial variance in marginalized advantages causes $\mathcal{C}_i(s)$ to spike (Fig. 2c), bounding the update step. This coupling-induced "safe hesitation" maintains $P(A) \approx 0.5$, **protecting probability mass on** $A$ and preserving the necessary exploration space for the joint policy to eventually lock onto the $(A, A)$ payoff. Beyond this specific initialization, evaluations across a broader range of initial states (Appendix H.3) confirm that CARPO expands the basin of attraction to the global optimum.

While this tabular analysis uses the exact $\mathcal{C}_i(s)$, Fig. 2c shows that the tractable approximation $\tilde{\mathcal{C}}_i(s)$ preserves the relative magnitude of exact $\mathcal{C}_i(s)$, validating its use as a scalable surrogate for deep RL.

### 4.2 Coupling-Adaptive Trust Regions for Diverse Coordination Levels

We evaluate the coupling-adaptive mechanism using Level-Based Foraging (LBF) with a $10 \times 10$ grid (Papoudakis et al., 2020). Existing bounds (Sun et al., 2022; Su & Lu, 2024) include a penalty proportional to agent count $N$. To expose what this count misses, we fix $N = 8$ and vary the coordination level, comparing A1F2 (requiring two agents to cooperate simultaneously) with A1F3 (requiring three). The same $N$ cannot capture the different coordination requirements of these tasks. We integrate CARPO into MAPPO to test the benefit of adapting to local coupling. Results for the fully decoupled A1F1 setting are in Appendix I.2.

**Optimal Clipping on Diverse Coordination Levels.** With the same agent count $N = 8$, MAPPO's best static clip shifts from $0.8$ in A1F2 to $0.5$ in A1F3 (Fig. 3a and 3b). The full clipping sweep is




(a) A1F2 (Medium Coordination) (b) A1F3 (High Coordination) (c) Cross-Environment Trade-off

Figure 3: **Performance across diverse coordination levels in LBF.** (a-b) MAPPO reaches its highest mean return at $\epsilon = 0.8$ in A1F2 and $\epsilon = 0.5$ in A1F3. (c) A single static clip faces a cross-environment trade-off. CARPO reaches a higher mean return than the best static MAPPO clip for each task by adapting $\epsilon$ to estimated coupling.

reported in Table 6. This shift shows that agent count alone cannot determine a suitable update range when coordination requirements change.

**Resolving the Trade-off via State-Level Adaptation.** Relying on a static clipping bound introduces a cross-environment dilemma. A permissive $\epsilon$ benefits A1F2 but can cause collapse in A1F3. A conservative $\epsilon$ stabilizes A1F3 but slows learning in A1F2. Fig. 3c shows this trade-off. By using state-dependent coupling to modulate clipping, CARPO reaches **the highest return in both tasks**.

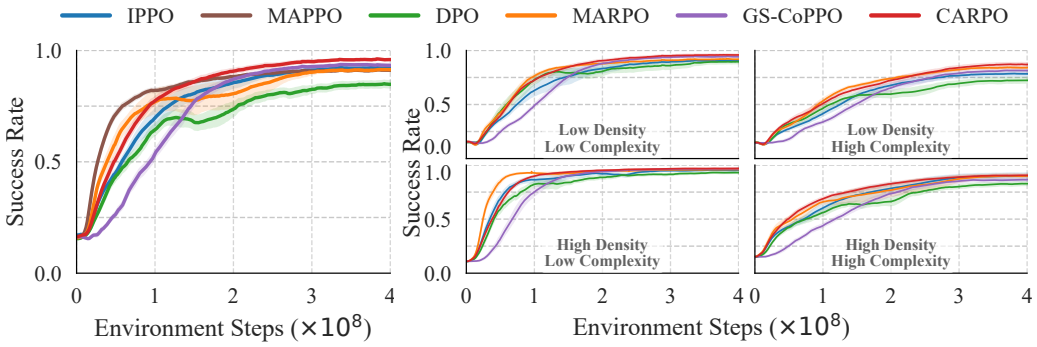
## 5 Experiments

Our empirical evaluation in PufferDrive (Cornelisse* et al., 2025) investigates CARPO's scalability and coupling mechanism in highly interactive autonomous driving. In this general-sum game, each agent navigates to a specific target destination while avoiding collisions and off-road boundaries. To ensure realistic multi-agent coordination rather than exploitative behaviors, we initialize environments using 10,000 real-world Waymo Open Dataset (Sun et al., 2020) scenarios and enforce kinematic action bounds (Appendix J.1).

**Baselines.** We compare our approach against five representative policy optimization baselines (details in Appendix J.3). (1) IPPO (De Witt et al., 2020) updates independent policies using standard PPO. (2) MAPPO (Yu et al., 2022) utilizes a centralized critic that ingests privileged environmental information to guide independent actor updates. (3) DPO (Su & Lu, 2024) optimizes a decentralized surrogate objective, employing an adaptive penalty to meet a predefined static KL target. (4) MARPO (Wu et al., 2026) incorporates a reflection mechanism by multiplying subsequent policy ratios to improve sample efficiency. (5) GS-CoPPO is a general-sum adaptation of CoPPO (Wu et al., 2021) that stabilizes joint updates by separately clipping other agents' ratio product. To ensure a fair comparison, all algorithms use identical actor networks with local observations. For value estimation, MAPPO's critic utilizes an expanded observation radius and accesses neighbors' goals, whereas the other baselines restrict their critics to the same local view as their actors. Furthermore, sequential-update algorithms such as HAPPO (Kuba et al., 2021) are excluded from our comparisons, as they are not applicable to general-sum games.

### 5.1 Performance in Complex Driving Environments

**Mixed Scenarios Performance.** We train all algorithms on a mixed dataset of 10,000 scenarios. As shown in Fig. 4a and Tab. 1, CARPO achieves the highest asymptotic success rate. Independent baselines learn rapidly initially but converge to lower success rates. GS-CoPPO updates agents jointly and learns more slowly. By adapting the clipping range inversely to estimated coupling, CARPO combines **rapid initial learning with stronger final performance**. It permits larger updates in loosely coupled states while restricting updates during complex interactions.




(a) Mixed Scenarios Performance (b) Performance under Diverse Densities and Complexities

Figure 4: **Performance in PufferDrive (10 seeds, mean ± std).** (a) Learning curves on the full scenario set, demonstrating CARPO's superior asymptotic performance. (b) Success rates across diverse traffic densities and interaction complexities. CARPO consistently outperforms baselines.

**Performance across Diverse Densities and Complexities.** We train all algorithms on four scenario subsets categorized by traffic density and intersection complexity (partitioning criteria in Appendix J.4). As shown in Fig. 4b, increased complexity reduces absolute success rates, while the relative rankings among baselines shift between subsets. CARPO achieves **the highest mean performance across all four configurations**. Table 11 compares CARPO with five fixed IPPO clipping values in the two low-density subsets. CARPO exceeds all five values in both subsets.

Table 1: Performance in PufferDrive. The reported values are averaged over the final 10% of training iterations. The column headers represent (Density, Complexity) of the traffic environment, respectively. **Bold** values show the best performance, underlined values the second best.

| Method | (Low, Low) | (High, Low) | (Low, High) | (High, High) | Mixed |
|---|---|---|---|---|---|
| IPPO | 0.906± 0.012 | 0.952± 0.005 | 0.784± 0.026 | 0.897± 0.027 | 0.921± 0.007 |
| MAPPO | 0.901± 0.003 | 0.962± 0.001 | 0.818± 0.017 | 0.898± 0.009 | 0.910± 0.003 |
| DPO | 0.895± 0.017 | 0.930± 0.007 | 0.722± 0.024 | 0.826± 0.016 | 0.848± 0.015 |
| MARPO | 0.919± 0.010 | 0.958± 0.003 | 0.841± 0.013 | 0.899± 0.016 | 0.912± 0.005 |
| GS-CoPPO | 0.942± 0.005 | 0.965± 0.002 | 0.814± 0.027 | 0.862± 0.015 | 0.935± 0.005 |
| CARPO (Ours) | **0.958**± 0.007 | **0.968**± 0.002 | **0.871**± 0.015 | **0.904**± 0.010 | **0.959**± 0.003 |

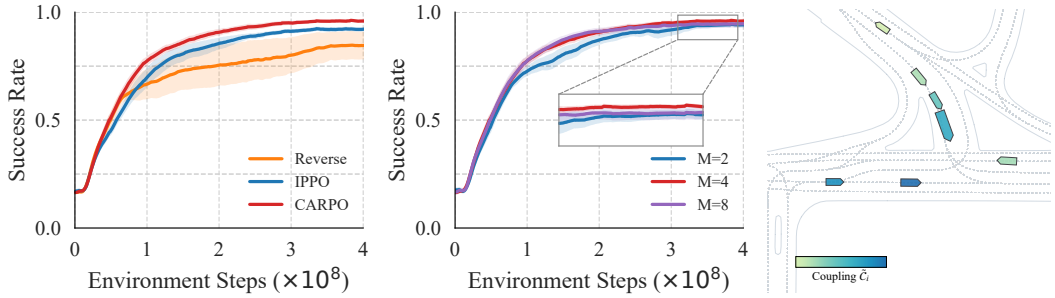
### 5.2 Ablation and Coupling Visualization

**Ablation.** We perform ablations to test the core components of CARPO. (1) **Inverse-CARPO** (Fig. 5a): We invert the ordering of clipping bounds within a batch, expanding the clipping range in highly coupled states instead of tightening it. The resulting performance drop supports the direction of CARPO's adaptation. States with stronger estimated coupling benefit from smaller update ranges. (2) **Vector Critic Neighborhood Size** ($M$) (Fig. 5b): We vary the maximum number of agents evaluated by the vector critic. The results show that a restricted $M$ omits highly-coupled neighbors, whereas an expansive $M$ accumulates estimation noise from loosely-coupled distant agents. Further ablations regarding the EMA normalization and the variance sampling size are provided in Appendix J.5.

**Visualizing State-Dependent Coupling.** We project the estimated coupling $\tilde{C}_i(s)$ onto an interchange scenario (Fig. 5c). High values align closely with physical traffic bottlenecks. Extended visualizations are provided in Appendix J.6.

## 6 Related Work

**Independent Learning and Non-stationarity.** Independent learning includes fully independent methods (De Witt et al., 2020) and Centralized Training with Decentralized Execution (CTDE) meth-




(a) Ablation: Inverse-CARPO (b) Ablation: Neighborhood Size $M$ (c) Coupling Visualization

Figure 5: **Ablations and coupling visualization.** (a-b) Ablations on the adaptation direction and neighborhood size. (c) Spatial visualization of estimated coupling.

ods (Yu et al., 2022; Lowe et al., 2017), which use a centralized critic but still perform independent actor updates. Concurrent updates make each agent's effective learning environment change over training (Papoudakis et al., 2019). Sequential schemes such as HAPPO (Kuba et al., 2021) address this with a one-by-one update order, sacrificing parallel efficiency. Stabilizing concurrent independent learning without this cost remains a challenge.

**Theoretical Bounds for Independent Learning.** To provide improvement guarantees for independent learning, recent studies connect non-stationarity with trust region constraints. Sun et al. (2022) explicitly identify a penalty that depends on the total agent count $N$, while DPO (Su & Lu, 2024) constructs a decentralized surrogate for joint policy improvement and retains this count dependence in its bound. As $N$ increases, these penalties can make the prescribed updates increasingly small. CARPO instead weights individual updates by state-dependent coupling and adapts the trust region accordingly. Under the interaction sparsity condition in Appendix B.1, the coupling coefficient remains bounded for each agent as $N$ grows.

## 7 Conclusion and Limitation

In this paper, we addressed the challenge of non-stationarity in independent self-play. Through an advantage centering identity, we showed that each policy update's contribution to the independent-to-joint objective gap depends on its magnitude and coupling strength. Building on this analysis, we proposed CARPO, which adapts trust regions according to estimated coupling. Empirical evaluations show that CARPO's "safe hesitation" prevents premature convergence in a coordination trap, while its adaptive clipping achieves strong performance across coordination levels and driving scenarios.

**Limitation** The $\mathcal{O}(1)$ scaling of each agent's coupling coefficient follows under the interaction sparsity condition. Without this condition, a single policy update may affect an increasing number of agents, and its contribution to the independent-to-joint objective gap may grow with $N$. CARPO is therefore best suited to large-scale systems with localized interaction structures.

### AI use statement

We used generative AI tools to improve the writing and grammar and to help debug code. We did not use them to formulate research questions, develop the theory or design experiments. We take responsibility for the manuscript.

## Reproducibility statement

Experimental settings, hyperparameters, and additional results are provided in Appendices G–J; implementation code is included in the supplementary materials.

## REFERENCES


Wei-Jer Chang, Akshay Rangesh, Kevin Joseph, Matthew Strong, Masayoshi Tomizuka, Yihan Hu, and Wei Zhan. Spacer: Self-play anchoring with centralized reference models. *arXiv preprint arXiv:2510.18060*, 2025.

Caroline Claus and Craig Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI*, 1998(746-752):2, 1998.

Daphne Cornelisse and Eugene Vinitsky. Human-compatible driving partners through data-regularized self-play reinforcement learning. *arXiv preprint arXiv:2403.19648*, 2024.

Daphne Cornelisse*, Spencer Cheng*, Pragnay Mandavilli, Julian Hunt, Kevin Joseph, Waël Doulazmi, Valentin Charraut, Aditya Gupta, Joseph Suarez, and Eugene Vinitsky. PufferDrive: A fast and friendly driving simulator for training and evaluating RL agents, 2025. URL `https://github.com/Emerge-Lab/PufferDrive`.

Marco Cusumano-Towner, David Hafner, Alex Hertzberg, Brody Huval, Aleksei Petrenko, Eugene Vinitsky, Erik Wijmans, Taylor Killian, Stuart Bowers, Ozan Sener, et al. Robust autonomy emerges from self-play. *arXiv preprint arXiv:2502.03349*, 2025.

Christian Schroeder De Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip HS Torr, Mingfei Sun, and Shimon Whiteson. Is independent learning all you need in the starcraft multi-agent challenge? *arXiv preprint arXiv:2011.09533*, 2020.

Joris Dinneweth, Abderrahmane Boubezoul, René Mandiau, and Stéphane Espié. Multi-agent reinforcement learning for autonomous vehicles: A survey. *Autonomous Intelligent Systems*, 2(1): 27, 2022.

Shariq Iqbal and Fei Sha. Actor-attention-critic for multi-agent reinforcement learning. In *International conference on machine learning*, pp. 2961–2970. PMLR, 2019.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pp. 267–274, 2002.

Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. Trust region policy optimisation in multi-agent reinforcement learning. *arXiv preprint arXiv:2109.11251*, 2021.

Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. *arXiv preprint arXiv:1702.03037*, 2017.

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pp. 157–163. Elsevier, 1994.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

Chengdong Ma, Aming Li, Yali Du, Hao Dong, and Yaodong Yang. Efficient and scalable reinforcement learning for large-scale network control. *Nature Machine Intelligence*, 6(9):1006–1020, 2024.

Georgios Papoudakis, Filippos Christianos, Arrasy Rahman, and Stefano V Albrecht. Dealing with non-stationarity in multi-agent deep reinforcement learning. *arXiv preprint arXiv:1906.04737*, 2019.

Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. *arXiv preprint arXiv:2006.07869*, 2020.

Guannan Qu, Yiheng Lin, Adam Wierman, and Na Li. Scalable multi-agent reinforcement learning for networked systems with average reward. *Advances in Neural Information Processing Systems*, 33:2074–2086, 2020.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015a.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015b.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Kefan Su and Zongqing Lu. A fully decentralized surrogate for multi-agent policy optimization. *Transactions on Machine Learning Research*, 2024.

Mingfei Sun, Sam Devlin, Jacob Beck, Katja Hofmann, and Shimon Whiteson. Trust region bounds for decentralized ppo under non-stationarity. *arXiv preprint arXiv:2202.00082*, 2022.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2446–2454, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Cuiling Wu, Yaozhong Gan, Junliang Xing, and Ying Fu. Marpo: A reflective policy optimization for multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29740–29748, 2026.

Zifan Wu, Chao Yu, Deheng Ye, Junge Zhang, Hankz Hankui Zhuo, et al. Coordinated proximal policy optimization. *Advances in Neural Information Processing Systems*, 34:26437–26448, 2021.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 35:24611–24624, 2022.

Yifan Zhong, Jakub Grudzien Kuba, Xidong Feng, Siyi Hu, Jiaming Ji, and Yaodong Yang. Heterogeneous-agent reinforcement learning. *Journal of Machine Learning Research*, 25(32): 1–67, 2024.

## APPENDIX TABLE OF CONTENTS


- **Appendix A:** Proof of Theorem 3.1 (Bound on Independent-to-Joint Objective Gap)
- **Appendix B:** Extended Theoretical Analysis on Joint Policy Improvement
- **Appendix C:** Policy Gradient Formulations of Surrogate Objectives
- **Appendix D:** Proof of the Lower Bound for the Coupling Coefficient Approximation
- **Appendix E:** Vector Critic Architecture
- **Appendix F:** Pseudocode of CARPO
- **Appendix G:** Cooperative Adaptive Cruise Control Experiment Details
- **Appendix H:** Implementation Details of Stochastic Climbing Game
- **Appendix I:** Implementation Details of Level-Based Foraging (LBF)
- **Appendix J:** Implementation Details of PufferDrive


## A PROOF OF THEOREM 3.1: BOUND ON INDEPENDENT-TO-JOINT OBJECTIVE GAP

In this section, we derive the bound on the independent-to-joint objective gap in Theorem 3.1 without using a maximum advantage coefficient for its first-order term.

*Proof.* **Step 1: Exact Decomposition of the Total Error.** We define the global objective gap $\Delta\mathcal{L}$ as the absolute difference between the joint and summed independent surrogate objectives:

$$\Delta\mathcal{L} \triangleq \left| \mathcal{L}_{\text{joint}}(\boldsymbol{\pi}_{\text{new}}) - \sum_{i\in\mathcal{N}} \mathcal{L}^i_{\text{ind}}(\pi^i_{\text{new}}) \right|. \tag{10}$$

We decompose the signed difference over evaluated agents $i$, so that $\Delta\mathcal{L} = |\sum_{i\in\mathcal{N}} E_i|$. Here, $E_i$ represents the deviation in agent $i$'s objective evaluation caused by the other agents $j$ updating from $\boldsymbol{\pi}^{-i}_{\text{old}}$ to $\boldsymbol{\pi}^{-i}_{\text{new}}$:

$$E_i \triangleq \mathbb{E}_{s\sim d_{\boldsymbol{\pi}_{\text{old}}}, a_i\sim\pi^i_{\text{new}}} \left[ \sum_{\boldsymbol{a}_{-i}} \left( \boldsymbol{\pi}^{-i}_{\text{new}}(\boldsymbol{a}_{-i}|s) - \boldsymbol{\pi}^{-i}_{\text{old}}(\boldsymbol{a}_{-i}|s) \right) A^i_{\boldsymbol{\pi}_{\text{old}}}(s, a_i, \boldsymbol{a}_{-i}) \right]. \tag{11}$$

**Step 2: Multinomial Expansion of Policy Differences.** Let the policy shift of any opponent $j$ be denoted as $\delta_j(a_j|s) \triangleq \pi^j_{\text{new}}(a_j|s) - \pi^j_{\text{old}}(a_j|s)$. By substituting $\pi^j_{\text{new}} = \pi^j_{\text{old}} + \delta_j$, the difference in the joint policy of all opponents can be decomposed via a multinomial expansion into first-order and higher-order terms (H.O.T.):

$$\boldsymbol{\pi}^{-i}_{\text{new}} - \boldsymbol{\pi}^{-i}_{\text{old}} = \underbrace{\sum_{j\neq i} \delta_j \prod_{k\neq i,j} \pi^k_{\text{old}}}_{\text{First-Order Terms}} + \underbrace{\left( \sum_{j,l\neq i, j<l} \delta_j \delta_l \prod_{k\neq i,j,l} \pi^k_{\text{old}} + \dots \right)}_{\text{H.O.T.}}. \tag{12}$$

Substituting this expansion into $E_i$ yields an exact decomposition $E_i = E_{i,1\text{st}} + E_{i,\text{H.O.T.}}$. The first-order error term isolates the pairwise impact of each agent $j$ on agent $i$:

$$E_{i,1\text{st}} \triangleq \sum_{j\neq i} \mathbb{E}_{s\sim d_{\boldsymbol{\pi}_{\text{old}}}, a_i\sim\pi^i_{\text{new}}} \left[ \sum_{a_j} \delta_j(a_j|s) \bar{A}^{(i,j)}_{\boldsymbol{\pi}_{\text{old}}}(s, a_i, a_j) \right], \tag{13}$$

where $\bar{A}^{(i,j)}_{\boldsymbol{\pi}_{\text{old}}}$ is the pairwise marginalized advantage. To avoid the approximation $\mathbb{E}_{a_i\sim\pi^i_{\text{new}}}[\cdot] \approx \mathbb{E}_{a_i\sim\pi^i_{\text{old}}}[\cdot]$, we further rewrite $E_{i,1\text{st}}$ as an exact identity by substituting $\pi^i_{\text{new}} = \pi^i_{\text{old}} + \delta_i$. This yields

$E_{i,\text{1st}} \equiv \hat{E}_{i,\text{1st}} + E_{i,\text{approx}}$. The residual difference is isolated into the second-order approximation error:

$$E_{i,\text{approx}} \triangleq \sum_{j \neq i} \mathbb{E}_{s \sim d_{\boldsymbol{\pi}_{\text{old}}}} \left[ \sum_{a_i} \delta_i(a_i|s) \sum_{a_j} \delta_j(a_j|s) \bar{A}^{(i,j)}_{\boldsymbol{\pi}_{\text{old}}}(s, a_i, a_j) \right], \tag{14}$$

which explicitly isolates the product of concurrent deviations $\delta_i$ and $\delta_j$. Consequently, the dominant first-order term, now exactly evaluated under the old policy distribution, is defined as:

$$\hat{E}_{i,\text{1st}} \triangleq \sum_{j \neq i} \mathbb{E}_{s \sim d_{\boldsymbol{\pi}_{\text{old}}}, a_i \sim \pi^i_{\text{old}}} \left[ \sum_{a_j} \delta_j(a_j|s) \bar{A}^{(i,j)}_{\boldsymbol{\pi}_{\text{old}}}(s, a_i, a_j) \right]. \tag{15}$$

**Step 3: The Advantage Identity Technique and Cauchy-Schwarz Bound.** To bound the absolute value $|\hat{E}_{i,\text{1st}}|$, we address the inner summation over $a_j$. Because the policy probability shift rigorously sums to zero, i.e., $\sum_{a_j} \delta_j(a_j|s) = 0$, we can formulate an *advantage centering identity* by subtracting the baseline expected advantage $\mathbb{E}_{a'_j \sim \pi^j_{\text{old}}}[\bar{A}^{(i,j)}(s, a_i, a'_j)]$ without altering the equality:

$$\sum_{a_j} \delta_j(a_j|s) \bar{A}^{(i,j)}(s, a_i, a_j) = \sum_{a_j} \delta_j(a_j|s) \left( \bar{A}^{(i,j)}(s, a_i, a_j) - \mathbb{E}_{a'_j \sim \pi^j_{\text{old}}}[\bar{A}^{(i,j)}(s, a_i, a'_j)] \right). \tag{16}$$

Applying the Cauchy-Schwarz inequality to this zero-mean formulation, we obtain:

$$\begin{aligned} \left| \sum_{a_j} \delta_j \bar{A}^{(i,j)} \right| &\leq \sqrt{\sum_{a_j} \frac{\delta_j(a_j|s)^2}{\pi^j_{\text{old}}(a_j|s)}} \cdot \sqrt{\sum_{a_j} \pi^j_{\text{old}}(a_j|s) \left( \bar{A}^{(i,j)} - \mathbb{E}_{a'_j}[\bar{A}^{(i,j)}] \right)^2} \\ &= \sqrt{\chi^2(\pi^j_{\text{new}} \| \pi^j_{\text{old}})[s]} \cdot \sqrt{\text{Var}_{a_j \sim \pi^j_{\text{old}}} \left( \bar{A}^{(i,j)}_{\boldsymbol{\pi}_{\text{old}}}(s, a_i, a_j) \right)}. \end{aligned} \tag{17}$$

Taking the expectation over $s$ and $a_i$, the bound for the evaluated agent $i$ is:

$$|\hat{E}_{i,\text{1st}}| \leq \sum_{j \neq i} \mathbb{E}_{s \sim d_{\boldsymbol{\pi}_{\text{old}}}} \left[ \sqrt{\chi^2_j(s)} \cdot \mathbb{E}_{a_i \sim \pi^i_{\text{old}}} \left[ \sqrt{\text{Var}_{a_j \sim \pi^j_{\text{old}}} \left( \bar{A}^{(i,j)}(s, a_i, a_j) \right)} \right] \right]. \tag{18}$$

**Step 4: Reindexing and Extracting the Coupling Coefficient.** The total first-order approximation error is bounded by the sum over all evaluated agents $i$: $\sum_i |\hat{E}_{i,\text{1st}}|$. To reveal the system-level impact of each updating agent, we swap the order of summation, transitioning from $\sum_i \sum_{j \neq i}$ to $\sum_j \sum_{i \neq j}$:

$$\sum_{i \in \mathcal{N}} |\hat{E}_{i,\text{1st}}| \leq \sum_{j \in \mathcal{N}} \mathbb{E}_{s \sim d_{\boldsymbol{\pi}_{\text{old}}}} \left[ \sqrt{\chi^2_j(s)} \underbrace{\sum_{i \neq j} \mathbb{E}_{a_i \sim \pi^i_{\text{old}}} \left[ \sqrt{\text{Var}_{a_j \sim \pi^j_{\text{old}}} \left( \bar{A}^{(i,j)}(s, a_i, a_j) \right)} \right]}_{= \mathcal{C}_j(s)} \right]. \tag{19}$$

Notice that the inner summation recovers the definition of the state-wise coupling coefficient $\mathcal{C}_j(s)$ (Definition 3.1), representing the aggregate variance impact that agent $j$'s policy shift imposes on the entire system. Since $j$ is merely a dummy index for the summation over all agents $\mathcal{N}$, we can equivalently write this sum using index $i$, fully aligning with the notation in Theorem 3.1.

The first-order term also admits a tighter coefficient. Before applying Cauchy–Schwarz in Step 3, exchange the old-policy expectation over the evaluated agent's action with the sum over the updating agent's action. Center the resulting function of the updating action and then apply Cauchy–Schwarz to obtain

$$\left| \sum_i \hat{E}_{i,\text{1st}} \right| \leq \sum_j \mathbb{E}_{s \sim d_{\boldsymbol{\pi}_{\text{old}}}} \left[ \tilde{\mathcal{C}}_j(s) \sqrt{\chi^2_j(s)} \right], \tag{20}$$

where $\tilde{\mathcal{C}}_j(s)$ is defined in Eq. equation 7. Thus the tractable surrogate directly controls the old-policy first-order term, even though it is a lower bound on $\mathcal{C}_j(s)$. Higher-order interactions and learned-estimator error remain separate.

**Step 5: Bounding the Higher-Order Terms.** The reward bound gives $|A^i_{\boldsymbol{\pi}_{\text{old}}}(s, \boldsymbol{a})| \leq B \triangleq 2R_{\max}/(1-\gamma)$. For each agent $j$, write $x_j(a_j|s) = \pi^j_{\text{new}}(a_j|s)/\pi^j_{\text{old}}(a_j|s) - 1$. The exact per-agent gap is the old-policy expectation of $A^i[(\prod_j(1+x_j)) - (1+x_i)]$. Its degree-one terms are precisely those bounded in Steps 3–4; every remaining monomial contains at least two distinct $x_j$. At each state, product-policy factorization, $|A^i| \leq B$, and Cauchy–Schwarz give, for any subset $S$ of $k$ agents,

$$\left| \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_{\text{old}}} \left[ A^i(s,\boldsymbol{a}) \prod_{j\in S} x_j(a_j|s) \right] \right| \leq B \prod_{j\in S} \mathbb{E}_{a_j\sim\pi^j_{\text{old}}} |x_j| \leq B \prod_{j\in S} \sqrt{\chi^2_j(s)} \leq B\epsilon^k_\pi. \tag{21}$$

There are at most $\binom{N}{k}$ degree-$k$ monomials for each evaluated agent. Summing over agents and degrees yields the explicit global residual

$$R_N(\epsilon_\pi) = NB \sum_{k=2}^{N} \binom{N}{k} \epsilon^k_\pi. \tag{22}$$

For fixed $N$, this residual is second order with an $N$-dependent coefficient. Combining it with Step 4 proves Theorem 3.1.

□

**Remark (Alternative Bound via Total Variation).** As demonstrated in Step 3, we utilize the Cauchy-Schwarz inequality to bound the objective gap via the $\chi^2$-divergence and the advantage variance. Alternatively, by bounding the error using the supremum norm, we can derive an analogous bound under the Total Variation (TV) divergence.

Exploiting the zero-sum property ($\sum_{a_j} \delta_j = 0$), the first-order error term can be directly bounded by the range of the marginalized advantage:

$$\left| \sum_{a_j} \delta_j(a_j|s) \bar{A}^{(i,j)}(s, a_i, a_j) \right| \leq \left( \max_{a_j} \bar{A}^{(i,j)}(s, a_i, a_j) - \min_{a_j} \bar{A}^{(i,j)}(s, a_i, a_j) \right) \cdot D_{\text{TV}}(\pi^j_{\text{new}} || \pi^j_{\text{old}})[s]. \tag{23}$$

Following the identical index-swapping procedure in Step 4, we can define a TV-based local coupling coefficient $\mathcal{C}^{\text{TV}}_i(s) \triangleq \sum_{j\neq i} \mathbb{E}_{a_j\sim\pi^j_{\text{old}}}[\max_{a_i} \bar{A}^{(j,i)}(s, a_j, a_i) - \min_{a_i} \bar{A}^{(j,i)}(s, a_j, a_i)]$. The global objective gap is then bounded by:

$$\Delta\mathcal{L} \leq \sum_{i\in\mathcal{N}} \mathbb{E}_{s\sim d_{\boldsymbol{\pi}_{\text{old}}}} \left[ \mathcal{C}^{\text{TV}}_i(s) \cdot D_{\text{TV}}(\pi^i_{\text{new}} || \pi^i_{\text{old}})[s] \right] + R_N(\epsilon_\pi). \tag{24}$$

Under Assumption B.1, the advantage range decays with distance, giving a local first-order coefficient for either TV or $\chi^2$ divergence.

# B Extended Theoretical Analysis on Joint Policy Improvement

Theorem 3.1 isolates a state-dependent coupling coefficient for each agent. We now formalize when this coefficient remains bounded as the population grows, then examine the additional condition needed to bound joint return improvement.

## B.1 Assumption on Interaction Sparsity in Networked MDPs

To formalize the physical intuition of interaction sparsity, we place agents on a graph with shortest-path distance $d(i, j)$. In traffic and other spatial systems, the immediate effects of actions are often concentrated near the acting agent. The advantage also reflects future actions, however, so local rewards and transitions alone do not guarantee that distant agents have little influence. Our condition therefore measures how an agent's action changes another agent's advantage after accounting for the policies of the other agents.

**Assumption B.1** (Interaction Sparsity)**.** There are constants $c_d, L > 0$, $d \geq 1$, and $\phi \in (0,1)$, independent of $N$, such that each distance-$m$ shell $\mathcal{N}_i^m = \{j : d(i,j) = m\}$ satisfies $|\mathcal{N}_i^m| \leq c_d m^d$. For any $i \neq j$, state $s$, action $a_j$, and product distribution $\boldsymbol{\mu}$ whose components are selected from the old or new policies, define

$$\bar{A}^{(j,i)}_{\boldsymbol{\mu}}(s, a_j, a_i) \triangleq \mathbb{E}_{\boldsymbol{a}_{-ji} \sim \boldsymbol{\mu}^{-ji}}[A^j_{\boldsymbol{\pi}_{\text{old}}}(s, a_j, a_i, \boldsymbol{a}_{-ji})]. \tag{25}$$

We assume the uniform sensitivity bound

$$\sup_{a_i, a_i'} |\bar{A}^{(j,i)}_{\boldsymbol{\mu}}(s, a_j, a_i) - \bar{A}^{(j,i)}_{\boldsymbol{\mu}}(s, a_j, a_i')| \leq L\phi^{d(i,j)}. \tag{26}$$

The condition measures how a change in agent $i$'s action affects agent $j$'s advantage after the other actions have been marginalized. It allows policies with global observations, but requires their induced influence to decay with distance. The condition covers the old-policy marginalization in Definition 3.1 and the hybrid policies used to bound the gap under simultaneous updates. In a one-step fixed-radius reward model, sensitivity vanishes beyond the reward neighborhood. Long-horizon systems also require decay of policy-mediated propagation; localized policies and transition influence provide related sufficient-condition templates (Qu et al., 2020).

**Lemma B.1** (Objective Gap Bound under Interaction Sparsity)**.** *Under Assumption B.1, consider old and new product policies with $\pi^i_{\text{new}}(\cdot|s) \ll \pi^i_{\text{old}}(\cdot|s)$ for every agent and $d_{\boldsymbol{\pi}_{\text{old}}}$-almost every state, and $\mathbb{E}_{s \sim d_{\boldsymbol{\pi}_{\text{old}}}} \sqrt{\chi^2_i(s)} < \infty$. Then the coupling coefficient satisfies*

$$\mathcal{C}_i(s) \leq K_{\text{loc}} \triangleq \frac{Lc_d}{2} \sum_{m=1}^{\infty} m^d \phi^m = \mathcal{O}(1). \tag{27}$$

*The sensitivity bound also holds at each step of a telescoping expansion, giving a bound on the exact gap without a separate higher-order residual:*

$$\Delta\mathcal{L} \leq K_{\text{loc}} \sum_{i \in \mathcal{N}} \mathbb{E}_{s \sim d_{\boldsymbol{\pi}_{\text{old}}}}[\sqrt{\chi^2_i(s)}]. \tag{28}$$

*Here $K_{\text{loc}}$ is a uniform upper bound used in the proof; CARPO still adapts to the state-dependent $\mathcal{C}_i(s)$. The global gap remains a sum over agents, with the state distribution fixed at $d_{\boldsymbol{\pi}_{\text{old}}}$.*

*Proof.* The range in Eq. equation 26 bounds the standard deviation with respect to $a_i$ by $L\phi^{d(i,j)}/2$. Summing over $j$ and using the shell-growth condition proves the first claim.

For the gap, fix an evaluated agent $j$ and order its opponents as $i_1, \ldots, i_{N-1}$. Let $\boldsymbol{\mu}_{j,t}$ use $\pi^j_{\text{new}}$ for agent $j$, $\pi^{i_r}_{\text{new}}$ for $r \leq t$, and $\pi^{i_r}_{\text{old}}$ for $r > t$. Define $\mathcal{L}_{j,t} = \mathbb{E}_{s \sim d_{\boldsymbol{\pi}_{\text{old}}}, \boldsymbol{a} \sim \boldsymbol{\mu}_{j,t}}[A^j_{\boldsymbol{\pi}_{\text{old}}}(s, \boldsymbol{a})]$. Thus $\mathcal{L}_{j,0} = \mathcal{L}^j_{\text{ind}}$ and $\mathcal{L}_{j,N-1} = \mathcal{L}^j_{\text{joint}}$, giving

$$\Delta\mathcal{L} = \left| \sum_j \sum_{t=1}^{N-1} (\mathcal{L}_{j,t} - \mathcal{L}_{j,t-1}) \right|. \tag{29}$$

At step $t$, only opponent $i_t$ changes policy. Write $g_{j,t}(s, a_j, a_{i_t}) = \bar{A}^{(j,i_t)}_{\boldsymbol{\mu}_{j,t}}(s, a_j, a_{i_t})$; the distribution over the other actions is the same at steps $t-1$ and $t$. Thus

$$\mathcal{L}_{j,t} - \mathcal{L}_{j,t-1} = \mathbb{E}_{s \sim d_{\boldsymbol{\pi}_{\text{old}}}, a_j \sim \pi^j_{\text{new}}} \left[ \mathbb{E}_{a_{i_t} \sim \pi^{i_t}_{\text{new}}} g_{j,t} - \mathbb{E}_{a_{i_t} \sim \pi^{i_t}_{\text{old}}} g_{j,t} \right]. \tag{30}$$

Equation equation 26 bounds the range of $g_{j,t}$ as a function of $a_{i_t}$. Applying the range–TV inequality gives

$$\begin{aligned} |\mathcal{L}_{j,t} - \mathcal{L}_{j,t-1}| &\leq L\phi^{d(i_t,j)} \mathbb{E}_{s \sim d_{\boldsymbol{\pi}_{\text{old}}}} D_{\text{TV}}(\pi^{i_t}_{\text{new}}, \pi^{i_t}_{\text{old}})[s] \\ &\leq \frac{L\phi^{d(i_t,j)}}{2} \mathbb{E}_{s \sim d_{\boldsymbol{\pi}_{\text{old}}}} \sqrt{\chi^2_{i_t}(s)}. \end{aligned} \tag{31}$$

Summing over the steps and evaluated agents, exchanging the sums over $j$ and $i_t$, and applying the shell-growth bound gives Eq. equation 28. □

### B.2 Conditional Localized Bound for Joint Return

The analysis of the objective gap above fixes the old state distribution. A bound on the true return additionally requires control of the state distribution shift for the policy class.

**Assumption B.2** (Localized State-Distribution Shift)**.** For the policy updates considered here, there are coefficients $0 \le \kappa_i \le \kappa$, independent of $N$, such that each agent's performance difference satisfies

$$J_i(\boldsymbol{\pi}_{\text{new}}) - J_i(\boldsymbol{\pi}_{\text{old}}) \ge \frac{1}{1-\gamma}\left[\mathcal{L}^i_{\text{joint}}(\boldsymbol{\pi}_{\text{new}}) - \kappa_i \sum_{j\in\mathcal{N}} \phi^{d(i,j)} \mathbb{E}_{s\sim d_{\boldsymbol{\pi}_{\text{old}}}} \sqrt{\chi^2_j(s)}\right]. \tag{32}$$

This condition describes spatial decay of the state-shift penalty, separately from action sensitivity at a fixed state. It can be checked through a localized transition-influence analysis for a specified policy class.

**Corollary B.2** (Conditional Joint Improvement Bound)**.** *Under Assumptions B.1 and B.2, define* $\nu^j_{\text{shift}} \triangleq \sum_i \kappa_i \phi^{d(i,j)}$*. Then*

$$J(\boldsymbol{\pi}_{\text{new}}) - J(\boldsymbol{\pi}_{\text{old}}) \ge \frac{1}{1-\gamma}\left[\sum_i \mathcal{L}^i_{\text{ind}}(\pi^i_{\text{new}}) - \sum_i \mathbb{E}_{s\sim d_{\boldsymbol{\pi}_{\text{old}}}}\left[(K_{\text{loc}} + \nu^i_{\text{shift}})\sqrt{\chi^2_i(s)}\right]\right]. \tag{33}$$

*Both* $K_{\text{loc}}$ *and* $\nu^i_{\text{shift}}$ *are bounded independently of* $N$*. The total return and penalty still sum over agents.*

*Proof.* Sum Eq. equation 32 over evaluated agents and exchange the indices in its state-shift penalty. Lemma B.1 bounds the difference between the joint surrogate and the sum of independent surrogates by $K_{\text{loc}} \sum_i \mathbb{E}_s \sqrt{\chi^2_i(s)}$. Combining the two inequalities proves Eq. equation 33. Finally, polynomial shell growth gives

$$\nu^i_{\text{shift}} \le \kappa\left(1 + c_d \sum_{m=1}^{\infty} m^d \phi^m\right) = \mathcal{O}(1). \tag{34}$$

□

The state-dependent $\mathcal{C}_i(s)$ in Theorem 3.1 explains CARPO's adaptation, while the uniform $K_{\text{loc}}$ in Eq. equation 28 bounds the objective gap under interaction sparsity.

## C Policy Gradient Formulations of Surrogate Objectives

In the self-play setting, all agents share an identical policy parameter. We define $\boldsymbol{\pi}_\theta$ as the joint policy where each component $\pi^i = \pi^i_\theta$. This section provides the formal derivation of the gradients for the joint and independent surrogate objectives with respect to the shared parameters $\theta$.

### C.1 Gradient of the Joint Surrogate Objective

The joint surrogate objective is derived from the performance difference lemma (Kakade & Langford, 2002) applied to the joint return $J(\boldsymbol{\pi}) = \sum_{i\in\mathcal{N}} J_i(\boldsymbol{\pi})$. Starting from the local approximation at $\theta_{\text{old}}$:

$$\mathcal{L}_{\text{joint}}(\theta) = \mathbb{E}_{s\sim d_{\theta_{\text{old}}}, \boldsymbol{a}\sim\boldsymbol{\pi}_\theta}\left[\sum_{i\in\mathcal{N}} A^i_{\theta_{\text{old}}}(s, \boldsymbol{a})\right]. \tag{35}$$

Taking the gradient with respect to $\theta$ and evaluating at $\theta = \theta_{\text{old}}$, we use the property $\nabla_\theta \boldsymbol{\pi}_\theta(\boldsymbol{a}|s) = \boldsymbol{\pi}_\theta(\boldsymbol{a}|s) \sum_{j\in\mathcal{N}} \nabla_\theta \log \pi_\theta^j(a^j|s)$:

$$\nabla_\theta \mathcal{L}_{\text{joint}}(\theta)\big|_{\theta=\theta_{\text{old}}} = \mathbb{E}_{s\sim d_{\theta_{\text{old}}}, \boldsymbol{a}\sim\boldsymbol{\pi}_{\theta_{\text{old}}}} \left[ \sum_{i\in\mathcal{N}} A^i_{\theta_{\text{old}}}(s,\boldsymbol{a}) \nabla_\theta \log \boldsymbol{\pi}_\theta(\boldsymbol{a}|s)\big|_{\theta_{\text{old}}} \right] \tag{36}$$

$$= \mathbb{E}_{s\sim d_{\theta_{\text{old}}}, \boldsymbol{a}\sim\boldsymbol{\pi}_{\theta_{\text{old}}}} \left[ \left( \sum_{i\in\mathcal{N}} A^i_{\theta_{\text{old}}}(s,\boldsymbol{a}) \right) \left( \sum_{j\in\mathcal{N}} \nabla_\theta \log \pi_\theta^j(a^j|s) \right) \right] \tag{37}$$

$$= \sum_{i\in\mathcal{N}} \sum_{j\in\mathcal{N}} \mathbb{E}_{s\sim d_{\theta_{\text{old}}}, \boldsymbol{a}\sim\boldsymbol{\pi}_{\theta_{\text{old}}}} \left[ A^i_{\theta_{\text{old}}}(s,\boldsymbol{a}) \nabla_\theta \log \pi_\theta^j(a^j|s) \right]. \tag{38}$$

Expanding the product, the gradient contains cross-terms $\mathbb{E}[A^i \nabla_\theta \log \pi_\theta^j(a^j|s)]$ for all pairs $(i,j) \in \mathcal{N} \times \mathcal{N}$.

### C.2 Gradient of the Independent Surrogate Objective

The independent surrogate objective approximates the joint improvement by assuming that for each agent $i$, the other agents' actions are sampled from the stationary policy $\pi_{\theta_{\text{old}}}$. The objective for a single agent $i$ is:

$$\mathcal{L}^i_{\text{ind}}(\theta) = \mathbb{E}_{s\sim d_{\theta_{\text{old}}}, a^i\sim\pi^i_\theta, \boldsymbol{a}^{-i}\sim\boldsymbol{\pi}^{-i}_{\theta_{\text{old}}}} \left[ A^i_{\theta_{\text{old}}}(s, a^i, \boldsymbol{a}^{-i}) \right]. \tag{39}$$

The gradient of the summed independent objectives $\sum_{i\in\mathcal{N}} \mathcal{L}^i_{\text{ind}}(\theta)$ with respect to $\theta$ is:

$$\nabla_\theta \sum_{i\in\mathcal{N}} \mathcal{L}^i_{\text{ind}}(\theta)\big|_{\theta=\theta_{\text{old}}} = \sum_{i\in\mathcal{N}} \mathbb{E}_{s\sim d_{\theta_{\text{old}}}, a^i\sim\pi^i_{\theta_{\text{old}}}, \boldsymbol{a}^{-i}\sim\boldsymbol{\pi}^{-i}_{\theta_{\text{old}}}} \left[ A^i_{\theta_{\text{old}}}(s, a^i, \boldsymbol{a}^{-i}) \nabla_\theta \log \pi^i_\theta(a^i|s) \right] \tag{40}$$

$$= \sum_{i\in\mathcal{N}} \mathbb{E}_{s\sim d_{\theta_{\text{old}}}, \boldsymbol{a}\sim\boldsymbol{\pi}_{\theta_{\text{old}}}} \left[ A^i_{\theta_{\text{old}}}(s,\boldsymbol{a}) \nabla_\theta \log \pi^i_\theta(a^i|s) \right]. \tag{41}$$

In this formulation, the gradient only includes terms where the advantage $A^i$ and the policy gradient $\nabla_\theta \log \pi_\theta(a^i)$ belong to the same agent index $i$. All cross-agent terms $(i \neq j)$ are omitted.

## D Proof of the Lower Bound for the Coupling Coefficient Approximation

In this section, we formally prove that the approximated coupling coefficient theoretically serves as a rigorous lower bound to the exact coupling coefficient $\mathcal{C}_i(s)$ presented in Definition 3.1.

**Proposition D.1.** *Let $\mathcal{C}_i(s)$ be the exact coupling coefficient and $\tilde{\mathcal{C}}_i(s)$ denote its approximation via the interchange of expectation and variance:*

$$\tilde{\mathcal{C}}_i(s) \triangleq \sum_{j\neq i} \sqrt{\text{Var}_{a_i\sim\pi^i_{\text{old}}} \left( \mathbb{E}_{a_j\sim\pi^j_{\text{old}}} \left[ \bar{A}^{(j,i)}_{\boldsymbol{\pi}_{\text{old}}}(s, a_j, a_i) \right] \right)}. \tag{42}$$

*Then, for any state $s$, $\tilde{\mathcal{C}}_i(s) \leq \mathcal{C}_i(s)$.*

*Proof.* For notational simplicity, let the marginalized advantage be denoted as $f(a_i, a_j) = \bar{A}^{(j,i)}_{\boldsymbol{\pi}_{\text{old}}}(s, a_j, a_i)$, and let $\mu(a_j) = \mathbb{E}_{a_i\sim\pi^i_{\text{old}}}[f(a_i, a_j)]$.

The standard deviation of a random variable corresponds to its $L_2$ norm after centering, i.e., $\sqrt{\text{Var}_{a_i}(X)} = \|X - \mathbb{E}[X]\|_{L_2}$. Using this equivalence, the exact coupling coefficient $\mathcal{C}_i(s)$ aggregates the expected standard deviation:

$$\mathcal{C}_i(s) = \sum_{j\neq i} \mathbb{E}_{a_j\sim\pi^j_{\text{old}}} \left[ \|f(\cdot, a_j) - \mu(a_j)\|_{L_2} \right]. \tag{43}$$

Conversely, the approximated coefficient $\tilde{\mathcal{C}}_i(s)$ computes the standard deviation of the expected advantage, which translates to the $L_2$ norm of the expectation:

$$\tilde{\mathcal{C}}_i(s) = \sum_{j \neq i} \left\| \mathbb{E}_{a_j \sim \pi^j_{\text{old}}} \left[ f(\cdot, a_j) - \mu(a_j) \right] \right\|_{L_2} . \tag{44}$$

According to Minkowski's integral inequality, the $L_2$ norm of an expected value is rigorously upper-bounded by the expected value of the $L_2$ norm. Applying this theorem to each term in the summation yields:

$$\left\| \mathbb{E}_{a_j \sim \pi^j_{\text{old}}} \left[ f(\cdot, a_j) - \mu(a_j) \right] \right\|_{L_2} \leq \mathbb{E}_{a_j \sim \pi^j_{\text{old}}} \left[ \left\| f(\cdot, a_j) - \mu(a_j) \right\|_{L_2} \right] . \tag{45}$$

Summing both sides of the inequality over all agents $j \neq i$, we conclude that $\tilde{\mathcal{C}}_i(s) \leq \mathcal{C}_i(s)$. □

## E Vector Critic Architecture

To process a varying number of interacting agents and maintain permutation equivariance, the vector critic employs a cross-attention mechanism. The feature of each individual neighbor, $s_j$ ($j \in \mathcal{N}_i$), is linearly projected to form an independent query vector. Meanwhile, the global contextual information, comprising the full state $s = (\{s_k\}_{k \in \mathcal{N}}, s_{\text{meta}})$ and agent $i$'s action $a_i$, is processed by a shared network to generate the key and value vectors (Vaswani et al., 2017). Because the cross-attention mechanism evaluates each query independently against the shared key-value memory (without inter-query attention), the evaluation for agent $j$ depends on $s_j$, $s$, and $a_i$. This structure not only guarantees that any permutation in the input sequence naturally translates to the identical permutation in the output vector, but also justifies the exact evaluation notation $A^{(j,i)}_{\psi}(s, s_j, a_i)$ used in Eq. equation 8.

While the cross-attention mechanism scales to an arbitrary number of agents, incorporating distant, loosely-coupled agents in sparse interaction environments can introduce spurious variance. Therefore, we define the neighborhood set $\mathcal{N}_i$ based on the environmental topology. In small-scale or fully-connected tasks (e.g., the Climbing Game and LBF), $\mathcal{N}_i$ encompasses all other agents ($|\mathcal{N}_i| = N-1$). In the large-scale PufferDrive, we limit the receptive field to the $M$ nearest neighbors to filter out interaction noise. If fewer than $M$ neighbors are present, we pad the inputs to length $M$ and apply a mask to exclude these padded terms during the loss computation.

The architecture of this vector critic, including the processing of contextual information and neighbor features, is illustrated in Fig. 6. The $j$-th element of the resulting output vector, $A^{(j,i)}_{\psi}$, estimates the marginalized advantage of agent $j$ conditioned on the state $s$ and agent $i$'s action $a_i$.

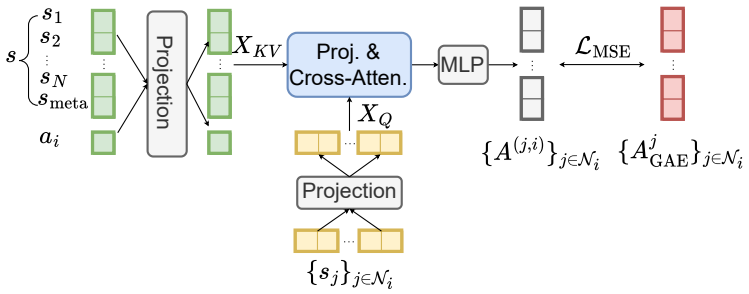



Figure 6: **The Vector Critic Architecture.** The network transforms contextual features $X_{KV}$ (state $s$ and action $a_i$) and neighbor queries $X_Q$ into a set of marginalized advantages $\{A^{(j,i)}\}_{j \in \mathcal{N}_i}$ through a cross-attention mechanism. This design supports a varying number of neighbors while ensuring permutation equivariance.

## F Pseudocode of CARPO

Algorithm 1 provides the full training pseudocode of CARPO. The algorithm augments a standard PPO update with two additional steps: estimating $\tilde{\mathcal{C}}_i(s)$ in Eq. equation 7, and adapting the clipping

**Algorithm 1** CARPO

**Require:** Old policy $\boldsymbol{\pi}_{\text{old}}$, value critic $V_\phi$, vector critic $A_\psi$, sample count $K$, $\epsilon_{\text{base}}, \epsilon_{\min}, \epsilon_{\max}, \alpha$

Initialize $\theta, \phi, \psi$ and EMA baseline $\bar{\tilde{\mathcal{C}}} > 0$
**while** not converged **do**
  Collect an on-policy batch $\mathcal{B}$ under $\boldsymbol{\pi}_{\text{old}}$
  Compute GAE targets $\hat{A}^i_{\text{GAE}}$ and return targets $\hat{R}^i$
  **for** each $(s_t, \boldsymbol{a}_t) \in \mathcal{B}$ and agent $i$ **do**
    Sample $\{\hat{a}_i^{(k)}\}_{k=1}^K \sim \pi^i_{\text{old}}(\cdot \mid o_i^t)$
    Estimate $\tilde{\mathcal{C}}_i(s_t)$ by Eq. equation 7
    Set $\epsilon_i(s_t)$ by Eq. equation 9
  **end for**
  Update $\bar{\tilde{\mathcal{C}}} \leftarrow (1-\alpha)\bar{\tilde{\mathcal{C}}} + \alpha\, \mathbb{E}_{(s_t,i)\sim\mathcal{B}}[\tilde{\mathcal{C}}_i(s_t)]$
  **for** each PPO minibatch update **do**
    Compute importance ratio $r_i^t(\theta) = \frac{\pi^i_\theta(a_i^t|o_i^t)}{\pi^i_{\text{old}}(a_i^t|o_i^t)}$
    Maximize the adaptive clipped surrogate

$$\mathcal{L}_{\text{actor}}(\theta) = \mathbb{E}\Big[\min\Big(r_i^t(\theta)\hat{A}^i_{\text{GAE}},\ \text{clip}\big(r_i^t(\theta), 1-\epsilon_i(s_t), 1+\epsilon_i(s_t)\big)\hat{A}^i_{\text{GAE}}\Big)\Big]$$

    Minimize the value loss $\mathcal{L}_{\text{Critic}}(\phi) = \mathbb{E}\Big[\big(V^i_\phi(s_t) - \hat{R}^i_t\big)^2\Big]$
    Minimize the vector-critic loss in Eq. equation 8
  **end for**
  $\boldsymbol{\pi}_{\text{old}} \leftarrow \boldsymbol{\pi}_\theta$
**end while**

range according to Eq. equation 9. The value critic and vector critic are then updated jointly with the policy.

# G Cooperative Adaptive Cruise Control Experiment

## G.1 Experiment Setup

The analysis is conducted in an open-platoon cooperative adaptive cruise control (CACC) system with local predecessor interactions. The system contains one external leader and $N$ controllable follower vehicles, where only the followers are treated as agents. All followers adopt a shared Gaussian policy class with state-independent variance. The objective of this experiment is to isolate the scaling behavior of the surrogate objectives under a fixed policy update, rather than evaluating asymptotic training performance.

Three predecessor-type tasks with reward-coupling radii $1$, $4$, and $16$ are considered. To isolate the effect of coupling range, all tasks use the same one-predecessor Gaussian policy mean, which depends on the local spacing error $e_i$, the speed error relative to cruise $z_i$, and the relative speed $\Delta v_i$. The tasks differ only in the stage reward. The $1$-hop task penalizes tracking errors with respect to the immediate predecessor. The $4$-hop and $16$-hop tasks add tracking penalties for predecessors at offsets $4$ and $16$, respectively. Increasing the predecessor depth enlarges only the reward-coupling radius; it does not turn the system into an all-to-all game as $N$ grows.

The base stage reward penalizes next-step spacing error, predecessor relative-speed mismatch, absolute speed error, and acceleration effort. The $4$-hop and $16$-hop variants incorporate corresponding longer-range tracking penalties while maintaining the same observation space and policy architecture. Therefore, in this task family, the coupling strength is determined by the environment's reward density rather than algorithmic architectural changes.

We define the independent-to-joint objective gap as

$$\Delta\mathcal{L} = |\mathcal{L}_{\text{joint}}(\boldsymbol{\pi}_{\text{new}}) - \sum_{i=1}^{N} \mathcal{L}^i_{\text{ind}}(\pi^i_{\text{new}})|. \tag{46}$$

The reported metric is the normalized objective gap $\Delta\mathcal{L}/N$. Given each sampled state, the surrogate objectives are evaluated exactly from the closed-form Gaussian expectation of the one-step linearized CACC model. For each topology and each platoon size $N \in \{5, 10, 20, 50, 100, 200\}$, 128 states are sampled from a Gaussian distribution over spacing errors, speed errors, relative speeds, leader speed error, and leader acceleration, and the resulting quantities are averaged across states.

### G.2 Objective Gap Scaling

Figure 7 compares the normalized objective gap with the first-order terms from two analyses. For the prior $(N-1)A_{\max}$ coefficient (Sun et al., 2022; Su & Lu, 2024), the plotted quantity is

$$P_{\text{global}} \triangleq \mathbb{E}_{s\sim d_{\boldsymbol{\pi}_{\text{old}}}}\left[\frac{1}{N}\sum_{i=1}^{N}(N-1)A^{\text{emp}}_{\max}(s)\cdot 2\,D_{\text{TV}}(\pi^i_{\text{new}}||\pi^i_{old})[s]\right], \tag{47}$$

where $A^{\text{emp}}_{\max}(s)$ is the largest sampled absolute advantage, not the global supremum required for a formal worst-case bound. The coupling-dependent first-order term is

$$P_{\text{coupling}} \triangleq \mathbb{E}_{s\sim d_{\boldsymbol{\pi}_{\text{old}}}}\left[\frac{1}{N}\sum_{i=1}^{N}\mathcal{C}_i(s)\sqrt{\chi^2_i(s)}\right], \tag{48}$$

which depends on the local interaction neighborhood. In Fig. 7, "CARPO Bound" plots $P_{\text{coupling}}$ and "Prior Bound" plots $P_{\text{global}}$. These curves compare first-order scaling under a fixed update; the complete bound in Theorem 3.1 also contains $R_N$. Since $2D_{\text{TV}} \le \sqrt{\chi^2}$, the two terms use comparable policy update scales.

As illustrated in Fig. 7, the normalized gap and coupling-dependent term remain nearly stable with platoon size, whereas $P_{\text{global}}$ increases. With the interaction radius fixed, adding vehicles extends the platoon without giving each vehicle proportionally more direct reward interactions. The experiment shows how the coupling term depends on interaction radius under a fixed policy update.

Table 2 gives representative values. With the shared policy fixed, enlarging the reward-coupling radius from 1 to 4 to 16 raises the level of both the gap and coupling-dependent term. Their scale follows the neighborhood size more closely than the total number of vehicles.

### G.3 Update Range Comparison

To compare the update ranges implied by the two coefficients, suppose the realized probability ratios satisfy $|\pi^i_{\text{new}}(a|s)/\pi^i_{\text{old}}(a|s) - 1| \le \epsilon$ for all $s, a$. Then

$$2\,D_{\text{TV}}(\pi^i_{\text{new}}||\pi^i_{\text{old}})[s] \le \sqrt{\chi^2_i(s)} \le \epsilon. \tag{49}$$

The plots use this common conditional ratio range to compare the coefficients of the first-order terms. PPO objective clipping alone does not impose the stated ratio constraint.

Set each coefficient times its corresponding update range to the same value $F > 0$.

For the prior analysis, the coefficient uses the sampled maximum advantage, yielding a single update range for the entire platoon:

$$\epsilon_{\text{global}} = \frac{F}{(N-1)\,A^{\text{emp}}_{\max}}. \tag{50}$$

For CARPO, the state-dependent coupling coefficient gives a range for each agent and state:

$$\epsilon^i_{\text{CARPO}}(s) = \frac{F}{\mathcal{C}_i(s)}. \tag{51}$$

To visualize this decentralized adaptation, the lower panels plot the distribution of these ranges across sampled states, summarized by their median and min-max envelope.

The $(N-1)A^{\text{emp}}_{\max}$ coefficient makes the corresponding range shrink with $N$, while the coupling coefficient gives a nearly stable range in these fixed-radius tasks.

| Hop radius | $N$ | True gap | Coupling term | Global proxy |
|---|---|---|---|---|
| 1 | 20 | 0.0091 | 0.0159 | 6.2814 |
| 1 | 100 | 0.0095 | 0.0169 | 31.0982 |
| 1 | 200 | 0.0097 | 0.0169 | 64.3915 |
| 4 | 20 | 0.0182 | 0.0293 | 7.9854 |
| 4 | 100 | 0.0198 | 0.0321 | 44.9209 |
| 4 | 200 | 0.0201 | 0.0323 | 91.5676 |
| 16 | 20 | 0.0245 | 0.0389 | 10.4391 |
| 16 | 100 | 0.0299 | 0.0467 | 58.1433 |
| 16 | 200 | 0.0306 | 0.0479 | 118.0273 |

Table 2: Representative normalized values from the 1-hop, 4-hop, and 16-hop CACC experiments. The coupling term omits the higher-order residual; the global proxy uses a sampled maximum advantage.

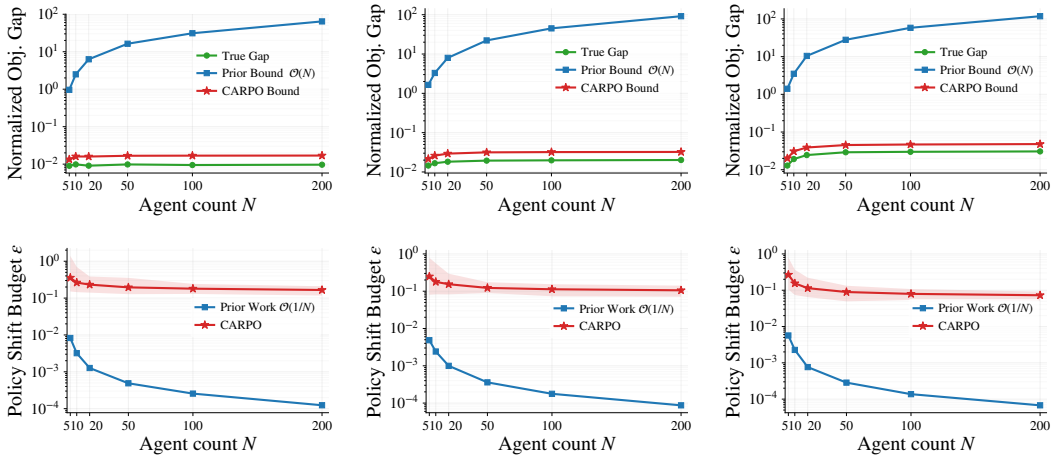



Figure 7: Scale-up behavior in predecessor-type CACC tasks with reward-coupling radii 1, 4, and 16. Top row: normalized objective gap, the prior first-order proxy ("Prior Bound") using a sampled maximum advantage, and the coupling-dependent first-order term ("CARPO Bound"). Bottom row: update ranges from the two coefficients at the same target value (labeled "Prior Work" and "CARPO"), with the state-dependent median and min–max envelope for coupling.

# H Implementation Details of Stochastic Climbing Game

## H.1 Algorithmic Implementation and Objective Function

In this tabular case study, we evaluate the convergence properties of CARPO using the Stochastic Climbing Game. To bridge the gap between our exact analytical bound (Theorem 3.1) and practical implementation, we eschew complex critic networks and optimize the coupling-adaptive objective directly.

Neglecting higher-order terms, the policy update for agent $i$ is performed by solving the following regularized optimization problem:

$$\pi^i_{k+1} = \arg\max_{\pi^i} \left( \mathbb{E}_{a^i \sim \pi^i, a^{-i} \sim \pi^{-i}_k} \left[ R(a^i, a^{-i}) \right] - \lambda \mathcal{C}_i(s) \sqrt{\chi^2 \left( \pi^i \| \pi^i_k \right) + \eta} \right), \tag{52}$$

where $\chi^2 \left( \pi^i \| \pi^i_k \right) = \sum_{a^i} (\pi^i(a^i) - \pi^i_k(a^i))^2 / \pi^i_k(a^i)$ is the Pearson $\chi^2$-divergence, and $\mathcal{C}_i(s)$ represents the local coupling coefficient.

To implement this update rule in a decentralized learning context, we compute the expectation over the agent's own policy $\pi^i$ analytically, while the impact of the opponent is estimated by sampling actions $a^{-i} \sim \pi^{-i}_k$. This sampling-based estimation simulates the inherent uncertainty and non-stationarity of

independent learning. Furthermore, to ensure numerical stability during gradient-based optimization, we introduce a small smoothing constant $\eta$ within the square-root term. This simplified framework allows us to isolate and verify the effect of the adaptive trust region without the confounding factors of neural function approximation. Because this smoothing artificially caps the penalty gradient near $\chi^2 = 0$, we introduce a multiplier $\lambda$ to restore the necessary repelling force of the trust region against extreme miscoordination penalties.

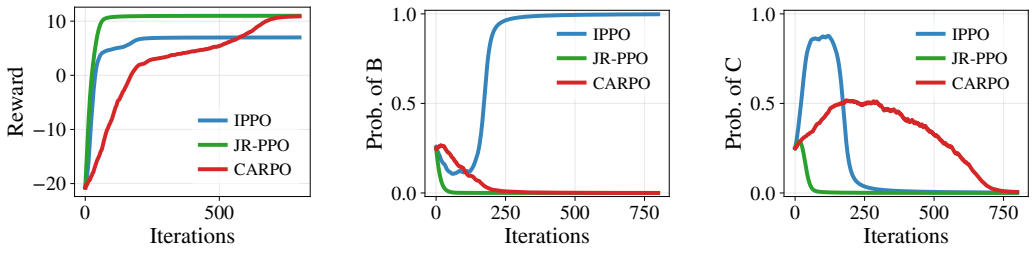
### H.2 Hyperparameter Configurations

The specific hyperparameter settings for the matrix game experiments are detailed in Table 3. We initialize both agents with symmetric action distributions to investigate the expansion of the basin of attraction towards the Pareto-optimal equilibrium. In our visualization of the learning dynamics, we specifically analyze the trajectory originating from initial conditions that typically lead to sub-optimal convergence in standard independent learning baselines.

Table 3: Experimental Hyperparameters for the Stochastic Climbing Game.

| **Hyperparameter** | **Value** |
|---|---|
| *Environment & Sampling* | |
| Action Estimation | Sampled from $\pi^{-i}$ |
| Opponent Sampling Batch Size | 1 |
| Seeds | 1–5 |
| *PPO Backbone* | |
| Learning Rate ($\alpha$) | 0.02 |
| Clip Coefficient ($\epsilon_{ppo}$) | 0.2 |
| Entropy Coefficient | 0.04 |
| Grad Norm Clipping | 0.8 |
| Update Epochs ($K$) | 2 |
| *CARPO Regularization* | |
| Smoothing Constant ($\eta$) | 0.0008 |
| Multiplier Constant ($\lambda$) | 8 |

### H.3 Extended Results




(a) Reward Dynamics (b) Policy Trajectory $P(B)$ (c) Policy Trajectory $P(C)$

Figure 8: **Extended Learning Dynamics in Stochastic Climbing Game.**

We provide an extended analysis of the learning dynamics and initialization sensitivity to supplement the main text.

Fig. 8 details the reward and policy trajectories. IPPO optimizes greedily, rapidly shifting probability mass to actions $B$ and $C$, which traps the agents in a suboptimal reward plateau ($\approx 7$). In contrast, CARPO demonstrates the "safe hesitation" phase: it preserves sufficient probability mass on $A$, allowing the agents to eventually discover and converge to the global optimum ($\approx 11$), matching the final performance of the JR-PPO oracle.

Fig. 9 and Fig. 10 quantify the basin of attraction for the global optimum. Assuming a symmetric initialization for the suboptimal actions ($P(B) = P(C)$), Fig. 9 tracks final rewards across varying initial probabilities of the optimal action $A$. IPPO requires a strong initial prior ($P(A) \geq 0.7$) to avoid the suboptimal trap. Both CARPO and JR-PPO significantly expand this basin, converging to the maximum reward from initializations as low as $P(A) \geq 0.45$.

Fig. 10 generalizes this evaluation across the entire symmetric policy simplex. The heatmaps confirm that IPPO fails across a wide range of initializations (the dark blue region in Fig. 10a). CARPO structurally mitigates this, converting a substantial portion of the simplex into a successful convergence zone (Fig. 10c). The relative difference maps (Fig. 10d-10f) indicate that CARPO consistently outperforms IPPO and maintains a near-zero performance gap compared to the JR-PPO baseline.

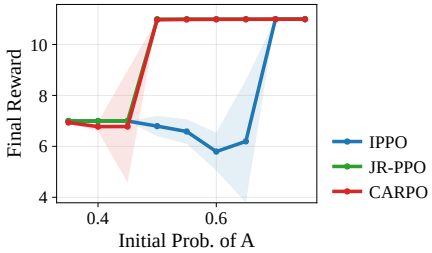
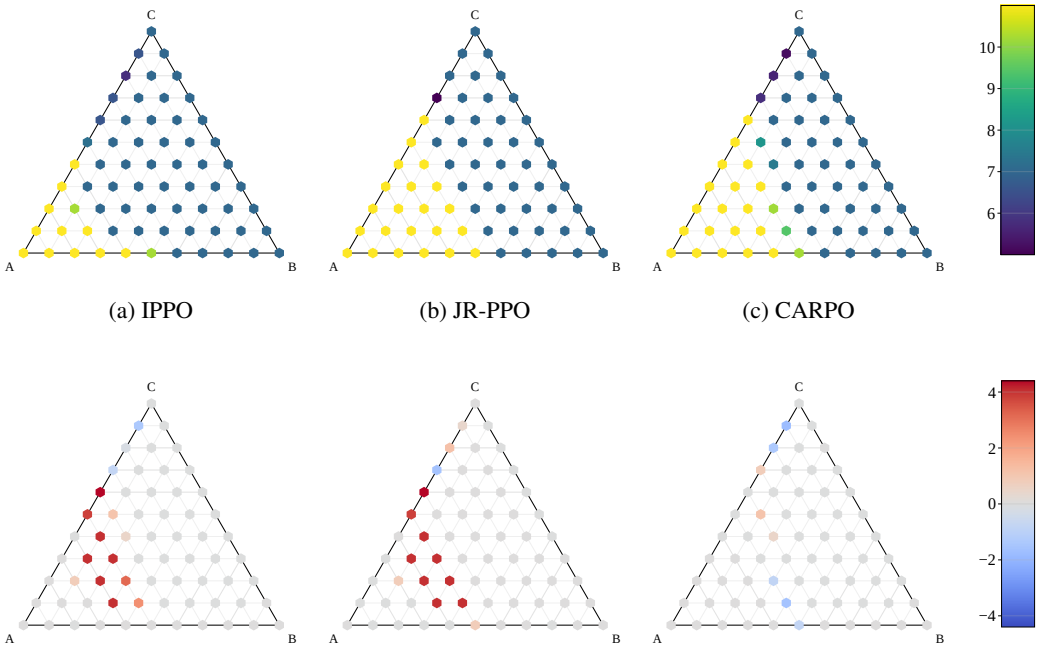



Figure 9: **Expanded Basin of Attraction under Symmetric Initialization ($P(B) = P(C)$).**




(a) IPPO (b) JR-PPO (c) CARPO

(d) CARPO vs. IPPO (e) JR-PPO vs. IPPO (f) CARPO vs. JR-PPO

Figure 10: **Performance Evaluation across the Policy Initialization Simplex.** Each point represents a symmetric initial policy distribution $(P(A), P(B), P(C))$. (a–c) Absolute final returns. IPPO is trapped in a suboptimal basin (dark blue regions) across a large portion of the simplex. In contrast, both JR-PPO and CARPO significantly expand the basin of attraction for the global optimum (yellow regions). (d–f) Relative performance differences calculated via direct subtraction. The red regions in (d) and (e) indicate the performance gain of CARPO and JR-PPO over IPPO. (f) shows the performance gap between CARPO and JR-PPO is near zero (white region) globally.

# I Implementation Details of Level-Based Foraging (LBF)

## I.1 Hyperparameter Configurations

Table 4 summarizes the shared environment and training setup based on EPyMARL[1], and Table 5 reports the algorithm-specific hyperparameters. Both MAPPO and CARPO use a parameter-shared recurrent actor with a centralized critic.

Table 4: Shared Environment and Training Setup for General-Sum LBF Experiments.

| **Hyperparameter** | **Value** |
|---|---|
| *Environment: General-Sum LBF* | |
| Environment | `lbforaging:Foraging-10x10-8p-4f-v3` |
| Grid / Agents / Foods | $10 \times 10$ / 8 agents / 4 foods |
| Environment Hierarchy | A1F1, A1F2, A1F3 |
| Food / Player Levels | food level = 1, 2, 3; player level = 1 |
| Game Type | General-sum (`common_reward=False`) |
| Episode Limit | 100 steps |
| *Training Backbone: EPyMARL PPO* | |
| Learning Rate | $3 \times 10^{-4}$ |
| PPO Update Epochs | 4 |
| GAE ($\lambda$) / Discount ($\gamma$) | 0.95 / 0.99 |
| Policy Network / Hidden Dim | GRU / 128 |
| Critic Network | Centralized MLP critic |
| Entropy Coefficient | 0.001 |
| Total Timesteps | $1 \times 10^6$ |
| Seeds | 1–10 |

Table 5: CARPO-Specific Hyperparameters in LBF.

| **Hyperparameter** | **Value** |
|---|---|
| Vector Critic Neighborhood Size $M$ | 7 (All) |
| Variance Sampling $K$ | 8 |
| Normalization Method | EMA (decay = 0.99) |
| Base Clipping Bound ($\epsilon_{\text{base}}$) | 0.20 |
| Adaptive Range $[\epsilon_{\min}, \epsilon_{\max}]$ | [0.1, 0.5] |
| Vector Critic Network Hidden Dim | 128 |

## I.2 Extended Results

We present the complete return performance of all static MAPPO variants and the proposed CARPO under different coordination levels in the LBF environment, as summarized in Table 6.

# J Implementation Details of PufferDrive

## J.1 Experimental Settings

The default configuration of the PufferDrive[2] environment provides a discrete action space for classic dynamics model that is highly permissive. Specifically, the original acceleration bounds range from $-4.0$ m/s$^2$ to $4.0$ m/s$^2$ (7 bins), and the steering angle ranges from $-1.0$ rad to $1.0$ rad (13 bins). An angle of $1.0$ rad ($\approx 57.3°$) per step allows for physically impossible hairpin turns at high speeds, while an acceleration of $4.0$ m/s$^2$ resembles sports-car kinematics. We observed that with these permissive bounds, agents tend to exploit unrealistic physics—such as instantly stopping or

[1] `https://github.com/uoe-agents/epymarl`

[2] `https://github.com/Emerge-Lab/PufferDrive`

Table 6: **Complete returns for all static MAPPO bounds and CARPO across varying coordination levels in LBF (mean ± std, 10 seeds)**. The best results per environment are in **bold**, and the best static MAPPO configuration per environment is highlighted with gray cells.

| **Method** | $\epsilon$ | **A1F1** (Low) | **A1F2** (Medium) | **A1F3** (High) |
|---|---|---|---|---|
| MAPPO | 0.1 | $1.000 \pm 0.000$ | $0.737 \pm 0.035$ | $0.167 \pm 0.028$ |
| | 0.2 | $1.000 \pm 0.000$ | $0.771 \pm 0.025$ | $0.209 \pm 0.041$ |
| | 0.3 | $1.000 \pm 0.000$ | $0.823 \pm 0.028$ | $0.249 \pm 0.058$ |
| | 0.4 | $1.000 \pm 0.000$ | $0.842 \pm 0.033$ | $0.294 \pm 0.041$ |
| | 0.5 | $1.000 \pm 0.000$ | $0.840 \pm 0.028$ | $0.318 \pm 0.047$ |
| | 0.6 | $1.000 \pm 0.000$ | $0.801 \pm 0.025$ | $0.223 \pm 0.032$ |
| | 0.7 | $1.000 \pm 0.000$ | $0.834 \pm 0.030$ | $0.274 \pm 0.050$ |
| | 0.8 | $1.000 \pm 0.000$ | $0.873 \pm 0.026$ | $0.163 \pm 0.029$ |
| | 0.9 | $1.000 \pm 0.000$ | $0.782 \pm 0.037$ | $0.231 \pm 0.047$ |
| | 1.0 | $1.000 \pm 0.000$ | $0.744 \pm 0.037$ | $0.204 \pm 0.042$ |
| CARPO (Ours) | | $1.000 \pm 0.000$ | $\mathbf{0.885} \pm 0.025$ | $\mathbf{0.337} \pm 0.052$ |

aggressively swerving to dodge collisions—thereby bypassing the need for genuine multi-agent coordination.

To enforce realistic driving behaviors and amplify the interactive challenge, we restricted the action space to reflect the kinematic limits of standard passenger vehicles:

1. **Acceleration:** The bounds are narrowed to $[-3.0, 2.0]$ m/s$^2$ and discretized into 9 bins:
$$\{-3.0, -2.0, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5, 2.0\}.$$
This asymmetric range captures the physical reality that standard vehicles possess stronger braking capabilities than forward acceleration.
2. **Steering:** The maximum steering angle is restricted to $[-0.5, 0.5]$ rad ($\approx \pm 28.6°$). We redistributed the 13 discrete bins as follows:
$$\{-0.50, -0.35, -0.20, -0.10, -0.05, -0.02, 0.00, 0.02, 0.05, 0.10, 0.20, 0.35, 0.50\}.$$
Crucially, we increased the action density around the center (e.g., $\pm 0.02, \pm 0.05$). This fine-grained granularity enables vehicles to make smooth micro-adjustments and maintain stable lane-keeping at higher speeds.

### J.2 Hyperparameter Configurations

Table 7 summarizes the environmental settings of PufferDrive and the common training hyperparameters shared across all methods. Table 8 lists the method-specific hyperparameters for CARPO and the baseline methods, including PPO, DPO, MARPO, and GS-CoPPO.

**Baseline Parameter Selection.** All PufferDrive baselines share the training backbone in Table 7. For each method, we evaluated its default update-control parameter and one smaller and one larger value with the same training budget. We selected the value with the highest mean success rate over the final $10\%$ of training on the mixed set of $10{,}000$ scenarios, then used that configuration in all four subsets without subset-specific retuning. Table 9 reports this matched-budget sweep; selected configurations were evaluated with $10$ seeds.

### J.3 Baseline

To ensure a fair evaluation in the PufferDrive, all other baselines (IPPO, DPO, MARPO, GS-CoPPO) utilize a shared actor-critic backbone conditioned on identical local egocentric observations. In contrast, MAPPO employs disjoint actor and critic networks to implement the CTDE paradigm. While MAPPO's actor uses the exact same local observation space as the other baselines, its centralized critic is granted an expanded spatial receptive field and privileged features.

Table 7: Shared Experimental Setup and PPO Backbone in PufferDrive.

| Hyperparameter | Value |
|---|---|
| *Environment: PufferDrive* | |
| Total Parallel Agents ($N$) | 93,200 ($5{,}825 \times 16$) |
| Simulation Step ($\Delta t$) | 0.1 s |
| Episode / Resample Horizon | 91 / 910 steps |
| *Training Backbone: PuffeRL* | |
| Learning Rate / Schedule | 0.003 / Cosine Annealing |
| Rollout Horizon / Batch Size | 90 / 8,388,000 agent-steps |
| Mini-batch Size / Update Epochs | 29,120 / 1 |
| GAE ($\lambda$) / Discount ($\gamma$) | 0.95 / 0.98 |
| Total Timesteps | $4 \times 10^8$ agent-steps |
| Seeds | 1–10 |

Table 8: Algorithm-Specific Hyperparameters in PufferDrive.

| Algorithm / Hyperparameter | Value |
|---|---|
| *IPPO/MAPPO* | |
| Clip Coefficient $\epsilon$ | 0.2 |
| *DPO* | |
| Target KL $d_{\text{target}}$ | 0.01 |
| Adaptive Factor $\eta_1$ | 1.5 |
| Adaptive Factor $\eta_2$ | 2.0 |
| *MARPO* | |
| $\alpha$ | 1.0 |
| $\beta$ | 0.05 |
| KL Bias | 0.05 |
| *GS-CoPPO* | |
| Clip Coefficient $\epsilon$ | 0.2 |
| Inner Clip Coefficient $\epsilon_{\text{in}}$ | 0.1 |
| *CARPO* | |
| Vector Critic Neighborhood Size $M$ | 4 |
| Variance Sampling $K$ | 4 |
| Normalization Method | EMA (decay = 0.99) |
| Base Clipping Bound ($\epsilon_{\text{base}}$) | 0.20 |
| Adaptive Range $[\epsilon_{\text{min}}, \epsilon_{\text{max}}]$ | [0.15, 0.25] |
| Vector Critic Network Hidden Dim | 128 |

Table 10 details the specific composition of these observation vectors. The fundamental differences lie in the neighbor perception: the local actor observation restricts visibility to a 50m radius with 7-dimensional kinematic features per neighbor, whereas the MAPPO critic expands this radius to 100m and appends 2-dimensional hidden navigation goals to each neighbor's feature vector.

Table 9: PufferDrive baseline parameter sweep. Entries give candidate value $\rightarrow$ final mean success rate (%); bold marks the selected value. Final evaluation variability is reported in Table 1.

| Method / parameter | Candidate values and success rates |
|---|---|
| IPPO / PPO clip | $0.15 \rightarrow 91.9$; $\mathbf{0.20 \rightarrow 92.1}$; $0.25 \rightarrow 91.1$ |
| MAPPO / PPO clip | $0.15 \rightarrow 90.5$; $\mathbf{0.20 \rightarrow 91.0}$; $0.25 \rightarrow 90.3$ |
| GS-CoPPO / inner clip | $0.05 \rightarrow 90.2$; $\mathbf{0.10 \rightarrow 93.5}$; $0.20 \rightarrow 82.5$ |
| DPO / target KL | $0.005 \rightarrow 55.6$; $\mathbf{0.01 \rightarrow 84.8}$; $0.02 \rightarrow 79.5$ |
| MARPO / KL bias | $0.025 \rightarrow 90.8$; $\mathbf{0.05 \rightarrow 91.2}$; $0.10 \rightarrow 91.0$ |

Table 10: Observation Space Composition. All evaluated algorithms share the base egocentric features. Only MAPPO's centralized critic utilizes the expanded privileged features.

| Observation Entity | Dimension | Feature Details |
|---|---|---|
| *Shared Base Features (Visible to all Actors and Independent Critics)* | | |
| Ego Vehicle | 8 | Relative goal position $(x, y)$, speed, width, length, collision state, respawn state, entity type. |
| Road Segments | $128 \times 7$ | Up to 128 segments. Features: relative $x, y$, length, width, heading $(\cos, \sin)$, entity type. |
| Local Neighbors | $31 \times 7$ | Up to 31 agents within a **50m radius**. Features: relative $x, y$, width, length, heading $(\cos, \sin)$, speed. |
| *Privileged Features (Visible ONLY to MAPPO Critic)* | | |
| Expanded Neighbors | $31 \times 9$ | Up to 31 agents within a **100m radius**. Features: local neighbor features + **hidden navigation goal** $(x, y)$. |

Let

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}$$

denote the importance ratio, and let

$$\hat{A}_t$$

be the advantage estimate.

**IPPO.** IPPO uses the standard PPO clipped surrogate objective:

$$\mathcal{L}^{\text{actor}}_{\text{IPPO}}(\theta) = \mathbb{E}_t\Big[\max\Big(-r_t(\theta)\hat{A}_t,\ -\text{clip}\big(r_t(\theta),\, 1-\epsilon,\, 1+\epsilon\big)\hat{A}_t\Big)\Big].$$

**DPO.** DPO replaces PPO clipping with an adaptive KL-regularized objective:

$$\mathcal{L}^{\text{actor}}_{\text{DPO}}(\theta) = -\mathbb{E}_t\Big[r_t(\theta)\hat{A}_t\Big] + \beta_t \bar{D}_{\text{KL}},$$

where $\bar{D}_{\text{KL}}$ is the empirical KL proxy and $\beta_t$ is adaptively updated to keep the KL near the target value $d_{\text{target}}$:

$$\beta_{t+1} \leftarrow \begin{cases} \beta_t \eta_2, & \text{if } \bar{D}_{\text{KL}} > \eta_1 d_{\text{target}}, \\ \beta_t / \eta_2, & \text{if } \bar{D}_{\text{KL}} < d_{\text{target}}/\eta_1, \\ \beta_t, & \text{otherwise.} \end{cases}$$

Thus, Target KL, $\eta_1$, and $\eta_2$ determine the target trust region and its adaptation strength.

**MARPO.** Its actor loss contains a standard clipped PPO term and an additional reflection term:

$$\mathcal{L}^{\text{actor}}_{\text{MARPO}}(\theta) = \mathcal{L}_0(\theta) + \alpha\, \mathcal{L}_1(\theta),$$

where

$$\mathcal{L}_0(\theta) = \mathbb{E}_t\Big[\max\Big(-r_t(\theta)\hat{A}_t,\ -\text{clip}\big(r_t(\theta),\, x_1,\, x_2\big)\hat{A}_t\Big)\Big],$$

and

$$\mathcal{L}_1(\theta) = \mathbb{E}_t\Big[\max\Big(-r_t(\theta) r_{t+1}(\theta)\hat{A}_t,\ -\text{clip}\big(r_t(\theta),\, x_1,\, x_2\big)\text{clip}\big(r_{t+1}(\theta),\, x_1,\, x_2\big)\hat{A}_t\Big)\Big].$$

Here, $x_1$ and $x_2$ are the lower and upper clipping bounds, obtained as the two solutions to

$$x - 1 - \log x = \delta_{\text{KL}},$$

with $x_1 \in (0, 1)$ and $x_2 \in (1, \infty)$. The target KL level $\delta_{\text{KL}}$ is updated online using an exponential moving average,

$$\delta_{\text{KL}} = \max\Big(D^{\text{bias}}_{\text{kl}},\ \beta\, \delta^{\text{prev}}_{\text{KL}} + (1-\beta)\, \widehat{D}_{\text{KL}}\Big),$$

where $\beta$ is the smoothing coefficient, $D^{\text{bias}}_{\text{kl}}$ is a lower bound on the target KL, and $\widehat{D}_{\text{KL}}$ is the current KL estimate.

**GS-CoPPO.** GS-CoPPO introduces a coordinated ratio that accounts for other agents in the same scene-time group. Let $r_t^{\text{ego}}(\theta)$ denote the ego-agent ratio and let $r_t^{\text{oth}}(\theta)$ denote the aggregated ratio of the other agents. CoPPO first applies inner clipping to the other-agent ratio,

$$\tilde{r}_t^{\text{oth}}(\theta) = \text{clip}\big(r_t^{\text{oth}}(\theta),\, 1-\epsilon_{\text{in}},\, 1+\epsilon_{\text{in}}\big),$$

then forms the coordinated ratio

$$r_t^{\text{co}}(\theta) = r_t^{\text{ego}}(\theta)\tilde{r}_t^{\text{oth}}(\theta),$$

and finally uses the PPO-style clipped objective

$$\mathcal{L}_{\text{CoPPO}}^{\text{actor}}(\theta) = \mathbb{E}_t\Big[\max\Big(-r_t^{\text{co}}(\theta)\hat{A}_t,\; -\text{clip}\big(r_t^{\text{co}}(\theta),\, 1-\epsilon,\, 1+\epsilon\big)\hat{A}_t\Big)\Big].$$

Hence, $\epsilon$ controls the outer clipping range, while $\epsilon_{\text{in}}$ controls the inner clipping of the other-agent joint ratio.

### J.4 Scenario Partitioning

We partition the training scenarios into four subsets based on traffic density and intersection complexity. To ensure that the partition reflects the interactions seen by the policy at initialization, we compute both metrics on runtime-active controllable vehicles. We measure complexity by the heading entropy of active vehicles, which increases when more motion directions are present. We measure density by the maximum number of active vehicles within a 15-meter neighborhood. We then split the scenario set using fixed thresholds on the two metrics (Fig. 11), producing four subsets: low-density low-complexity, high-density low-complexity, low-density high-complexity, and high-density high-complexity. Representative scenarios for each partition are visualized in Fig. 12. This procedure yields 4962, 2290, 1311, and 1094 scenarios in the four subsets, respectively.

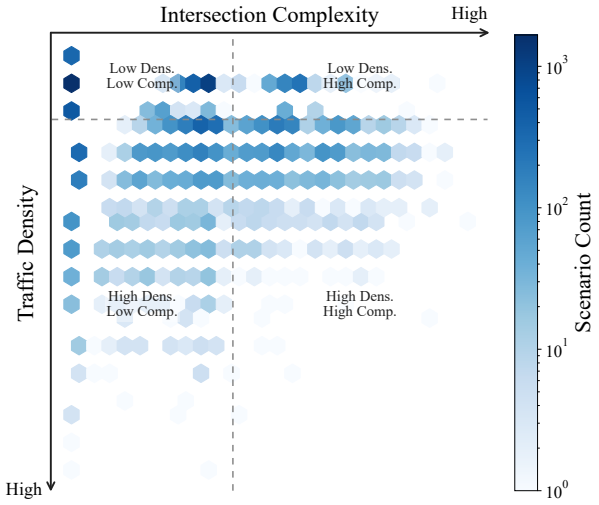



Figure 11: **Scenario partitioning by traffic density and intersection complexity.** Each hexagon shows the number of scenarios with the same density-complexity pattern. We use heading entropy as the complexity metric and the maximum number of active vehicles within a 15-meter neighborhood as the density metric. The two thresholds divide the full scenario set into four subsets.

### J.5 Extended Ablation Studies

In this section, we provide extended ablation studies on the variance sampling size $K$ and the coupling normalization strategy. Table. 12 summarizes the final success rates across all evaluated configurations.

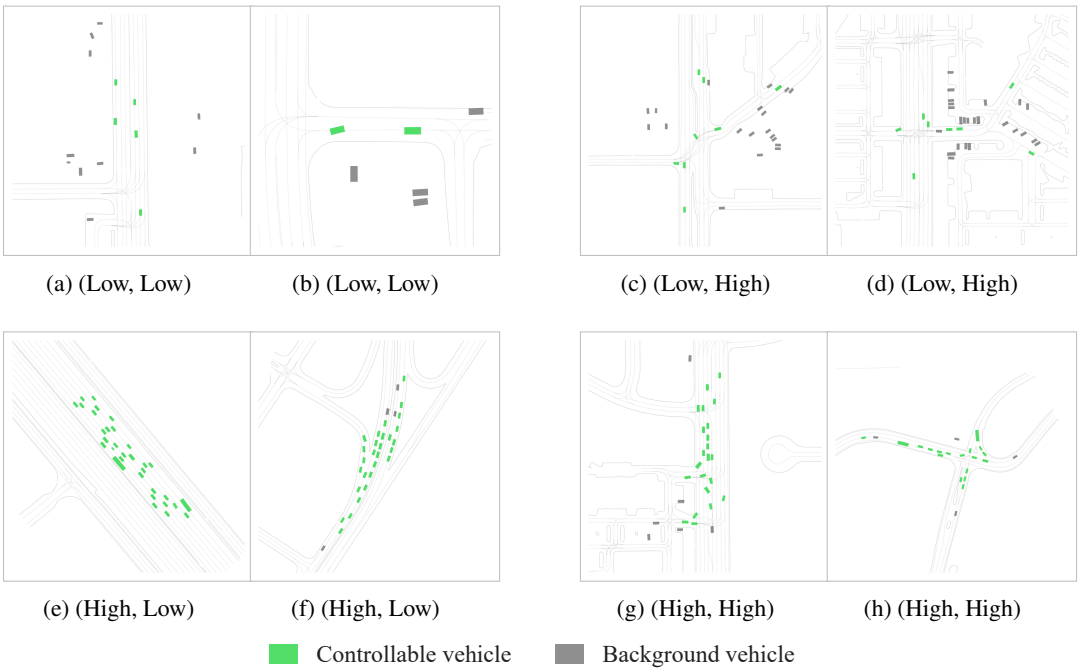
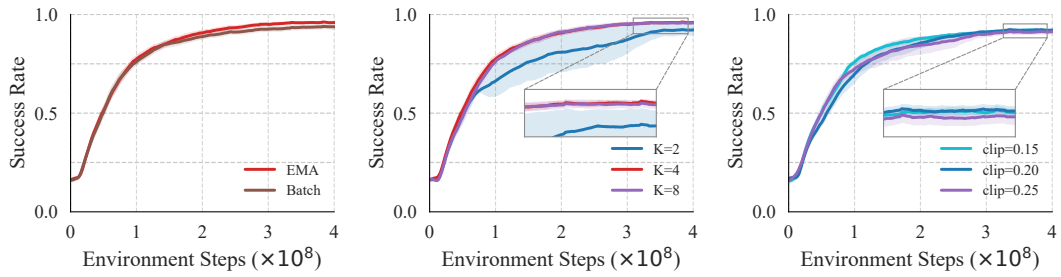



Figure 12: **Representative scenarios from the four partition subsets.** The tuple labels below each subfigure indicate (Traffic Density, Intersection Complexity). The low-density low-complexity subset mainly contains sparse scenes with simple motion patterns. The high-density low-complexity subset contains crowded but structurally simple interactions. The low-density high-complexity subset contains fewer vehicles but more diverse motion directions. The high-density high-complexity subset combines dense traffic with complex multi-direction interactions.




(a) Coupling Normalization (b) Variance Sampling Size ($K$) (c) IPPO Trust Region ($\epsilon$)

Figure 13: **Extended Ablation Studies.**

**Coupling Normalization Strategy.** We investigate the Exponential Moving Average (EMA) normalization used in CARPO. The adaptation in Eq. equation 9 compares instantaneous coupling with a historical baseline. As shown in Fig. 13a, the EMA baseline yields a higher final success rate than batch-level normalization. It also prevents the clipping range from depending entirely on a single batch mean; common multiplicative estimation bias cancels in the ratio.

**Variance Sampling Size ($K$).** We evaluate the number of action samples $K$ used to approximate the marginalized advantage variance for the coupling coefficient. As shown in Fig. 13b, setting $K = 2$ results in a high-variance learning curve and degraded asymptotic performance, indicating that too few samples fail to provide a reliable estimate of the local coupling strength. Increasing the sample size to K=4 stabilizes the estimation and maximizes the success rate. Further increasing the sample size to $K = 8$ results in nearly identical performance to $K = 4$, indicating that the approximation quality saturates beyond four samples.

**IPPO Trust Region ($\epsilon$).** The mixed-scenario sweep in Fig. 13c compares fixed IPPO clipping values $\epsilon \in \{0.15, 0.20, 0.25\}$, with 0.20 giving the highest mean success rate. CARPO varies its clipping range within $[0.15, 0.25]$ and exceeds the best fixed IPPO configuration in this sweep. We also evaluate five fixed values in each low-density subset (Table 11); CARPO exceeds every value tested in both subsets. These comparisons support adapting the clipping range to the state instead of selecting one fixed value from the tested ranges.

Table 11: IPPO clipping sweep in two PufferDrive subsets (success rate in %, 10 seeds, mean $\pm$ std). Each row compares the same IPPO method at five fixed $\epsilon$ values with CARPO trained in that subset.

| **Density, complexity** | $\epsilon = 0.10$ | 0.15 | 0.20 | 0.25 | 0.30 | **CARPO** |
|---|---|---|---|---|---|---|
| Low, low | $90.7 \pm 0.7$ | $89.3 \pm 0.6$ | $90.6 \pm 1.2$ | $91.7 \pm 0.8$ | $91.6 \pm 0.6$ | $\mathbf{95.8 \pm 0.7}$ |
| Low, high | $78.2 \pm 1.5$ | $78.6 \pm 3.2$ | $78.4 \pm 2.6$ | $77.3 \pm 2.2$ | $77.4 \pm 1.3$ | $\mathbf{87.1 \pm 1.5}$ |

Table 12: Ablation Study Results

| **Ablation** | **Performance** |
|---|---|
| *Reverse CARPO* | |
| Reverse | $0.845_{\pm 0.062}$ |
| IPPO | $0.921_{\pm 0.007}$ |
| CARPO (Ours) | $\mathbf{0.959}_{\pm 0.003}$ |
| *Vector Critic Neighborhood Size $M$* | |
| $M = 2$ | $0.941_{\pm 0.009}$ |
| $M = 4$ | $\mathbf{0.959}_{\pm 0.003}$ |
| $M = 8$ | $0.945_{\pm 0.013}$ |
| *Coupling Normalization Strategy* | |
| Batch | $0.937_{\pm 0.010}$ |
| EMA | $\mathbf{0.959}_{\pm 0.003}$ |
| *Variance Sampling Size $K$* | |
| $K = 2$ | $0.921_{\pm 0.026}$ |
| $K = 4$ | $\mathbf{0.959}_{\pm 0.003}$ |
| $K = 8$ | $0.957_{\pm 0.006}$ |
| *IPPO Trust Region $\epsilon$* | |
| $\epsilon = 0.15$ | $0.919_{\pm 0.005}$ |
| $\epsilon = 0.20$ | $\mathbf{0.921}_{\pm 0.010}$ |
| $\epsilon = 0.25$ | $0.911_{\pm 0.011}$ |

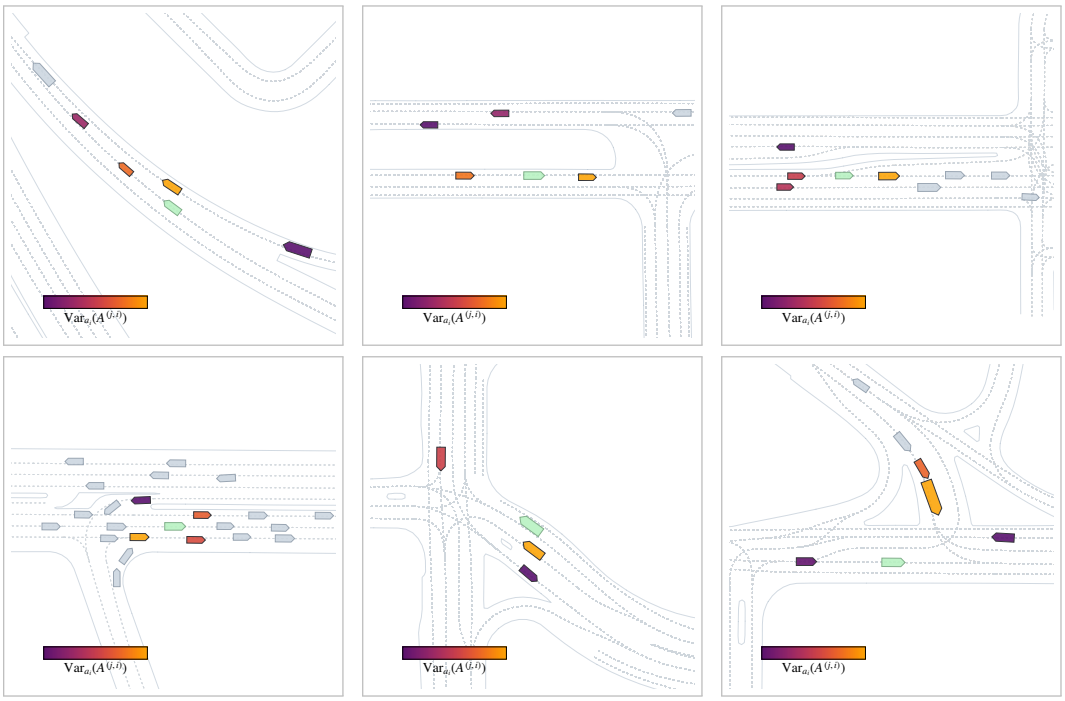


Figure 14: **Extended Visualizations of Ego-centric Variance.** We visualize the ego-centric variance $\mathrm{Var}_{a_i}(A^{(j,i)})$ across diverse traffic topologies. The green vehicle denotes the ego agent associated with action $a_i$.

### J.6 Visualizing Ego-centric Variance and Global Coupling Map

To supplement the mechanism analysis, Fig. 14 and Fig. 15 provide an extended gallery of visualizations across diverse traffic topologies.

Fig. 14 illustrates the marginalized advantage variance $\mathrm{Var}_{a_i}(A^{(j,i)})$ from the ego vehicle's perspective, showing how an individual vehicle's policy uncertainty impacts conflicting neighbors. Across all distinct geometries, the variance impact remains sparse and localized, impacting immediately conflicting neighbors.

Correspondingly, Fig. 15 displays the aggregated global coupling maps. We can see how localized impacts accumulate into major bottlenecks. The estimated coupling $\tilde{\mathcal{C}}_i(s)$ reaches its highest values in dense interaction areas, showing that CARPO safely takes smaller update steps wherever coordination is challenging.

### J.7 Hardware and Computational Cost

The experiments were conducted on a server equipped with dual Intel Xeon Silver 4210 CPUs (20 cores, 40 threads in total), 125 GB of RAM, and a single NVIDIA A100-PCIE-40GB GPU. Table 13 reports the total training time for each evaluated algorithm under a fixed budget of $417,231,360$ ($\approx 4.17 \times 10^8$) environment samples.

Table 13: Training time comparison across algorithms under a fixed budget of $4.17 \times 10^8$ samples.

| **Algorithm** | IPPO | MAPPO | DPO | MARPO | GS-CoPPO | CARPO |
|---|---|---|---|---|---|---|
| **Time (Minutes)** | 36.9 | 60.5 | 38.6 | 45.9 | 39.2 | 58.0 |

To compare training cost at a common performance level, we use the first logged checkpoint where the mean mixed-scenario success rate in Fig. 4a reaches $90\%$. IPPO, MAPPO, and CARPO reach




Figure 15: **Extended Visualizations of Global Coupling Maps.** We visualize the estimated coupling $\tilde{C}_i(s)$ across diverse traffic topologies.

this level after $2.68 \times 10^8$, $2.55 \times 10^8$, and $1.90 \times 10^8$ environment samples, respectively. Scaling each method's full-run time in Table 13 by the fraction of samples used gives estimated times of about 23.7, 36.9, and 26.4 minutes in the same order. CARPO uses fewer samples than both baselines and reaches the target earlier than MAPPO, while IPPO reaches it 2.7 minutes sooner. At the full budget, CARPO attains a $95.9\%$ success rate compared with $92.1\%$ for IPPO and $91.0\%$ for MAPPO (Table 1).