ReCord: Replay Coordination for Safe and Robust Population-Based Training in Autonomous Driving

Hyeon-Chang Jeon [1] **Kyung-Joong Kim** [1]

Abstract

Autonomous driving policies trained with self-play reinforcement learning (RL) can generalize to unseen scenarios, but they are trained primarily through interactions among copies of the same policy. As a result, they may fail to prepare for diverse and unfamiliar partner behaviors, which is safety-critical in autonomous driving, where other agents can be aggressive, non-reactive, or otherwise different from those seen during training. Population-based training (PBT) addresses this limitation by training the ego policy with diverse pre-trained partners. However, conventional PBT typically executes partner policies online during ego training, making them reactive to the ego policy. We refer to this standard setting as reactive-PBT. To address this limitation, we propose Replay Coordination (ReCord), which trains the ego policy on fixed trajectories replayed from a diverse partner population. By removing online partner adaptation, ReCord encourages robust coordination without relying on partners' yielding behavior. In both a matrix game and a multi-agent driving simulator, ReCord outperforms reactive-PBT, especially against non-reactive or weakly reactive partners, including replayed human trajectories, while remaining competitive under reactive evaluation.

1. Introduction

With the availability of large-scale real-world datasets (Ettinger et al., 2021; Caesar et al., 2020; Chang et al., 2019; Wilson et al., 2021) and fast simulation environments (Gulino et al., 2023; Kazemkhani et al., 2025; Cornelisse et al., 2025a), reinforcement learning (RL) based self-play has made significant progress in multi-agent autonomous driving (Cornelisse & Vinitsky, 2025; Cornelisse et al., 2025b; Cusumano-Towner et al., 2025; Wang et al., 2026; Chang et al., 2026). Self-play based approaches have demonstrated strong performance, often achieving nearly 100% success rates with very low collision rates in both training and test scenarios (Cornelisse et al., 2025b). However, these results primarily demonstrate generalization to unseen scenarios, whereas safe deployment in real-world traffic also requires autonomous vehicles to coordinate reliably with previously unseen partners.

This gap arises because self-play (SP) typically uses the same policy for all agents in the environment (Carroll et al., 2019; Bard et al., 2020). While SP can produce highly optimized driving policies, it exposes the ego policy to a limited distribution of partner behaviors during training. As a result, the learned policy can be brittle when interacting with partners outside the self-play distribution, including human drivers, agents trained with different RL algorithms, or even policies trained with different random seeds under the same algorithm.

This brittleness is particularly important in autonomous driving, where rare and extreme driving styles can create safety-critical coordination failures. We refer to such unseen or atypical partners as *long-tail* agents. In particular, we use the term from two complementary perspectives. First, long-tail agents include policies that differ from those encountered during training, thereby creating a zero-shot generalization challenge for the ego policy. Second, they include agents whose driving styles deviate from nominal behavior in coordination-critical ways, such as aggressive high-speed driving, frequent lane changes, or weak responses to other agents.

A promising approach is population-based training (PBT), which exposes the ego policy to diverse partner policies during training (Lanctot et al., 2017; Jaderberg et al., 2017; Carroll et al., 2019). In a common two-stage setup (Strouse et al., 2021; Zhao et al., 2023; Lou et al., 2023; Loo et al., 2023; Sarkar et al., 2023), a diverse partner population is first constructed and then kept fixed during ego-policy learning. However, although the partner pool is fixed, sampled partners are still executed online in simulation and remain re-

[1]Department of AI Convergence, Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea.. Correspondence to: Kyung-Joong Kim <kjkim@gist.ac.kr>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

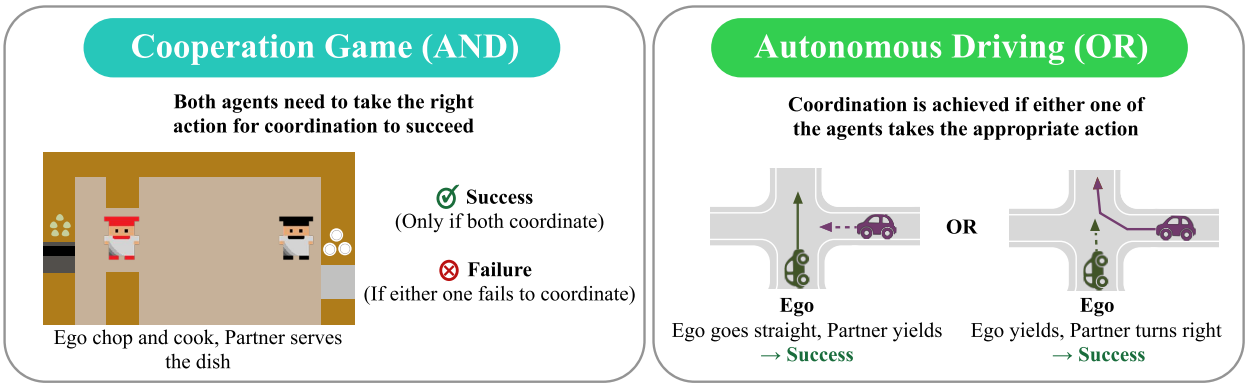

Figure 1. **AND vs. OR coordination.** Cooperative games typically require AND-style coordination, where all agents must execute their roles correctly. In autonomous driving, safe coordination often has an OR-style structure: either the ego vehicle or the partner vehicle can prevent failure by yielding or avoiding the conflict.

active to the observed state. We refer to this standard setting as *reactive-PBT*. While reactive-PBT improves robustness by broadening partner diversity, it has two limitations. First, as illustrated in Figure 1, autonomous driving differs from conventional cooperative games: safe coordination often requires only one agent to take the appropriate yielding or avoidance action. Therefore, reactive partners can resolve conflicts by yielding, reducing the ego policy's opportunity to learn conservative responses on its own (see Section 3). Second, executing partner policies online at every timestep is computationally expensive.

To mitigate these issues, we propose Replay Coordination (ReCord), a simple framework that modifies the use of partner populations during ego-policy learning. Instead of executing partner policies online as in reactive-PBT, ReCord first pre-records trajectories from a diverse partner population and replays them during training. This directly addresses the first limitation: because replayed partners cannot adapt to the ego policy, they cannot resolve conflicts by yielding online, forcing the ego policy to learn adaptive responses on its own. At the same time, ReCord preserves the main benefit of PBT by exposing the ego policy to diverse partner behaviors. It also addresses the second limitation by replacing repeated partner-policy inference with lightweight trajectory replay, making training substantially more efficient.

To validate these advantages, we conduct experiments in both a simple matrix game and a real-world driving simulation. The matrix game provides a controlled setting for isolating the effect of partner reactiveness during training. The driving simulation evaluates zero-shot coordination against diverse styles of unseen partners under both reactive and non-reactive interaction settings. These partners include policies trained with different random seeds, policies trained under different reward functions, and replayed human trajectories. Beyond comparing ReCord with reactive-PBT, we further ask whether the benefits of ReCord are specific to a particular partner selection strategy or hold more broadly. To answer this, we evaluate ReCord across multiple partner selection methods. Finally, we vary both the training population and the evaluation interaction setting to understand when ReCord is most effective and when reactive interaction remains useful.

Our results show that ReCord is most effective when the evaluation partners become more unfamiliar and less reactive. In particular, ReCord achieves lower collision rates and higher success scores than reactive-PBT in long-tail settings with non-reactive or weakly reactive partners, including replayed human trajectories. At the same time, ReCord remains competitive when evaluated with reactive unseen partners from a similar distribution, suggesting that removing partner adaptiveness during training does not substantially sacrifice performance in standard zero-shot coordination settings.

We further find that the effectiveness of ReCord depends strongly on the training population. ReCord benefits most from diverse populations that include behaviors deviating from nominal driving, indicating that population diversity and long-tail coverage are key factors for robust recorded-play training. Overall, while PBT-based methods substantially outperform self-play, ReCord provides the strongest benefits in non-reactive and long-tail partner settings. Our contributions are as follows:

- **ReCord improves robustness to long-tail partners.** We show that ReCord ReCord provides the largest gains in non-reactive and behaviorally mismatched settings, while remaining competitive with reactive-PBT under nominal reactive evaluation. This leads to lower collision rates and higher success rates, suggesting that removing partner adaptiveness encourages safer ego policies that do not rely on favorable partner reactions.

- **Population diversity is key to effective ReCord.** By comparing ReCord and reactive-PBT across different training populations, we find that population diversity is a key factor behind ReCord's gains. Diverse partners create richer coordination challenges during training, improving safety and robustness under unseen partner behaviors.
- **ReCord is computationally efficient.** ReCord avoids online partner-policy inference by replaying fixed trajectories, resulting in nearly $3\times$ faster simulation throughput than reactive-PBT.

2. Related Work

2.1. Reinforcement Learning in Autonomous Driving

RL-based approaches in autonomous driving have largely focused on improving generalization through self-play, starting with unseen scenarios within the same benchmark and gradually extending to broader zero-shot settings, such as unseen cities or other benchmarks. Previous studies (Vinitsky et al., 2022; Kazemkhani et al., 2025; Cornelisse et al., 2025b) showed that RL policies can generalize to unseen scenarios in a Waymo Open Motion Dataset (WOMD): for instance, Nocturne reported an 80% success rate in unseen scenarios (Vinitsky et al., 2022), and GPUDrive (Kazemkhani et al., 2025) demonstrated a nearly 100% success rate with collision rates below 1% using variants of PPO (Cornelisse et al., 2025b) with carefully designed reward combinations.

More recent studies have pushed this line of work further, beyond scenario-level generalization. NOMAD (Wang et al., 2026) showed that an RL agent can drive in entirely unseen cities without human trajectories by leveraging a scenario generator for new urban environments. GIGAFLOW (Cusumano-Towner et al., 2025) extended zero-shot evaluation even further, achieving strong performance across multiple benchmarks, including CARLA (Dosovitskiy et al., 2017), nuPlan (Caesar et al., 2021), and WOMD (Ettinger et al., 2021), through large-scale training and augmented map information. **Despite this progress, existing RL-based approaches primarily evaluate generalization over maps, scenarios, or environments under self-play, while zero-shot coordination with unseen partners remains relatively underexplored in autonomous driving.**

2.2. Safe and Robust Driving

Many data-driven behavior cloning methods have successfully learned human driving behaviors and achieved strong performance in nominal scenarios (Philion et al., 2024; Wu et al., 2024; Zhang et al., 2025). However, long-tail scenarios remain a major open challenge (Xu et al., 2025; Wang et al., 2025b), as they occur rarely in real-world data yet are critical for safety. Although collecting such cases from the real world would be the most direct solution, it is difficult in practice because these events are rare, costly to capture, and inherently risky. Simulation provides a promising alternative, as it avoids these limitations while enabling scalable data generation. In particular, unlike rule-based agents and planners, which often rely on predefined state transitions or handcrafted decision rules and struggle to generalize to unknown situations and uncertainty (Schwarting et al., 2018; Bouchard et al., 2022), RL agents can learn complex interactive behaviors well suited to challenging multi-agent traffic scenarios (Cusumano-Towner et al., 2025; Cornelisse et al., 2025b). **In this paper, we target to learn the long-tail scenarios, especially focused on coordinating with long-tail partners.**

2.3. Learning with Pre-trained and Non-reactive Partners

A common strategy in population-based multi-agent learning is to train an ego agent with a pool of pre-trained partner policies. Using pre-trained population approaches (Strouse et al., 2021; Zhao et al., 2023; Lou et al., 2023) improves generalization by sampling partners from diverse policy populations. However, these partners are still executed online during training. Thus, although the partner pool is frozen, the interaction remains reactive at the behavioral level.

Non-reactive partners have also been widely used in autonomous driving, especially in log-replay evaluation. In log-replay or closed-loop simulation, background vehicles follow recorded human trajectories and do not respond to the ego vehicle's counterfactual actions (Gulino et al., 2023; Caesar et al., 2021). This protocol is useful for evaluating policies against realistic human behavior, but it is usually treated as an evaluation setting rather than a training mechanism. **In contrast, ReCord uses non-reactive partner behavior during ego-policy training. Rather than learning to rely on partners that may yield or adapt online, the ego policy is encouraged to adapt to non-reactive partners that continue their recorded behaviors.**

3. Simplified Matrix Game

To validate ReCord in a controlled and interpretable setting, we introduce a modified matrix game, similar to those in (Elhenawy et al., 2015; Wang et al., 2021), as shown in Table 1. This game captures the core coordination dilemma in driving: while safe coordination is possible when at least one agent yields, failures occur when both agents choose to go. We additionally assume that all rational agents prefer to avoid collisions; therefore, even aggressive populations may still yield in reactive settings when they infer that continuing to go would cause a crash. This allows us to isolate the effects of reactive and non-reactive partner interactions on

		Partner	
		Yield	**Go**
Ego	**Yield**	0	0
	Go	1	2

(a) Safe

		Partner	
		Yield	**Go**
Ego	**Yield**	-0.2	0
	Go	1	-1

(b) Risky

Table 1. **Simplified matrix game:** There are two states in this matrix game: (a) Safe and (b) Risky. In (a), since the situation is safe, the optimum point is (Go, Go). However, in (b), if one of the agents does not yield, the collision occurs.

the ego policy's learned yielding behavior.

3.1. Population Setup

We model opponent populations with three latent types $t \in \{0, 1, 2\}$ and two states $s \in \{\texttt{safe}, \texttt{risky}\}$. Type and state are sampled independently:

$$\Pr(t) = [0.30,\ 0.40,\ 0.30], \qquad \Pr(s) = [0.55,\ 0.45].$$

For each (s, t) pair, the opponent has a base probability $p_{\text{base}}(s, t)$ of choosing $\texttt{Go}$ ($a_{\text{other}} = 1$):

$$p_{\text{base}}(\texttt{safe}, t) = [0.80,\ 0.92,\ 0.98],$$

Given ego go-probability p_{ego}, the reactive opponent uses

$$p_{\text{go}}^{\text{reactive}}(s, i) = \text{clip}\big(p_{\text{go}}^{\text{base}}(s, i) - \alpha\, p_{\text{ego}},\ 0.01,\ 0.99\big)\,,$$

where α controls how strongly the opponent reacts to the ego policy. Thus, α also determines the behavioral gap between the reactive and replay settings: a larger α makes the reactive opponent more adaptive to the ego policy, while the replay opponent remains fixed, resulting in a larger discrepancy between the two settings.

4. ReCord

In this section, we introduce ReCord, a simple and effective method for robust and safe driving. Unlike reactive-PBT, which executes sampled partner policies online during ego-policy training, ReCord first records partner behaviors in advance and then replays them during training. Formally, let $\mathcal{P} = \pi_1, \ldots, \pi_K$ denote a pre-trained population of K partner policies. Consider an episode with one ego agent and M partner agents. In reactive-PBT, each partner agent is assigned a policy sampled independently from the population, with replacement:

$$\pi^m \sim \text{Uniform}(\mathcal{P}), \quad m = 1, \ldots, M.$$

At each timestep t, the ego policy produces its action from its observation,

$$a_t^e \sim \pi_\theta(\cdot \mid o_t^e),$$

while each partner agent m executes its assigned policy online in the simulator:

$$a_t^m \sim \pi^m(\cdot \mid o_t^m), \quad m = 1, \ldots, M.$$

Equivalently, the joint partner action is

$$\mathbf{a}_t^{-e} = (a_t^1, \ldots, a_t^M),$$

where each a_t^m is generated online from the current observation. Since the partner observations o_t^m depend on the evolving simulator state and the ego policy's behavior, the sampled partners remain reactive throughout training.

In contrast, ReCord samples and replays pre-recorded partner trajectories instead of executing partner policies online. To construct the pre-recorded trajectories, we perform R population rollouts. In each rollout r, for all N agents in the scene, we independently sample a policy from the population with replacement:

$$\pi^{(r,i)} \sim \text{Uniform}(\mathcal{P}), \quad i = 1, \ldots, N.$$

We then execute these sampled policies online in the simulator and store the resulting joint action sequence. Each pre-recorded trajectory is defined as

$$\tau^r = (\mathbf{a}_0^r, \mathbf{a}_1^r, \ldots, \mathbf{a}_T^r)\,,$$

where T is the episode length and

$$\mathbf{a}_t^r = \left(a_t^{(r,1)}, a_t^{(r,2)}, \ldots, a_t^{(r,N)}\right)$$

denotes the joint action of all N agents at timestep t. For each rollout r, the corresponding pre-recorded trajectory τ_r and set of pre-recorded trajectories τ are defined as

$$\tau_r = (\mathbf{a}_0^r, \mathbf{a}_1^r, \ldots, \mathbf{a}_T^r)\,, \tau = \quad \{\tau_1, ..., \tau_r\}$$

During ego-policy training, ReCord samples one recorded trajectory $\tau_r \sim \tau$ from the trajectory set at the beginning of each episode. Unlike reactive-PBT, where partner actions are generated online from partner observations, ReCord directly replays partner actions from the sampled trajectory.

5. Experiments

We evaluate self-play, reactive-PBT, and ReCord under multiple population settings. In the matrix game, we vary the partner population's probability of taking risky actions. In PufferDrive, we construct four populations from nominal, lane-breaker, fast, and slow policies: L+N as the default nominal/lane-breaker mixture, ON as nominal-only, OL as lane-breaker-only, and Mixed as the combination of all four policy types. Population details are provided in Appendix A.

5.1. Matrix Game

We start with the simplified matrix game in Section 3, using two different PBT settings: reactive-PBT and ReCord. For training, we used the REINFORCE algorithm and trained it for 4K steps for 16 seeds.

Evaluation Metrics. Our evaluation aims to understand the difference in learning outcomes between reactive-PBT and ReCord. To this end, we measure three metrics. First, we define a **collision** as the case in which both agents choose *go* in the risky state. Second, we measure the **average return in the risky state** obtained by the ego policy. Lastly, we evaluate the **probability of choosing *go*** in each state in order to characterize how the learned policy behaves under different traffic conditions.

After training, we evaluate each policy against both reactive (reactive evaluation) and non-reactive (non-reactive evaluation) opponents. In addition, to examine how performance changes with the opponent's sensitivity to the ego policy, we vary α and report the corresponding performance across evaluation settings.

5.2. PufferDrive

PufferDrive (Cornelisse et al., 2025a) is a real-world data-driven multi-agent driving simulator built on PufferLib (Suarez, 2025). It is based on the WOMD, which contains 400K real-world driving scenes with diverse objects, including vehicles, pedestrians, and road elements. Each scene provides a 9-second trajectory, corresponding to 91 simulation steps with a control interval of 0.1 seconds.

The simulator achieves up to 300K simulation steps per second (SPS) and supports up to 128 controllable agents; in our experiments, we limit this number to 32 vehicles. Importantly, PufferDrive also supports human-compatible evaluation through features such as log-replay for zero-shot coordination with human trajectories.

Environment Setting. Each episode lasts 910 steps; agents reset to their initial positions upon reaching the goal, maps are resampled every 910 steps, and collisions are non-terminating (Kazemkhani et al., 2025; Cornelisse et al., 2025b). Observations are partially observable and ego-centric, containing ego-state, nearby vehicles (up to 32), and map features, and actions are chosen from 91 discrete steering–acceleration combinations.

Following (Cornelisse et al., 2025b), we use

$$r_t^{nominal} = \beta_1 \mathbf{1}\{c_t\} + \beta_2 \mathbf{1}\{o_t\} + \beta_3 \mathbf{1}\{g_t\}, \quad (1)$$

where c_t, o_t, and g_t denote collision, off-road, and goal-reaching, respectively, with $\beta_1 = -0.75$, $\beta_2 = -0.75$, and $\beta_3 = 1.0$. To construct long-tail populations, we augment the reward as

$$r_t = r_t^{nominal} + \beta_4 s_t + \beta_5 l_t, \quad (2)$$

$$s_t = \begin{cases} \mathbf{1}\{v_t \geq v_{\text{thres}}\}, & \text{fast} \\ \mathbf{1}\{v_t \leq v_{\text{thres}}\}, & \text{slow} \end{cases} \qquad l_t = \frac{\Delta\theta_t}{\pi} + \frac{d_t}{4}. \quad (3)$$

We set $v_{\text{thres}} = 40\,\text{m/s}$ for fast mode and $12\,\text{m/s}$ for slow mode, where $\Delta\theta_t$ is the heading difference to the closest lane direction and d_t is the distance to the closest lane segment.

Evaluation Settings. To evaluate zero-shot coordination, we consider four partner settings that vary both behavioral distribution shift and partner reactivity:

- **Unseen Seed (US):** Partners are drawn from the same policy family but trained with different random seeds.
- **Unseen Reward (UR):** Partners are trained with different reward functions, inducing behavioral mismatch and long-tail interactions.
- **UR + Replay (URR):** We use the same unseen-reward partners as in UR, but replay their pre-recorded trajectories during evaluation, removing online reactivity.
- **Log-Replay (LR):** Partners replay human trajectories, providing a natural, non-reactive long-tail evaluation setting.

Metrics. We report four metrics: collision per agent, success score, off-road per agent, and lane-alignment rate. Collision per agent and off-road per agent denote the average number of collisions and off-road events per agent over 910 steps, respectively. Success score is the fraction of agents that reach the goal without any collision or off-road event.

Methods. We compare three methods to assess the effectiveness of PBT for zero-shot coordination and to analyze the behavior of recorded-PBT. All methods use PPO and are trained for 1 billion environment steps. **Self-Play (SP)** trains all controllable agents jointly with the nominal reward. **Reactive-PBT** designates 25% of agents as ego agents in each scenario, while the remaining agents are controlled online by sampled pre-trained population policies. The ego/partner assignment is resampled whenever the map is resampled. **ReCord** uses the same ego-agent assignment, but replaces online partner-policy execution with trajectory replay. Specifically, we pre-record 50 trajectories per scenario from the pre-trained population, and during training, non-ego agents replay one randomly selected recorded trajectory.

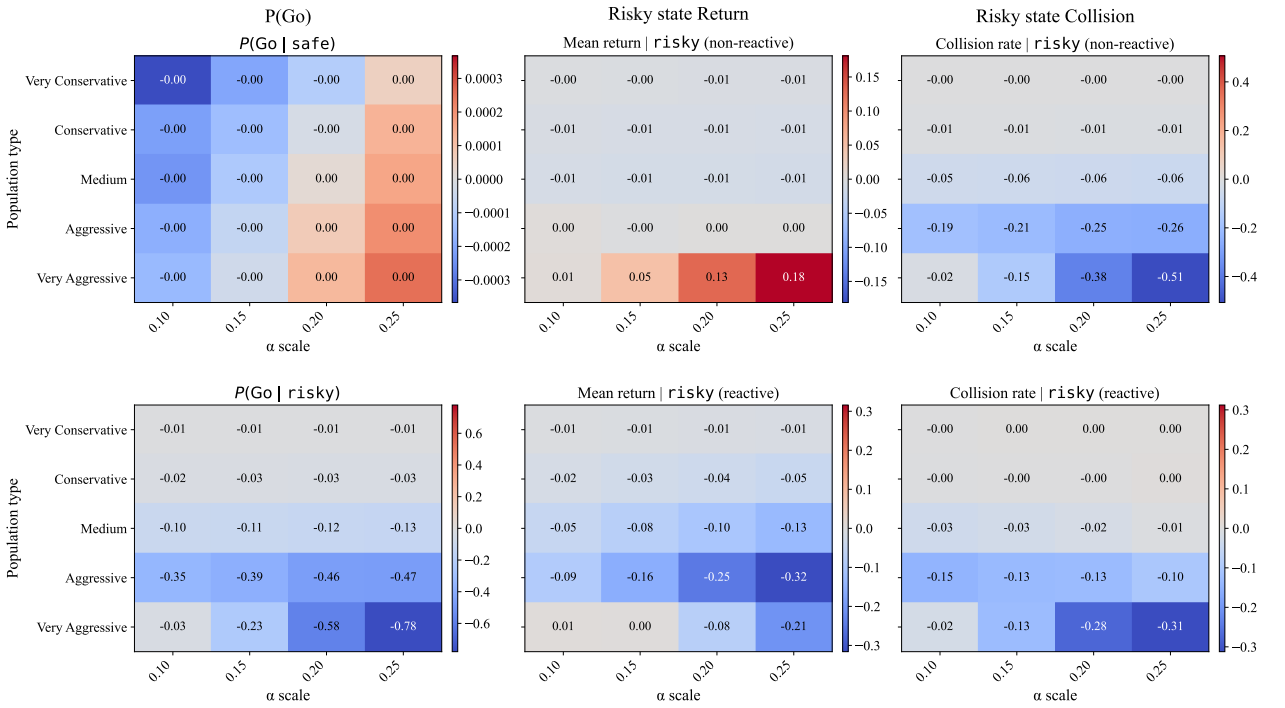

Figure 2. Toy-game heatmaps results: **ReCord** − **Reactive-PBT** across population families and risky-state reactivity scales α. $P(\mathrm{Go})$ captures policy aggressiveness, where lower values indicate more conservative behavior. **Return** captures average payoff in risky states, where higher is better. **Collision Rate** captures the frequency of the unsafe $(\mathrm{go}, \mathrm{go})$ outcome in risky states, where lower is better.

6. Results

6.1. Matrix Game

Figure 2 shows three consistent patterns. First, the two methods behave almost identically in the safe state, but differ clearly in the risky state. In particular, ReCord consistently learns a lower $p(\mathrm{go})$ than reactive-PBT in risky situations, indicating a more conservative policy. This gap is relatively small for conservative populations, but grows substantially as the population becomes more aggressive, reaching nearly -0.8 for the *Very Aggressive* population at $\alpha = 0.25$. This result suggests that ReCord more strongly suppresses risky *go* decisions when the partner population is less nominal and more aggressive.

Second, this behavioral difference translates directly into a safety benefit. ReCord consistently achieves lower collision rates than reactive-PBT under both non-reactive and reactive evaluation. Under non-reactive evaluation, the collision-rate gap is negative in nearly all settings and becomes particularly large for aggressive populations, reaching up to -0.51 in the *Very Aggressive* case. A similar pattern also appears under reactive evaluation: although the partner can still adapt online, ReCord continues to reduce collisions, with the advantage becoming more pronounced as the partner population becomes more aggressive. Overall, these results indicate that ReCord yields a robust reduction in unsafe $(\mathrm{go}, \mathrm{go})$ outcomes.

Third, the return results reveal a clear dependence on the evaluation setting. Under non-reactive evaluation, ReCord generally achieves comparable or higher return than reactive-PBT, and its advantage increases as the population becomes more aggressive and the reactivity scale α grows. For example, in the *Very Aggressive* population, the return gap increases from 0.01 to 0.18 as α increases from 0.1 to 0.25. In contrast, under reactive evaluation, ReCord usually obtains a lower return than reactive-PBT, and this disadvantage also becomes larger as α grows. This suggests that ReCord is particularly effective against less-responsive, long-tail opponents, whereas reactive-PBT can achieve a higher task return when the evaluation partner remains adaptive. Taken together, the matrix-game results show that ReCord trades some return in reactive settings for substantially safer coordination, while providing both safety and return benefits in non-reactive settings.

6.2. PufferDrive

We first compare zero-shot coordination performance in the L+N population, which mixes nominal and aggressive agents. For US evaluation, we use nominal-style agents with unseen random seeds, while for UR and URR evaluation, we use a mixture of slow and fast agents to introduce behavioral mismatch.

Table 2. Comparison on zero-shot coordination performance. The PBT was trained by the L+N population. (mean ± std across seeds).

Evaluation	Metric	Reactive-PBT	ReCord	Self-play
Log-replay (LR)	**Collision per agent ↓**	0.646 ± 0.027	$\mathbf{0.373 \pm 0.039}$	0.874 ± 0.014
	Off-road per agent ↓	0.157 ± 0.002	$\mathbf{0.139 \pm 0.007}$	0.165 ± 0.033
	Success score ↑	0.921 ± 0.005	$\mathbf{0.943 \pm 0.005}$	0.920 ± 0.002
	Lane alignment rate ↑	$\mathbf{0.885 \pm 0.005}$	0.878 ± 0.010	0.852 ± 0.005
Unseen seeds (US)	**Collision per agent ↓**	$\mathbf{0.096 \pm 0.001}$	0.104 ± 0.004	0.180 ± 0.006
	Off-road per agent ↓	$\mathbf{0.133 \pm 0.005}$	0.136 ± 0.005	0.138 ± 0.026
	Success score ↑	$\mathbf{0.977 \pm 0.001}$	0.975 ± 0.004	0.973 ± 0.002
	Lane alignment rate ↑	$\mathbf{0.902 \pm 0.007}$	0.887 ± 0.010	0.867 ± 0.007
Unseen rewards (UR)	**Collision per agent ↓**	0.145 ± 0.005	$\mathbf{0.128 \pm 0.006}$	0.233 ± 0.009
	Off-road per agent ↓	0.135 ± 0.005	$\mathbf{0.134 \pm 0.003}$	0.140 ± 0.025
	Success score ↑	$\mathbf{0.970 \pm 0.000}$	0.962 ± 0.003	0.966 ± 0.002
	Lane alignment rate ↑	$\mathbf{0.900 \pm 0.006}$	0.885 ± 0.010	0.865 ± 0.007
UR + replay (URR)	**Collision per agent ↓**	0.283 ± 0.025	$\mathbf{0.210 \pm 0.009}$	0.380 ± 0.010
	Off-road per agent ↓	0.144 ± 0.005	$\mathbf{0.141 \pm 0.004}$	0.145 ± 0.024
	Success score ↑	$\mathbf{0.970 \pm 0.001}$	0.962 ± 0.003	0.966 ± 0.002
	Lane alignment rate ↑	$\mathbf{0.900 \pm 0.007}$	0.884 ± 0.010	0.865 ± 0.007

As shown in Table 2, both reactive-PBT and ReCord substantially outperform self-play across all zero-shot coordination settings, achieving much lower collision rates and higher lane alignment rates, which indicates more stable driving overall. However, the relative strengths of the two PBT variants vary with the coordination mode. In the US setting, reactive-PBT is slightly better than ReCord on all metrics, suggesting that online reactivity remains beneficial when coordinating with nominal but unseen partners. In contrast, in the UR and URR setting, ReCord achieves fewer collisions while maintaining comparable off-road performance, indicating a safer policy in UR and URR mode. The clearest advantage of ReCord appears in the log-replay setting, where it achieves the highest success score and the lowest total collision rate (off-road and vehicle) among all methods. **Overall, these results suggest that PBT itself improves robustness over self-play by exposing the ego policy to diverse partners during training, while the additional benefit of ReCord becomes most evident in long-tail settings involving non-reactive (LR) or behaviorally mismatched partners (UR and URR).**

6.3. The effect of ReCord depends on population

ReCord is not uniformly effective across all population settings as in Section 6.1. To understand its effect more precisely, we ask two questions: (1) Does ReCord become more effective when the training population is more long-tailed, and (2) Does it become more effective when the population is more diverse?

To answer the first question, we train the reactive-PBT and ReCord with **Only Nominal (ON)** and **Only Lane (OL)** populations, representing nominal-only and long-tail-only settings, respectively. We then evaluate zero-shot coordination using unseen-seed agents from the same population type and unseen-reward agents drawn from populations excluded during training. As in Table 3, the relative advantage of ReCord becomes more pronounced as the evaluation partners are far from the trained population (US → UR → URR → LR). Population type primarily affects the relative advantage between ReCord and reactive-PBT. Training with the ON population consistently yields larger gains for ReCord, especially in collision reduction and lane alignment. In contrast, training with the OL population makes Reactive-PBT substantially more competitive, particularly under reactive unseen-partner evaluation, where the collision advantage of ReCord diminishes, and the lane alignment advantage often reverses.

For the second question, we compare **Lane + Nominal (L+N)** and **Mixed** populations to examine the role of diversity. In this case, we focus on log-replay zero-shot coordination to measure how diversity in the training population changes the relative performance of reactive-PBT and ReCord. Clearly, as population diversity grows, LR evaluation consistently improves zero-shot coordination performance for both reactive-PBT and ReCord. Compared with training on a single population, training on a mixed population reduces collision and off-road events while increasing the success rate in both methods. Notably, ReCord on the Mixed population achieves the best overall performance, yielding the lowest collision and off-road rates and the highest success score across all settings. This suggests that ex-

Table 3. Difference between ReCord and reactive-PBT across ON and OL populations. Values denote ReCord minus reactive-PBT; negative values are better for ↓ metrics, and positive values are better for ↑ metrics.

Evaluation	Metric	Only Nominal (ON)	Only Lane Breaker (OL)
Log-replay (LR)	**Collision per agent ↓**	$\mathbf{-0.405 \pm 0.057}$	-0.158 ± 0.028
	Off-road per agent ↓	-0.008 ± 0.022	$\mathbf{-0.009 \pm 0.028}$
	Success score ↑	$+0.014 \pm 0.012$	$\mathbf{+0.015 \pm 0.014}$
	Lane alignment rate ↑	$\mathbf{+0.016 \pm 0.017}$	-0.012 ± 0.007
Unseen seeds (US)	**Collision per agent ↓**	$\mathbf{+0.001 \pm 0.009}$	$+0.025 \pm 0.004$
	Off-road per agent ↓	$\mathbf{+0.020 \pm 0.020}$	$+0.030 \pm 0.023$
	Success score ↑	$\mathbf{-0.004 \pm 0.001}$	-0.011 ± 0.001
	Lane alignment rate ↑	$\mathbf{+0.011 \pm 0.017}$	-0.016 ± 0.004
Unseen rewards (UR)	**Collision per agent ↓**	$\mathbf{-0.034 \pm 0.012}$	-0.009 ± 0.012
	Off-road per agent ↓	$+0.013 \pm 0.021$	$\mathbf{+0.006 \pm 0.025}$
	Success score ↑	-0.008 ± 0.002	$\mathbf{-0.008 \pm 0.002}$
	Lane alignment rate ↑	$\mathbf{+0.011 \pm 0.018}$	-0.015 ± 0.004
UR + replay (URR)	**Collision per agent ↓**	$\mathbf{-0.089 \pm 0.016}$	-0.031 ± 0.017
	Off-road per agent ↓	$+0.012 \pm 0.021$	$\mathbf{+0.003 \pm 0.027}$
	Success score ↑	$\mathbf{-0.008 \pm 0.002}$	-0.008 ± 0.002
	Lane alignment rate ↑	$\mathbf{+0.009 \pm 0.018}$	-0.016 ± 0.005

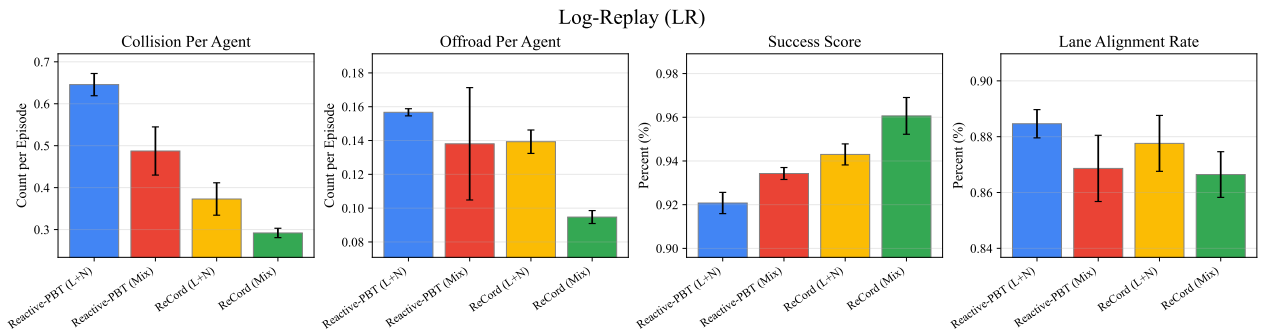

Figure 3. Log-replay zero-shot coordination performance across L+N and Mixed populations for reactive-PBT and ReCord.

posure to a more diverse set of partner behaviors improves robustness to non-reactive human trajectories. However, the benefits of diversity are not uniform across all metrics: lane alignment does not improve with mixing and is, in fact, slightly lower than in the single-population setting. Overall, these results indicate that diverse populations primarily improve the safety–success trade-off under log replay, with this effect being strongest for replay-based training.

6.4. Sample Efficiency of ReCord

The reactive-PBT is less sample-efficient than self-play for two reasons. First, because part of the simulation is occupied by pre-trained partner agents rather than the ego policy, fewer ego-policy samples are collected under the same simulation budget. Second, reactive partners must perform online inference at every timestep, which slows training. In contrast, ReCord simply replays pre-recorded trajectories, avoiding online partner inference. As shown in Figure 4, ReCord achieves more than three times higher steps per second (SPS) than reactive-PBT under the same setting.

7. Conclusion

We presented ReCord, a simple and effective approach for zero-shot coordination in autonomous driving. By replaying the pre-recorded trajectories, our approach encourages the ego policy to learn more conservative behaviors and avoids over-reliance on partner adaptation. From our experiments, ReCord showed its strongest advantage in non-reactive settings, which the partner policy does not yield, so forcing the ego policy to adapt. From the population-comparison experiment, we found that ReCord benefits increase as population diversity increases. In addition, ReCord provides a substantial practical benefit, achieving more than 3× higher

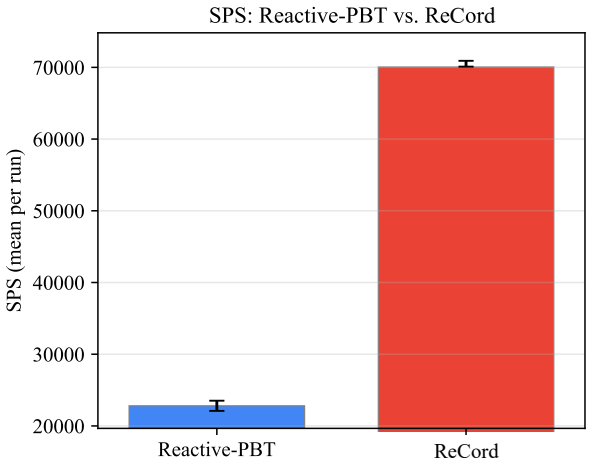

Figure 4. SPS comparison across three different seeds.

training speed than reactive-PBT.

Our study remains an interesting direction for future work. First, the population design remains simple, modeling long-tail behavior primarily through speed preferences and lane alignment, whereas real-world long-tail driving is much broader (Wang et al., 2025b; Xu et al., 2025). We expect that improving long-tail design would help build a rich population. Second, diverse population generation (Wang et al., 2024) could improve robustness by increasing behavioral diversity. Finally, our partner-selection strategy is purely random, and it remains unclear whether more advanced methods (Ruhdorfer et al., 2025; Wang et al., 2025a; Chaudhary et al., 2025) would further improve ReCord in the replay-based setting.

References

Bard, N., Foerster, J. N., Chandar, S., Burch, N., Lanctot, M., Song, H. F., Parisotto, E., Dumoulin, V., Moitra, S., Hughes, E., et al. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence*, 280:103216, 2020.

Bouchard, F., Sedwards, S., and Czarnecki, K. A rule-based behaviour planner for autonomous driving. In *International joint conference on rules and reasoning*, pp. 263–279. Springer, 2022.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Caesar, H., Kabzan, J., Tan, K. S., Fong, W. K., Wolff, E., Lang, A., Fletcher, L., Beijbom, O., and Omari, S. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021.

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., and Dragan, A. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32, 2019.

Chang, M.-F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8748–8757, 2019.

Chang, W.-J., Rangesh, A., Joseph, K., Strong, M., Tomizuka, M., yihan hu, and Zhan, W. SPACer: Self-play anchoring with centralized reference models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Q5H3NCy18S.

Chaudhary, P., Liang, Y., Chen, D., Du, S. S., and Jaques, N. Improving human-ai coordination through adversarial training and generative models. *arXiv e-prints*, pp. arXiv–2504, 2025.

Cornelisse, D. and Vinitsky, E. Human-compatible driving agents through data-regularized self-play reinforcement learning. In *Reinforcement Learning Conference*, 2025.

Cornelisse, D., Cheng, S., Mandavilli, P., Hunt, J., Joseph, K., Doulazmi, W., Charraut, V., Gupta, A., Suarez, J., and Vinitsky, E. PufferDrive: A fast and friendly driving simulator for training and evaluating RL agents. https://github.com/Emerge-Lab/PufferDrive, 2025a. Version 2.0.0. Equal contribution by the first two authors.

Cornelisse, D., Pandya, A., Joseph, K., Suárez, J., and Vinitsky, E. Building reliable sim driving agents by scaling self-play. *arXiv preprint arXiv:2502.14706*, 2025b.

Cusumano-Towner, M. F., Hafner, D., Hertzberg, A., Huval, B., Petrenko, A., Vinitsky, E., Wijmans, E., Killian, T. W., Bowers, S., Sener, O., et al. Robust autonomy emerges from self-play. In *Forty-second International Conference on Machine Learning*, 2025.

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., and Koltun, V. Carla: An open urban driving simulator. In *Conference on robot learning*, pp. 1–16. PMLR, 2017.

Elhenawy, M., Elbery, A. A., Hassan, A. A., and Rakha, H. A. An intersection game-theory-based traffic control algorithm in a connected vehicle environment. In *2015 IEEE 18th international conference on intelligent transportation systems*, pp. 343–347. IEEE, 2015.

Ettinger, S., Cheng, S., Caine, B., Liu, C., Zhao, H., Pradhan, S., Chai, Y., Sapp, B., Qi, C. R., Zhou, Y., et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9710–9719, 2021.

Gulino, C., Fu, J., Luo, W., Tucker, G., Bronstein, E., Lu, Y., Harb, J., Pan, X., Wang, Y., Chen, X., et al. Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research. *Advances in Neural Information Processing Systems*, 36:7730–7742, 2023.

Jaderberg, M., Dalibard, V., Osindero, S., Czarnecki, W. M., Donahue, J., Razavi, A., Vinyals, O., Green, T., Dunning, I., Simonyan, K., et al. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*, 2017.

Kazemkhani, S., Pandya, A., Cornelisse, D., Shacklett, B., and Vinitsky, E. Gpudrive: Data-driven, multi-agent driving simulation at 1 million fps. In *The Thirteenth International Conference on Learning Representations*, 2025.

Lanctot, M., Zambaldi, V., Gruslys, A., Lazaridou, A., Tuyls, K., Pérolat, J., Silver, D., and Graepel, T. A unified game-theoretic approach to multiagent reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

Loo, Y., Gong, C., and Meghjani, M. A hierarchical approach to population training for human-ai collaboration. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 3011–3019, 2023.

Lou, X., Guo, J., Zhang, J., Wang, J., Huang, K., and Du, Y. Pecan: Leveraging policy ensemble for context-aware zero-shot human-ai coordination. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 679–688, 2023.

Philion, J., Peng, X. B., and Fidler, S. Trajeglish: Traffic modeling as next-token prediction. In *The Twelfth International Conference on Learning Representations*, 2024.

Ruhdorfer, C., Bortoletto, M., Oei, V., Penzkofer, A., and Bulling, A. Unsupervised partner design enables robust ad-hoc teamwork. *arXiv preprint arXiv:2508.06336*, 2025.

Sarkar, B., Shih, A., and Sadigh, D. Diverse conventions for human-ai collaboration. *Advances in neural information processing systems*, 36:23115–23139, 2023.

Schwarting, W., Alonso-Mora, J., and Rus, D. Planning and decision-making for autonomous vehicles. *Annual Review of Control, Robotics, and Autonomous Systems*, 1 (1):187–210, 2018.

Strouse, D., McKee, K., Botvinick, M., Hughes, E., and Everett, R. Collaborating with humans without human data. *Advances in neural information processing systems*, 34:14502–14515, 2021.

Suarez, J. PufferLib 2.0: Reinforcement learning at 1m steps/s. *Reinforcement Learning Journal*, 6:1378–1388, 2025.

Vinitsky, E., Lichtlé, N., Yang, X., Amos, B., and Foerster, J. Nocturne: a scalable driving benchmark for bringing multi-agent learning one step closer to the real world. *Advances in Neural Information Processing Systems*, 35: 3962–3974, 2022.

Wang, C., Rahman, A., Cui, J., Sung, Y., and Stone, P. Rotate: Regret-driven open-ended training for ad hoc teamwork. In *Second Coordination and Cooperation in Multi-Agent Reinforcement Learning Workshop*, 2025a.

Wang, H., Meng, Q., Chen, S., and Zhang, X. Competitive and cooperative behaviour analysis of connected and autonomous vehicles across unsignalised intersections: A game-theoretic approach. *Transportation research part B: methodological*, 149:322–346, 2021.

Wang, X., Zhang, S., Zhang, W., Dong, W., Chen, J., Wen, Y., and Zhang, W. Zsc-eval: An evaluation toolkit and benchmark for multi-agent zero-shot coordination. *Advances in Neural Information Processing Systems*, 37: 47344–47377, 2024.

Wang, Y., Luo, W., Bai, J., Cao, Y., Che, T., Chen, K., Chen, Y., Diamond, J., Ding, Y., Ding, W., et al. Alpamayo-r1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. *arXiv preprint arXiv:2511.00088*, 2025b.

Wang, Z., Rahmani, S., Cornelisse, D., Sarkar, B., Goldie, A. D., Foerster, J. N., and Whiteson, S. Learning to drive in new cities without human demonstrations. *arXiv preprint arXiv:2602.15891*, 2026.

Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J. K., et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Wu, W., Feng, X., Gao, Z., and Kan, Y. Smart: Scalable multi-agent real-time motion generation via next-token prediction. *Advances in Neural Information Processing Systems*, 37:114048–114071, 2024.

Xu, R., Lin, H., Jeon, W., Feng, H., Zou, Y., Sun, L., Gorman, J., Tolstaya, K., Tang, S., White, B., et al. Wod-e2e: Waymo open dataset for end-to-end driving in challenging long-tail scenarios. *arXiv preprint arXiv:2510.26125*, 2025.

Zhang, Z., Karkus, P., Igl, M., Ding, W., Chen, Y., Ivanovic, B., and Pavone, M. Closed-loop supervised fine-tuning of tokenized traffic models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5422–5432, 2025.

Zhao, R., Song, J., Yuan, Y., Hu, H., Gao, Y., Wu, Y., Sun, Z., and Yang, W. Maximum entropy population-based training for zero-shot human-ai coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 6145–6153, 2023.

A. Population Setting

A.1. Matrix Game

Table 4. Risky-state base probability $p_{go}^{base}(s, i)$ ith of policy in population name P

Population P	$i = 0$	$i = 1$	$i = 2$
Very Conservative	0.20	0.25	0.30
Conservative	0.35	0.40	0.45
Medium	0.45	0.50	0.55
Aggressive	0.50	0.55	0.60
Very Aggressive	0.60	0.65	0.70

$p_{\text{base}}(\texttt{risky}, t)$ depends on the population type as in Table 4.

A.2. PufferDrive

We first train four types of partner policies, as summarized in Table 5, each with eight different random seeds: Using these policies, we construct four training populations, as shown in Table 6. The default population is L+N, which consists of nominal and lane-breaker policies. To analyze the effect of population composition, we additionally construct ON, which contains only nominal policies, and OL, which contains only lane-breaker policies. Finally, we construct a Mixed population that includes all policy types.

Table 5. Policy types defined by reward coefficients.

	Nominal	**Fast**	**Slow**	**Lane Breaker**
β_4	0	0.03	0.03	0
β_5	0	0	0	0.02

Table 6. Population settings in PufferDrive experiment.

	Nominal	**Fast**	**Slow**	**Lane Breaker**
L+N	4	0	0	4
ON	4	0	0	0
OL	0	0	0	4
Mixed	2	2	2	2

B. Population Results

In this section, we evaluate the population's performance to show that it is comparably optimized yet exhibits different styles. To do this, we run the self-play evaluation for 10,000 test scenarios and measure the success rate, collision, off-road, and lane alignment rate.

Table 7. Population self-play in PufferDrive (mean ± std across models).

	L+N	**ON**	**OL**	**Mixed**
Lane alignment	0.577 ± 0.253	$\mathbf{0.825 \pm 0.011}$	0.338 ± 0.033	0.663 ± 0.203
Collisions per agent	0.117 ± 0.007	$\mathbf{0.112 \pm 0.006}$	0.124 ± 0.003	0.116 ± 0.016
Off-road per agent	0.097 ± 0.017	0.095 ± 0.009	0.108 ± 0.004	$\mathbf{0.089 \pm 0.016}$
Success Rate	0.963 ± 0.021	$\mathbf{0.983 \pm 0.002}$	0.934 ± 0.028	0.942 ± 0.045