DISCOVERING SAFETY SPECTRA IN AUTONOMOUS DRIVING VIA CONSTRAINED ADVERSARIAL SELF-PLAY

Anonymous authors
Paper under double-blind review

ABSTRACT

Naturalistic traffic data contain few safety-critical interactions, limiting their ability to expose the weaknesses of autonomous driving policies during training. Adversarial scenario generation can target these weaknesses, but it faces two challenges: optimizing only for safety violations may produce implausibly extreme behavior, while scenarios generated against a fixed policy may lose relevance as the protagonist evolves. To address these challenges, we introduce **Constrained Self-Play with Adversarial Rationality (Co-SPAR)**, a closed-loop framework that couples prior-constrained adversary optimization with region-wise protagonist optimization. On the adversary side, a deviation budget around a behavioral prior defines nested adversary classes with different attack strengths. Within each class, the adversary maximizes the constraint decay function (CDF), which measures the protagonist's future violation risk, to identify the worst admissible adversary. On the protagonist side, the resulting robust CDF defines a robust feasible region that certifies safety against every adversary in the class. This region guides the protagonist to reduce violation risk outside it and improve task return inside it while preserving safety. We prove that at each adversarial strength, exact policy iteration monotonically expands the robust feasible region, improves task performance within it, and converges to the optimal CDF. Across strengths, we establish a robust feasible region spectrum and prove that a protagonist optimal at a stronger strength retains its safety certificates against weaker adversary classes. Experiments show Co-SPAR advances the attack–realism Pareto frontier with a controllable adversarial spectrum and learns protagonist policies that outperform adversarial training baselines across nominal and adversarial traffic. Cross-strength evaluations demonstrate the predicted robust feasible region ordering and stronger-to-weaker safety transfer.

1 INTRODUCTION

Training autonomous driving policies for long-tail scenarios requires experience with safety-critical interactions, which are sparsely represented in naturalistic driving data. Collecting these interactions at scale on public roads is costly and may itself introduce safety risks (Kalra & Paddock, 2016; Zhao et al., 2016; Chen et al., 2024; Xu et al., 2022). Adversarial scenario generation addresses this gap in simulation by controlling surrounding traffic agents to expose policy weaknesses (Koren et al., 2018; Wang et al., 2021; Hanselmann et al., 2022; Ruan et al., 2026). The generated scenarios provide targeted tests of the current policy. For training, however, scenario generation must keep pace with an evolving policy and continue to expose its current weaknesses.

Existing adversarial scenario generation typically holds the driving policy fixed while searching for scenarios that expose its weaknesses (Wang et al., 2021; Hanselmann et al., 2022; Chen et al., 2024; Rempe et al., 2022). As the driving policy improves during training, its safety weaknesses shift, and scenarios tailored to an earlier policy may lose their value as training data. This limitation motivates closed-loop adversarial training, in which surrounding traffic agents and the driving policy adapt to each other through self-play (Zhang et al., 2024; Nie et al., 2026b).

Effective closed-loop self-play requires controlling both the traffic challenges posed by surrounding agents and how the driving policy learns from them. When optimized only to induce safety violations,

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.

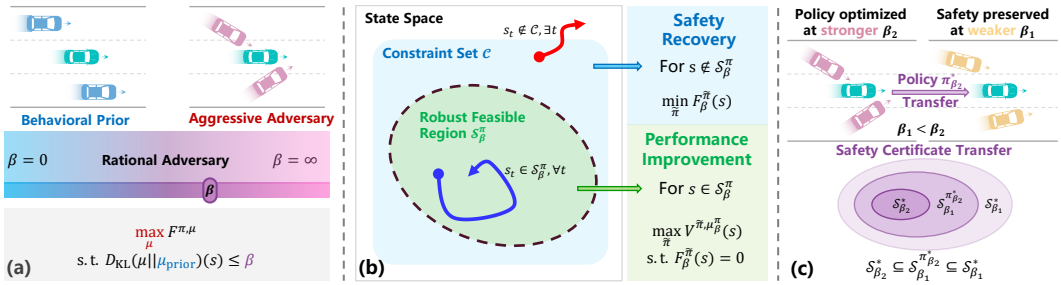


Figure 1: Overview of Co-SPAR. **(a) Prior-constrained adversary optimization:** at strength β, the adversary maximizes the CDF subject to a KL constraint around the behavioral prior. **(b) Region-wise protagonist optimization:** the robust CDF partitions the state space: the protagonist reduces future violation risk outside the robust feasible region and improves task return inside it while preserving safety. **(c) Safety certificate transfer:** a protagonist optimal at the stronger strength preserves the safety certificates for states in the feasible region $\mathcal{S}^*_{\beta_2}$ when evaluated at any weaker strength $\beta_1 < \beta_2$.

these agents may learn overly aggressive or implausible behavior (Chen et al., 2024; Ruan et al., 2026; Nie et al., 2026b). Conversely, keeping them close to nominal traffic behavior may leave the current policy's safety weaknesses unexposed. These cases motivate controlling adversarial strength by how far surrounding agents may depart from a common reference for traffic behavior (Rempe et al., 2022; Zhang et al., 2024). The driving policy, in turn, has two objectives across states: reducing constraint violation risk where safety has not been established and improving task performance where safety can be preserved.

We propose **Constrained Self-Play with Adversarial Rationality (Co-SPAR)**, a closed-loop self-play framework that turns surrounding traffic agents into adaptive sparring partners for the driving policy (Figure 1). These partners expose the policy's evolving safety weaknesses while remaining anchored to a behavioral prior. The deviation budget β controls how far these agents may depart from the prior. As β increases, the resulting adversary classes are nested. For each adversary class, Co-SPAR evaluates future violation risk through the constraint decay function (CDF) under the worst adversary in the class. A sublevel set of the resulting robust CDF defines a robust feasible region and provides a safety certificate against every adversary in the class. This region guides training to reduce future violation risk in states outside the region and improve task performance in states inside it. At any fixed adversarial strength, exact policy iteration monotonically expands the region and improves task performance within it. Across adversarial strengths, the robust feasible regions shrink as the adversary class grows, forming the robust feasible region spectrum. We also establish stronger-to-weaker transfer of safety certificates: a protagonist optimal at the stronger strength retains the certificates of its robust feasible region when evaluated against any weaker class.

The main contributions of this work are:

- We develop Co-SPAR as a closed-loop algorithm for autonomous driving that couples prior-constrained adversary optimization with region-wise protagonist optimization. The adversary is optimized under a deviation budget that controls attack strength, while the protagonist uses the robust CDF to reduce violation risk outside the robust feasible region and improve task return inside it.
- We show that sublevel sets of the robust CDF certify safety against every adversary in the corresponding class. At a fixed strength, we develop an operator view of exact Co-SPAR policy iteration and prove monotone expansion of the robust feasible region, improvement of task value within it, and convergence to the optimal CDF. Across strengths, we establish the robust feasible region spectrum and show that a protagonist optimal at a stronger strength retains its safety certificates against weaker classes.
- Experiments show that Co-SPAR produces a controllable adversarial spectrum with a stronger attack–realism Pareto frontier and higher protagonist driving success than baselines across nominal and adversarial traffic. Cross-strength evaluations demonstrate the predicted feasible region ordering and stronger-to-weaker safety transfer.

2 PRELIMINARIES

2.1 PROBLEM SETUP

We consider closed-loop autonomous driving interaction under a self-play framework with protagonist and adversary policy populations. For analysis, we use a two-role abstraction with one protagonist policy π and one adversary policy μ. At the implementation level, these role policies are instantiated by interacting decentralized policy collections; Appendix C.4 gives the corresponding population objective and its decomposition into individual updates.

We formalize this interaction as a constrained asymmetric Markov game $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{U}, P, r, c, \gamma)$, where $a \in \mathcal{A}$ is the protagonist action, $u \in \mathcal{U}$ is the adversary action, and $c(s) \in \{0, 1\}$ indicates state-wise safety violation, with constraint set $\mathcal{C} := \{s \in \mathcal{S} \mid c(s) = 0\}$. The asymmetry lies in the optimization roles: the protagonist is trained to maximize task return subject to safety constraints, whereas the adversary is trained only to worsen the protagonist's safety. The task value of the protagonist is the discounted return:

$$V^{\pi,\mu}(s) := \mathbb{E}_{\pi,\mu}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \;\middle|\; s_0 = s\right]. \tag{1}$$

The safety constraint of the protagonist is to ensure that, from a given state, the infinite horizon closed-loop trajectory never reaches a violation state:

$$s_t \in \mathcal{C}, \forall t \geq 0, \forall s_0 \in \mathcal{S}_{\text{init}}, \tag{2}$$

where $\mathcal{S}_{\text{init}}$ is the set of initial states under the initial distribution d.

2.2 FEASIBILITY FUNCTION

To aggregate this infinite horizon constraint into a single objective, a feasibility function, whose zero-level set represents the states that remain safe under future interaction, is introduced. We instantiate this feasibility function with the constraint decay function (CDF) (Yang et al., 2023), which captures the first violation time under the interaction between the protagonist and the adversary:

$$F^{\pi,\mu}(s) := \mathbb{E}_{\pi,\mu}\left[\gamma^{N(\tau)} \mid s_0 = s\right]. \tag{3}$$

Here $N(\tau) := \inf\{t \geq 0 \mid s_t \notin \mathcal{C}\}$ is the time of the first safety violation along trajectory τ. Thus $F^{\pi,\mu}(s) = 0$ if and only if the probability of ever reaching a violation from s is zero, while larger values reflect earlier violations under the discount scale. Thus, the infinite horizon safety constraint in Eq. 2 is equivalent to:

$$F^{\pi,\mu}(s) = 0, \forall s \in \mathcal{S}_{\text{init}}. \tag{4}$$

3 CONSTRAINED SELF-PLAY WITH ADVERSARIAL RATIONALITY

We develop **CO**nstrained **S**elf-**P**lay with **A**dversarial **R**ationality (Co-SPAR) through two coupled policy optimizations. In Co-SPAR, adversarial rationality is defined by how far an adversary policy may deviate from a behavioral prior. The prior-constrained adversary optimization controls this deviation. The region-wise protagonist optimization reduces future violation risk outside the robust feasible region and improves task return inside it while preserving safety.

3.1 PRIOR-CONSTRAINED ADVERSARY OPTIMIZATION

To expose safety weaknesses while anchoring adversarial behavior to observed traffic, we constrain the adversary with a behavioral prior. We enforce this restriction through a state-wise KL budget β. Let μ_{prior} denote this prior, obtained by behavior cloning from realistic driving data, and let $\beta \geq 0$ specify the allowed deviation from it. The behavioral interpretation of this constraint depends on how well μ_{prior} represents nominal traffic behavior over the states considered. For each state s, we define the prior-constrained policy set

$$\mathfrak{M}_\beta(s) := \left\{\mu(\cdot \mid s) \;\middle|\; D_{\text{KL}}\big(\mu(\cdot \mid s) \,\|\, \mu_{\text{prior}}(\cdot \mid s)\big) \leq \beta\right\}, \tag{5}$$

and let $\mathfrak{M}_\beta := \{\mu | \mu(\cdot \mid s) \in \mathfrak{M}_\beta(s), \forall s \in \mathcal{S}\}$ denote the induced prior-constrained adversary policy class. Because $\mathfrak{M}_{\beta_1} \subseteq \mathfrak{M}_{\beta_2}$ whenever $\beta_1 < \beta_2$, increasing β yields a larger and potentially stronger adversary class.

For a fixed protagonist policy π, the adversary update seeks the worst policy in $\mathfrak{M}_\beta$

$$\mu^\pi_\beta := \arg\max_{\mu \in \mathfrak{M}_\beta} F^{\pi,\mu}. \tag{6}$$

The corresponding robust CDF is defined by $F^\pi_\beta := F^{\pi,\mu^\pi_\beta}$. Therefore, strict safety against every $\mu \in \mathfrak{M}_\beta$ is equivalent to $F^\pi_\beta(s) = 0$. This robust CDF naturally induces the robust ε-feasible region and, at zero level, its strict variant:

$$\mathcal{S}^\pi_{\beta,\varepsilon} := \{s \in \mathcal{S} \mid F^\pi_\beta(s) \le \varepsilon\}, \qquad \mathcal{S}^\pi_\beta := \mathcal{S}^\pi_{\beta,0}. \tag{7}$$

Proposition 1 shows that these sublevel sets provide probabilistic safety certificates against every adversary in $\mathfrak{M}_\beta$; the zero-level set further certifies infinite horizon safety. The pairwise, robust, and optimal CDFs, their corresponding risky operators, and their fixed-point characterization are given in Appendix A.2.

3.2 Region-wise Protagonist Optimization

The robust CDF partitions the state space into states that are already certified safe and states that still require safety improvement, inducing two corresponding protagonist update objectives. Given the current protagonist policy and μ^π_β, the protagonist reduces future violation risk outside the robust feasible region and improves return inside it subject to the corresponding feasibility constraint. Outside $\mathcal{S}^\pi_\beta$, the protagonist update prioritizes safety recovery by minimizing the robust CDF:

$$\pi' = \arg\min_{\tilde\pi} F^{\tilde\pi}_\beta(s), \qquad \forall s \notin \mathcal{S}^\pi_\beta, \tag{8}$$

while inside $\mathcal{S}^\pi_\beta$, it improves task return under μ^π_β while preserving robust safety:

$$\pi' = \arg\max_{\tilde\pi} V^{\tilde\pi,\mu^\pi_\beta}(s) \quad \text{s.t.} \quad F^{\tilde\pi}_\beta(s) = 0, \qquad \forall s \in \mathcal{S}^\pi_\beta. \tag{9}$$

The policy iteration formulation of this update is developed in Appendix A.1 and A.4. Section 4 establishes monotone expansion of the robust feasible region and monotone task improvement within it.

3.3 Practical Algorithm

We instantiate the prior-constrained adversary optimization and the region-wise protagonist optimization as a practical policy gradient algorithm with function approximation. The protagonist uses a policy network π_θ, a task-value critic V_ψ, and a CDF critic F_{ϕ^π}. The adversary uses a policy network μ_ξ and a CDF critic F_{ϕ^μ}. Because the protagonist and adversary condition on different observations, we use separate CDF networks F_{ϕ^π} and F_{ϕ^μ} to provide the estimates required by their respective policy updates. We approximate the theoretical feasible region $\mathcal{S}^\pi_\beta$ by the protagonist-side sublevel set $\widehat{\mathcal{S}} := \{s \in \mathcal{S} \mid F_{\phi^\pi}(s) \le \epsilon\}$, where $\epsilon > 0$ serves as a tolerance for critic estimation errors.

To implement the region-wise protagonist objective, we construct a single protagonist advantage whose CDF penalty depends on the estimated robust feasible region. We estimate the task and CDF advantages using GAE-style recursions, whose temporal-difference definitions are given in Appendix C.2. Let $A^{V,\pi}$ denote the protagonist task advantage, and let $A^{F,\pi}$ denote the CDF advantage estimated for π. Using the estimated feasible region, we define the protagonist's region-wise advantage as

$$\bar A^\pi := \mathbf{1}[F_{\phi^\pi}(s) \le \epsilon]\left(A^{V,\pi} - w(s)A^{F,\pi}\right) - \mathbf{1}[F_{\phi^\pi}(s) > \epsilon]\, w_{\max} A^{F,\pi}, \tag{10}$$

where $w(s_t) := C(1/(\alpha - r(s_t)) - 1/\alpha)$ is the barrier weight induced by the feasibility ratio $r(s_t) := \min(F_{\phi^\pi}(s_t)/\epsilon, 1)$, with $C := w_{\max}\alpha(\alpha-1)$ and $\alpha > 1$. This barrier weight increases the CDF penalty as a state approaches the feasible region boundary and reaches $w_{\max}$ at the boundary.

For the adversary, $A^{F,\mu}$ denotes a separate CDF advantage under μ. Maximizing this advantage exposes earlier safety violations, while the dual variable penalizes deviation from μ_{prior} and regulates

the empirical KL around β. We optimize the protagonist and adversary with the following clipped surrogate objectives:

$$\begin{aligned} L_\pi(\theta) &:= \mathbb{E}\left[\min\Big(\rho_\pi\, \bar{A}^\pi, \mathrm{clip}(\rho_\pi, 1-\delta, 1+\delta)\bar{A}^\pi\Big)\right], \\ L_\mu(\xi) &:= \mathbb{E}\left[\min\Big(\rho_\mu\, A^{F,\mu}, \mathrm{clip}(\rho_\mu, 1-\delta, 1+\delta)A^{F,\mu}\Big)\right] - \eta_\beta \mathbb{E}\left[D_{\mathrm{KL}}\big(\mu_\xi \,\|\, \mu_{\mathrm{prior}}\big)\right], \end{aligned} \tag{11}$$

where $\rho_\pi := \pi_\theta(a \mid s)/\pi_{\theta_{\mathrm{old}}}(a \mid s)$ and $\rho_\mu := \mu_\xi(u \mid s)/\mu_{\xi_{\mathrm{old}}}(u \mid s)$. Here $\eta_\beta > 0$ is the multiplier for the adversary policy constraint. We increase it when the sampled adversary KL exceeds β and decrease it otherwise:

$$\eta_\beta \leftarrow \eta_\beta \lambda_\eta^{\mathrm{sgn}(\mathbb{E}[D_{\mathrm{KL}}(\mu_\xi \,\|\, \mu_{\mathrm{prior}})]-\beta)}, \tag{12}$$

where $\lambda_\eta > 1$ is the multiplicative adjustment factor. Implementation details and KL training traces are given in Appendices C.3 and D.5. The full training loop is summarized in Algorithm 1.

4 Theoretical Guarantees

In this section, we establish the theoretical foundations of Co-SPAR. At a fixed adversarial strength β, a robust feasible region provides safety certificates, and exact Co-SPAR policy iteration expands this region, improves task value within it, and converges to the optimal CDF. The CDF and task operators underlying this policy iteration are developed in Appendix A. Across adversarial strengths, nested adversary classes induce the robust feasible region spectrum; a protagonist optimal at a stronger strength retains its safety certificates against weaker adversary classes.

Safety certificates for robust feasible regions. We first state the probabilistic safety certificate provided by membership in a robust ε-feasible region. For every adversary in $\mathfrak{M}_\beta$, this guarantee bounds the probability of reaching a violation by horizon H.

Proposition 1 (Safety certificate for a robust ε-feasible region)**.** *If $s \in \mathcal{S}^\pi_{\beta,\varepsilon}$, then for every adversary policy $\mu \in \mathfrak{M}_\beta$ and every horizon $H \geq 0$,*

$$\mathbb{P}^{\pi,\mu}_s\big(N(\tau) \leq H\big) \leq \varepsilon/\gamma^H. \tag{13}$$

Proof. Proof in Appendix A.3.2.

At the zero level, $\varepsilon = 0$ yields the strict safety certificate: if $s \in \mathcal{S}^\pi_\beta$, then $\mathbb{P}^{\pi,\mu}_s(N(\tau) < \infty) = 0$ for every $\mu \in \mathfrak{M}_\beta$; see Appendix A.3.2 for the proof.

Monotone expansion of the robust feasible region. We next analyze the policy iteration dynamics at a fixed adversarial strength β. The first guarantee concerns monotone expansion of the current robust feasible region.

Theorem 1 (Monotone expansion of the robust feasible region)**.** *Let $\{\pi_k\}_{k\geq 0}$ be the sequence generated by Co-SPAR policy iteration in Proposition 6. Then*

$$F^{\pi_{k+1}}_\beta(s) \leq F^{\pi_k}_\beta(s), \qquad \forall s \in \mathcal{S}, \tag{14}$$

and consequently, defining $\mathcal{S}_k := \{s \in \mathcal{S} \mid F^{\pi_k}_\beta(s) = 0\}$, we have

$$\mathcal{S}_k \subseteq \mathcal{S}_{k+1}. \tag{15}$$

Proof. Proof in Appendix A.4.5.

This monotonicity follows from the region-wise update: it reduces the robust CDF outside the current feasible region while preserving feasibility within it. Within the preserved region, the same update also yields monotone task improvement under the adversary used to define the current robust CDF. The formal statement and proof are given in Theorem 3 in Appendix A.4.6. Beyond monotone expansion, Corollary 2 in Appendix A.4.5 establishes that the robust CDF sequence converges uniformly to the optimal CDF $F^*_\beta(s) := \min_\pi F^\pi_\beta(s)$, which induces the optimal ε-feasible region $\mathcal{S}^*_{\beta,\varepsilon} := \{s \in \mathcal{S} \mid F^*_\beta(s) \leq \varepsilon\}$ and its strict variant $\mathcal{S}^*_\beta$.

Robust feasible region spectrum. The preceding analysis fixes the adversarial strength β and shows that policy iteration expands the robust feasible region and converges to F^*_β. We now vary β and use the nested adversary classes to compare the corresponding robust CDFs and robust feasible regions. As β increases, the adversary class expands. Consequently, both the robust CDF and the optimal CDF can only increase, while the feasible regions can only shrink. The spectrum result orders CDFs across adversarial strengths.

Proposition 2 (Robust feasible region spectrum). *If $\beta_1 < \beta_2$, then*

$$F^{\pi}_{\beta_1} \le F^{\pi}_{\beta_2}, \qquad F^{*}_{\beta_1} \le F^{*}_{\beta_2}. \tag{16}$$

Consequently, for every $\varepsilon \ge 0$,

$$\mathcal{S}^{\pi}_{\beta_2,\varepsilon} \subseteq \mathcal{S}^{\pi}_{\beta_1,\varepsilon}, \qquad \mathcal{S}^{*}_{\beta_2,\varepsilon} \subseteq \mathcal{S}^{*}_{\beta_1,\varepsilon}. \tag{17}$$

Proof. Proof in Appendix B.1.

Safety certificate transfer. We next evaluate a policy that is optimal at the stronger strength under the weaker adversary class. The resulting comparison takes the form of a pointwise sandwich.

Theorem 2 (Safety certificate transfer). *Let $\beta_1 < \beta_2$ and let $\pi^\star_{\beta_2}$ be any protagonist policy satisfying $F^{\pi^\star_{\beta_2}}_{\beta_2} = F^{*}_{\beta_2}$. Then*

$$F^{*}_{\beta_1} \le F^{\pi^\star_{\beta_2}}_{\beta_1} \le F^{*}_{\beta_2}. \tag{18}$$

Consequently, for every $\varepsilon \ge 0$,

$$\mathcal{S}^{*}_{\beta_2,\varepsilon} \subseteq \mathcal{S}^{\pi^\star_{\beta_2}}_{\beta_1,\varepsilon} \subseteq \mathcal{S}^{*}_{\beta_1,\varepsilon}. \tag{19}$$

Proof. Proof in Appendix B.2.

The left inclusion gives stronger-to-weaker transfer: every state in the optimal feasible region at β_2 remains certified when $\pi^\star_{\beta_2}$ is evaluated against the weaker adversary class at β_1. The right inclusion places this transferred region inside the optimal feasible region at β_1.

5 Experiments

Our experiments evaluate adversary attack–realism trade-offs, protagonist driving performance, and cross-strength safety behavior of Co-SPAR.

- **Q1.** Do Co-SPAR adversary policies yield a controllable attack–realism spectrum and a stronger Pareto frontier than competing methods?
- **Q2.** How do protagonists trained with Co-SPAR compare in driving success, task progress, and safety across naturalistic and adversarial traffic?
- **Q3.** Do the cross-strength CDF and policy evaluations exhibit the feasible region contraction and stronger-to-weaker safety transfer predicted by the theory?

Benchmark and protocol. We train in PufferDrive (Cornelisse et al., 2025) on 10,000 scenarios from the Waymo Open Motion Dataset (WOMD) (Ettinger et al., 2021) and evaluate on 1,000 scenarios held out from training. The protagonist aims to reach its route goal while avoiding collisions and off-road violations. For each held-out initial state, we run 32 rollouts with actions sampled from the evaluated policies. Vehicles are sampled into protagonist and adversary populations for each scene, with at least one protagonist retained. Both roles are evaluated on the same held-out scenarios, with the shared role proportion and optimization settings reported in Table 3.

Comparison methods. Across adversary and protagonist evaluations, we compare Co-SPAR with four baselines that differ in how they train adversaries and protagonists. (1) FREA (Chen et al., 2024) uses protagonist-side feasibility estimates to guide adversarial optimization toward critical yet feasible interactions. (2) RARL (Pinto et al., 2017) trains a protagonist against a destabilizing adversary

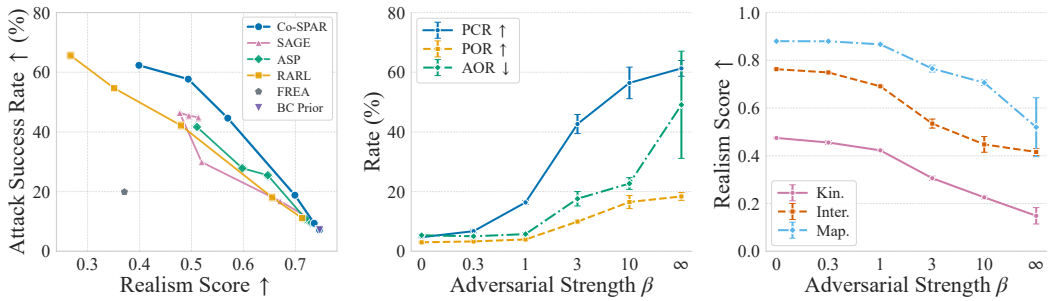


Figure 2: **Attack–realism trade-offs across a controllable adversarial spectrum.** **(a)** Attack success versus adversary realism across methods. **(b)** Protagonist collision rate (PCR), protagonist off-road rate (POR), and adversary off-road rate (AOR) across β. **(c)** Kinematic, interactive, and map-based WOSAC realism components across β. Results are equally averaged over Log Replay, BC Prior, HR-PPO, and BC Trained protagonists.

through a zero-sum minimax objective. (3) ASP (Zhang et al., 2024) uses asymmetric teacher–student self-play, in which the teacher challenges the student while demonstrating a solution to the same scenario. (4) SAGE (Nie et al., 2026a) uses test-time weight interpolation between two endpoint policies to control the attack–realism trade-off. To compare attack–realism trade-offs, we vary the available strength controls for Co-SPAR, RARL, ASP, and SAGE; FREA has no corresponding control and therefore contributes a single point. We additionally include HR-PPO (Cornelisse & Vinitsky, 2024) in the protagonist evaluation as a non-adversarial baseline. BC Prior denotes the behavior-cloned traffic policy μ_{prior}, while Log Replay uses the trajectories recorded in each scenario.

5.1 Adversary Policy Evaluation

Adversary evaluation protocol and metrics. We evaluate the adversaries from every method, together with BC Prior, against four protagonist conditions: Log Replay, BC Prior, HR-PPO, and BC Trained, the Co-SPAR protagonist trained against the fixed BC Prior ($\beta = 0$). We compare attack success rate (ASR) with the Waymo Open Sim Agents Challenge (WOSAC) realism score and its components (Montali et al., 2023), equally averaged over the four protagonist conditions (Figure 2). Full results for each protagonist condition are reported in Appendix D.2.

Adversarial spectrum and Pareto frontier. At comparable realism, Co-SPAR achieves higher attack success than competing baselines, yielding a stronger **attack–realism Pareto frontier** (Figure 2(a)). As adversarial strength β increases, protagonist collision and off-road rates rise, indicating stronger attacks, while the adversary off-road rate rises and the WOSAC realism components decline (Figure 2(b,c)). Varying β thereby traces a controllable adversarial spectrum along this frontier.

Spatial CDF predictions. The learned adversary-side CDF exhibits spatially localized high-risk regions (Figure 3). As β increases, these high CDF regions become more pronounced. Additional spatial CDF visualizations are shown in Figure 9.

5.2 Protagonist Policy Evaluation

Protagonist evaluation protocol and metrics. We evaluate protagonist performance under Log Replay, BC Prior, and each method's strongest adversary. We report the average of each metric across these adversaries. Success rate (SR) is the primary overall metric; safe route progress (SRP) and safety violation rate (SVR) measure task progress and safety violations, respectively. For methods with multiple protagonist configurations, we select the one with the highest SR on a Log Replay validation set; the selected settings are listed in Table 4. The adversary configurations and full per-adversary results are given in Appendix D.3.

Task performance and safety across traffic conditions. Co-SPAR achieves the highest SR and SRP under all three traffic conditions while attaining the second-lowest SVR (Table 1). Under Log

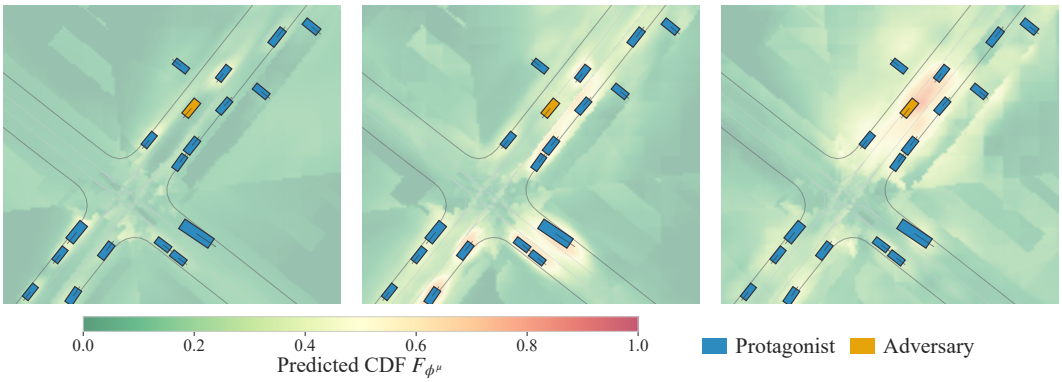


Figure 3: **Spatial CDF predictions across adversarial strengths.** We evaluate the adversary-side CDF critic after placing the orange adversarial vehicle at candidate road positions and aligning its heading with the road at each position. Color encodes the predicted future violation risk, treating all other vehicles as protagonists. The panels show $\beta = 0.3, 1, 3$ from left to right.

Table 1: Protagonist performance across nominal and adversarial traffic. SR requires goal completion without a safety violation; SRP and SVR denote safe route progress and safety violation rate. Strong Adv. Avg. averages results against each method's strongest evaluated adversary setting. Values are percentages; **bold** and underlined values mark the best and second-best results, respectively.

Protagonist	Log Replay			BC Prior			Strong Adv. Avg.		
	SR ↑	SRP ↑	SVR ↓	SR ↑	SRP ↑	SVR ↓	SR ↑	SRP ↑	SVR ↓
HR-PPO	60.8	83.7	**6.7**	60.5	83.5	**6.0**	56.6	79.2	**13.8**
FREA	66.1	86.7	9.7	66.0	86.7	8.8	61.5	82.1	16.5
RARL	69.7	88.2	8.6	69.4	88.1	8.0	64.4	83.4	15.7
ASP	69.6	88.7	8.5	69.3	88.6	8.0	63.9	83.5	16.2
SAGE	68.3	86.7	8.4	68.1	86.7	7.8	63.0	81.9	15.7
Co-SPAR	**82.2**	**93.4**	7.5	**82.3**	**93.6**	7.0	**75.8**	**88.4**	15.1

Replay and BC Prior, Co-SPAR's SR exceeds 82%, more than **12 pp** above the best baseline in each condition. Against strong adversaries, its SR is 75.8%, a **+11.4 pp** improvement over the best baseline. HR-PPO attains the lowest SVR but substantially lower SR across all three conditions.

5.3 Cross-Strength Evaluation

Robust feasible region spectrum. To visualize the feasible region spectrum, we use a car-following interaction with a two-dimensional state defined by relative distance and velocity. The shaded area below each boundary in Figure 4(a) marks the optimal feasible region at that adversarial strength. As β increases, the boundary moves downward and the shaded optimal feasible region contracts, consistent with the set inclusion in Proposition 2. The CDF fields are shown in Figure 7.

Cross-strength safety transfer. Theorem 2 establishes stronger-to-weaker safety certificate transfer for a protagonist optimal at β_2: against the weaker β_1 adversary class, its robust CDF lies between the optimal CDFs at the two strengths. We compare three cross-play pairs: both trained at β_1, the protagonist trained at β_2 against the adversary trained at β_1, and both trained at β_2. Figure 4(b) shows distributions of CDF values estimated by Monte Carlo at different initial states, with $\beta_1 = 0.3$ and $\beta_2 \to \infty$. A curve farther right represents greater violation risk, and the strong-to-weak transfer curve lies between the other two. Figure 4(c) reports SVR for the same pairings as β_1 varies with $\beta_2 \to \infty$ fixed. At each tested weaker strength β_1, the strong-to-weak SVR is higher than when both policies are trained at β_1 and lower than when both are trained at β_2. The full cross-play matrix is shown in Figure 8. The observed CDF and SVR ordering demonstrates safety transfer for the learned policies.

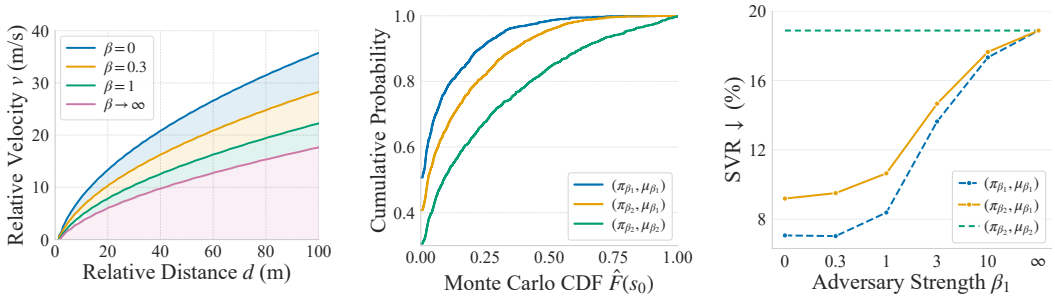


Figure 4: **Feasible region spectrum and safety transfer across adversarial strengths. (a)** Feasible region spectrum in a car-following interaction for $\beta \in \{0, 0.3, 1, \infty\}$, with shaded regions shrinking as β increases. **(b)** Distributions of Monte Carlo estimated CDF values at sampled initial states for the protagonist–adversary cross-play pairings $(\pi_{\beta_1}, \mu_{\beta_1})$, $(\pi_{\beta_2}, \mu_{\beta_1})$, and $(\pi_{\beta_2}, \mu_{\beta_2})$, with $\beta_1 = 0.3$ and $\beta_2 \to \infty$. **(c)** Safety violation rates for these three pairings as β_1 varies, with $\beta_2 \to \infty$ fixed.

Table 2: Ablations of adversarial strength and protagonist update. SR and SVR denote success rate and safety violation rate. All values are percentages.

(a) Adversarial Strength Ablation

Protagonist	Log Replay		Strong Adv. Avg.	
	SR ↑	SVR ↓	SR ↑	SVR ↓
BC Trained	81.4	7.7	74.7	15.3
Adv Trained	81.0	9.7	75.3	15.3
Co-SPAR	**82.2**	**7.5**	**75.8**	**15.1**

(b) Protagonist Update Ablation

Protagonist	Log Replay		Strong Adv. Avg.	
	SR ↑	SVR ↓	SR ↑	SVR ↓
Reward Penalty	60.7	8.4	56.4	15.4
CDF Penalty	79.7	7.8	73.4	15.4
Co-SPAR	**82.2**	**7.5**	**75.8**	**15.1**

5.4 Ablation Studies

Effect of adversarial strength. We compare Co-SPAR with its BC Trained ($\beta = 0$) and Adv Trained ($\beta \to \infty$) endpoints, which train against the fixed BC Prior and an adversary without the prior constraint, respectively. BC Trained has higher SR than Adv Trained under Log Replay, whereas Adv Trained has higher SR against strong adversaries (Table 2(a)). Co-SPAR attains higher SR than either endpoint under both Log Replay and strong adversaries.

Effect of the region-wise update. We compare the region-wise update with two uniform alternatives: Reward Penalty adds a safety penalty to the task reward, while CDF Penalty learns a separate CDF to penalize predicted risk (Table 2(b)). The CDF penalty achieves higher SR than the reward penalty, while the region-wise update attains the highest SR and lowest SVR in both settings.

6 Conclusion

We introduced Co-SPAR, which couples prior-constrained traffic adversaries with region-wise protagonist optimization. The deviation budget defines a spectrum of adversary classes and robust feasible regions. At a fixed strength, exact policy iteration expands the feasible region, improves task value within it, and converges to the optimal CDF. Across strengths, nested adversary classes order the robust feasible regions, and a protagonist optimal at the stronger strength retains its safety certificates against weaker classes. Co-SPAR traces a controllable adversarial spectrum with a stronger attack–realism Pareto frontier and achieves higher protagonist driving success than baselines across nominal and adversarial traffic. Cross-strength evaluations demonstrate the robust feasible region ordering and stronger-to-weaker safety transfer in learned policies.

AI use statement

In this work, we used generative AI tools for text refinement and code implementation assistance, but not for formulating scientific claims, novel ideas, or experimental designs. All outputs were manually reviewed: we verified the refined text to ensure the original scientific meaning was preserved, and

tested all assisted code for logic and correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Implementation details and experimental configurations are provided in Appendices C and D.1. The supplementary material includes code and launch scripts for the experiments.

REFERENCES

Wei-Jer Chang, Francesco Pittaluga, Masayoshi Tomizuka, Wei Zhan, and Manmohan Chandraker. Safe-sim: Safety-critical closed-loop traffic simulation with diffusion-controllable adversaries. In *European conference on computer vision*, pp. 242–258. Springer, 2024.

Keyu Chen, Yuheng Lei, Hao Cheng, Haoran Wu, Wenchao Sun, and Sifa Zheng. Frea: Feasibility-guided generation of safety-critical scenarios with reasonable adversariality. *arXiv preprint arXiv:2406.02983*, 2024.

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. `https://github.com/Emerge-Lab/PufferDrive`, 2025. Version 2.0.0.

Wenhao Ding, Baiming Chen, Minjun Xu, and Ding Zhao. Learning to collide: An adaptive safety-critical scenarios generating method. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2243–2250. IEEE, 2020.

Wenhao Ding, Haohong Lin, Bo Li, and Ding Zhao. Causalaf: Causal autoregressive flow for safety-critical driving scenario generation. In Karen Liu, Dana Kulic, and Jeff Ichnowski (eds.), *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 812–823. PMLR, 14–18 Dec 2023a.

Wenhao Ding, Chejian Xu, Mansur Arief, Haohong Lin, Bo Li, and Ding Zhao. A survey on safety-critical driving scenario generation—a methodological perspective. *IEEE Transactions on Intelligent Transportation Systems*, 24(7):6971–6988, 2023b.

Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, 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.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Niklas Hanselmann, Katrin Renz, Kashyap Chitta, Apratim Bhattacharyya, and Andreas Geiger. King: Generating safety-critical driving scenarios for robust imitation via kinematics gradients. In *European Conference on Computer Vision*, pp. 335–352. Springer, 2022.

Kunkun Hao, Wen Cui, Yonggang Luo, Lecheng Xie, Yuqiao Bai, Jucheng Yang, Songyang Yan, Yuxi Pan, and Zijiang Yang. Adversarial safety-critical scenario generation using naturalistic human driving priors. *IEEE Transactions on Intelligent Vehicles*, 9(9):5392–5406, 2023.

Nidhi Kalra and Susan M Paddock. Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transportation research part A: policy and practice*, 94:182–193, 2016.

Mark Koren, Saud Alsaif, Ritchie Lee, and Mykel J Kochenderfer. Adaptive stress testing for autonomous vehicles. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1–7. IEEE, 2018.

Sampo Kuutti, Saber Fallah, and Richard Bowden. Arc: Adversarially robust control policies for autonomous vehicles. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pp. 522–529. IEEE, 2021.

Zeyang Li, Chuxiong Hu, Yunan Wang, Yujie Yang, and Shengbo Eben Li. Safe reinforcement learning with dual robustness. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10876–10890, 2024.

Xiaobai Ma, Katherine Driggs-Campbell, and Mykel J Kochenderfer. Improved robustness and safety for autonomous vehicle control with adversarial reinforcement learning. *arXiv preprint arXiv:1903.03642*, 2019.

Nico Montali, John Lambert, Paul Mougin, Alex Kuefler, Nicholas Rhinehart, Michelle Li, Cole Gulino, Tristan Emrich, Zoey Yang, Shimon Whiteson, et al. The waymo open sim agents challenge. *Advances in Neural Information Processing Systems*, 36:59151–59171, 2023.

Tong Nie, Yuewen Mei, Yihong Tang, Junlin He, Jie Sun, Haotian Shi, Wei Ma, and Jian Sun. Steerable adversarial scenario generation through test-time preference alignment. In *International Conference on Learning Representations*, volume 2026, pp. 120964–121003, 2026a.

Tong Nie, Yihong Tang, Junlin He, Yuewen Mei, Jie Sun, Lijun Sun, Wei Ma, and Jian Sun. Adv-0: Closed-loop min-max adversarial training for long-tail robustness in autonomous driving. *arXiv preprint arXiv:2603.15221*, 2026b.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International conference on machine learning*, pp. 2817–2826. PMLR, 2017.

Joshua Ransiek, Johannes Plaum, Jacob Langner, and Eric Sax. Goose: Goal-conditioned reinforcement learning for safety-critical scenario generation. In *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 2651–2658. IEEE, 2024.

Davis Rempe, Jonah Philion, Leonidas J Guibas, Sanja Fidler, and Or Litany. Generating useful accident-prone driving scenarios via a learned traffic prior. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17284–17294. IEEE, 2022.

Yangang Ren, Letian Tao, Jingliang Duan, Honglin Li, Beiyan Jiang, and Shengbo Eben Li. Driving at realistic urban intersections with adversarial reinforcement learning. In *2024 IEEE International Automated Vehicle Validation Conference (IAVVC)*, pp. 1–6. IEEE, 2024.

Qiyu Ruan, Yuxuan Wang, He Li, Zhenning Li, and Cheng-zhong Xu. Scenepilot: Controllable boundary-driven critical scenario generation for autonomous driving. *arXiv preprint arXiv:2605.21168*, 2026.

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

Akifumi Wachi. Failure-scenario maker for rule-based agent using multi-agent adversarial reinforcement learning and its application to autonomous driving. *arXiv preprint arXiv:1903.10654*, 2019.

Jingkang Wang, Ava Pun, James Tu, Sivabalan Manivasagam, Abbas Sadat, Sergio Casas, Mengye Ren, and Raquel Urtasun. Advsim: Generating safety-critical scenarios for self-driving vehicles. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9904–9913. IEEE, 2021.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

Chejian Xu, Wenhao Ding, Weijie Lyu, Zuxin Liu, Shuai Wang, Yihan He, Hanjiang Hu, Ding Zhao, and Bo Li. Safebench: A benchmarking platform for safety evaluation of autonomous vehicles. *Advances in Neural Information Processing Systems*, 35:25667–25682, 2022.

Yujie Yang, Zhilong Zheng, Shengbo Eben Li, Wei Xu, Jingjing Liu, Xianyuan Zhan, and Ya-Qin Zhang. Feasible policy iteration for safe reinforcement learning. *arXiv preprint arXiv:2304.08845*, 2023.

Yujie Yang, Yuanxu Sun, Wenyu Li, Beiyan Jiang, Tao Zhang, and Shengbo Eben Li. Feasible policy optimization for safe reinforcement learning. 2026.

Dongjie Yu, Haitong Ma, Shengbo Li, and Jianyu Chen. Reachability constrained reinforcement learning. In *International conference on machine learning*, pp. 25636–25655. PMLR, 2022.

Chris Zhang, Sourav Biswas, Kelvin Wong, Kion Fallah, Lunjun Zhang, Dian Chen, Sergio Casas, and Raquel Urtasun. Learning to drive via asymmetric self-play. In *European Conference on Computer Vision*, pp. 149–168. Springer, 2024.

Linrui Zhang, Zhenghao Peng, Quanyi Li, and Bolei Zhou. Cat: Closed-loop adversarial training for safe end-to-end driving. In *Conference on Robot Learning*, pp. 2357–2372. PMLR, 2023.

Ding Zhao, Henry Lam, Huei Peng, Shan Bao, David J LeBlanc, Kazutoshi Nobukawa, and Christopher S Pan. Accelerated evaluation of automated vehicles safety in lane-change scenarios based on importance sampling techniques. *IEEE transactions on intelligent transportation systems*, 18(3): 595–607, 2016.

APPENDIX

Appendix Contents

- Appendix A: Policy Iteration Foundation of Co-SPAR
- Appendix B: Robust Feasible Region Spectrum and Safety Certificate Transfer
- Appendix C: Implementation Details
- Appendix D: Additional Experimental Results
- Appendix E: Related Work

A POLICY ITERATION FOUNDATION OF CO-SPAR

This appendix develops the policy iteration foundation of Co-SPAR for a protagonist facing adversaries in $\mathfrak{M}_\beta$. The pairwise task value $V^{\pi,\mu}$ and pairwise CDF $F^{\pi,\mu}$ are the basic task and feasibility objects. The adversary affects Co-SPAR through feasibility evaluation, while reward optimization remains protagonist-side. Throughout this analysis, we assume finite state, protagonist-action, and adversary-action spaces and stationary Markov policies. The relevant protagonist and adversary policy classes are nonempty, their pointwise extrema are attained, and $\mathfrak{M}_\beta$ is s-rectangular (Wiesemann et al., 2013), consistent with its state-wise definition in Section 3. The robust CDF F^π_β is obtained by maximizing the pairwise CDF over this class, and the optimal CDF F^*_β is obtained by minimizing F^π_β across protagonist policies. At iteration k, V^π_k denotes the task value under the adversary associated with the current policy π_k. Sections 3.1 and 3.2 introduce the adversary and protagonist updates, respectively. In the policy iteration analysis, π_k and π_{k+1} denote the policies before and after the update at iteration k, respectively.

A.1 CO-SPAR AS POLICY ITERATION OVER FEASIBLE REGIONS

At the policy iteration level, the safety requirement retains the pairwise CDF formulation from Section 2:

$$F^{\pi,\mu}(s) = 0, \qquad \forall \mu \in \mathfrak{M}_\beta,\ \forall s \in \mathcal{S}_{\text{init}}, \tag{20}$$

where $\mathcal{S}_{\text{init}} := \text{supp}(d)$. The protagonist task value is the pairwise value $V^{\pi,\mu}$ defined in Section 2. Adversaries in $\mathfrak{M}_\beta$ enter the policy iteration formulation through this safety requirement and through the adversarial context used for task improvement within the current region. Equivalently, the safety constraint is represented by the robust CDF

$$F^\pi_\beta(s) := \max_{\mu \in \mathfrak{M}_\beta} F^{\pi,\mu}(s). \tag{21}$$

The feasible region is therefore not given a priori. It is induced by the current protagonist policy through this maximization.

This structure makes policy improvement inherently region-wise. If the current policy π_k induces the robust CDF $F_k := F^{\pi_k}_\beta$ and strict feasible region

$$\mathcal{S}_k := \mathcal{S}^{\pi_k}_\beta = \{s \in \mathcal{S} \mid F^{\pi_k}_\beta(s) = 0\}, \tag{22}$$

then the state space splits into two regions. The adversary evaluates the current policy before policy improvement applies different objectives across the two regions. For fixed π_k, adversarial feasibility evaluation solves

$$\max_{\mu \in \mathfrak{M}_\beta} F^{\pi_k,\mu}(s), \qquad \forall s \in \mathcal{S}. \tag{23}$$

Its value is the current robust CDF

$$F_k(s) = F^{\pi_k}_\beta(s) = \max_{\mu \in \mathfrak{M}_\beta} F^{\pi_k,\mu}(s), \qquad \forall s \in \mathcal{S}. \tag{24}$$

Given this robust CDF, region-wise policy improvement separates across regions:

$$\min_\pi F^\pi_\beta(s), \qquad \forall s \notin \mathcal{S}_k, \tag{25}$$

$$\max_{\pi} V_k^{\pi}(s) \quad \text{s.t.} \quad F_\beta^{\pi}(s) = 0, \qquad \forall s \in \mathcal{S}_k. \tag{26}$$

Within $\mathcal{S}_k$, the protagonist optimizes task return subject to the current feasibility constraint under a maximizing adversary for F_k. Outside $\mathcal{S}_k$, the protagonist should reduce the current robust CDF so that the feasible region can expand. The operator definitions that make this split precise are introduced in the remainder of this appendix. The practical algorithm is described in Appendix C.

A.2 CDFs and Risky Operators

This section formalizes β-adversarial feasibility evaluation in Co-SPAR. We first distinguish pairwise, robust, and optimal CDFs and then define the risky operators whose fixed points characterize them. For brevity, policy and adversary optimizations are written with $\max / \min$. We identify CDFs and task value functions with vectors in $\mathbb{R}^{|\mathcal{S}|}$ equipped with the sup norm.

A.2.1 Pairwise, robust, and optimal CDFs

Definition 1 (Pairwise, robust, and optimal CDFs)**.** *For a fixed protagonist policy π and adversary policy μ, the pairwise CDF is*

$$F^{\pi,\mu}(s) := \mathbb{E}_{\pi,\mu}\left[\gamma^{N(\tau)} \mid s_0 = s\right]. \tag{27}$$

For a fixed protagonist policy π, the robust CDF is

$$F_\beta^{\pi}(s) := \max_{\mu \in \mathfrak{M}_\beta} F^{\pi,\mu}(s). \tag{28}$$

The optimal CDF is

$$F_\beta^{*}(s) := \min_{\pi} F_\beta^{\pi}(s). \tag{29}$$

These three objects characterize safety under a fixed protagonist–adversary pair, for a protagonist against the entire class $\mathfrak{M}_\beta$, and after optimization over protagonist policies, respectively.

Remark 1 (Stationary worst adversary in $\mathfrak{M}_\beta$)**.** *Because $\mathfrak{M}_\beta$ is s-rectangular, the state-wise maximizers in $\mathcal{R}_\beta^{\pi}$ can be selected independently across states. Under the finite-space and attainment assumptions above, the maximum is attained by a stationary adversary policy in $\mathfrak{M}_\beta$:*

$$F_\beta^{\pi} = \max_{\mu \in \mathfrak{M}_\beta} F^{\pi,\mu}. \tag{30}$$

A.2.2 Risky self-consistency and Bellman operators

Definition 2 (Risky self-consistency and Bellman operators)**.** *The risky self-consistency operator of a protagonist-adversary policy pair is*

$$(\mathcal{R}^{\pi,\mu}F)(s) := c(s) + (1 - c(s))\gamma \mathbb{E}_{a\sim\pi(\cdot|s)}\mathbb{E}_{u\sim\mu(\cdot|s)}\mathbb{E}_{s'\sim P(\cdot|s,a,u)}[F(s')]. \tag{31}$$

The robust risky self-consistency operator of a protagonist policy takes the worst pairwise continuation over $\mathfrak{M}_\beta(s)$:

$$(\mathcal{R}_\beta^{\pi}F)(s) := c(s) + (1 - c(s))\gamma \max_{\mu \in \mathfrak{M}_\beta(s)} \mathbb{E}_{u\sim\mu(\cdot|s)}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathbb{E}_{s'\sim P(\cdot|s,a,u)}[F(s')], \tag{32}$$

and the robust risky Bellman operator is

$$(\mathcal{R}_\beta^{*}F)(s) := \min_{\pi}(\mathcal{R}_\beta^{\pi}F)(s). \tag{33}$$

The two self-consistency operators evaluate a fixed protagonist–adversary pair and a fixed protagonist against the adversary class $\mathfrak{M}_\beta$, respectively, while the robust risky Bellman operator introduces protagonist optimization.

A.2.3 MONOTONICITY AND CONTRACTION

The next proposition collects the common dynamic-programming properties shared by all three operator levels.

Proposition 3 (Monotonicity and contraction of the risky operators)**.** *For any bounded functions $F_1, F_2 : \mathcal{S} \to \mathbb{R}$ with $F_1 \le F_2$ pointwise, the three risky operators are monotone:*

$$\mathcal{R}^{\pi,\mu}F_1 \le \mathcal{R}^{\pi,\mu}F_2, \qquad \mathcal{R}^{\pi}_{\beta}F_1 \le \mathcal{R}^{\pi}_{\beta}F_2, \qquad \mathcal{R}^{*}_{\beta}F_1 \le \mathcal{R}^{*}_{\beta}F_2. \tag{34}$$

Moreover, each operator is a γ-contraction in the sup norm:

$$\|\mathcal{R}^{\pi,\mu}F_1 - \mathcal{R}^{\pi,\mu}F_2\|_\infty \le \gamma\|F_1 - F_2\|_\infty, \tag{35}$$

$$\|\mathcal{R}^{\pi}_{\beta}F_1 - \mathcal{R}^{\pi}_{\beta}F_2\|_\infty \le \gamma\|F_1 - F_2\|_\infty, \tag{36}$$

$$\|\mathcal{R}^{*}_{\beta}F_1 - \mathcal{R}^{*}_{\beta}F_2\|_\infty \le \gamma\|F_1 - F_2\|_\infty. \tag{37}$$

Proof. Monotonicity of the risky self-consistency operator follows from monotonicity of expectation. The robust risky self-consistency operator preserves this order because pointwise maximization preserves inequalities, and the robust risky Bellman operator preserves it because pointwise minimization does the same. For the contraction bounds, the collision term cancels and the continuation term is multiplied by γ. For any two indexed families, $|\max_i x_i - \max_i y_i|$ and $|\min_i x_i - \min_i y_i|$ are bounded by $\sup_i |x_i - y_i|$, so taking a maximum or minimum over policies cannot increase the corresponding sup-norm difference. Thus all three operators are γ-contractions. □

A.2.4 FIXED POINTS OF THE RISKY OPERATORS

Proposition 4 (Fixed points of the risky operators)**.** *Under the finite-space, nonemptiness, and attainment assumptions above, the pairwise, robust, and optimal CDFs are the unique fixed points of their corresponding risky operators:*

$$F^{\pi,\mu} = \mathcal{R}^{\pi,\mu}F^{\pi,\mu}, \qquad F^{\pi}_{\beta} = \mathcal{R}^{\pi}_{\beta}F^{\pi}_{\beta}, \qquad F^{*}_{\beta} = \mathcal{R}^{*}_{\beta}F^{*}_{\beta}. \tag{38}$$

Moreover, the robust and optimal CDFs satisfy

$$F^{\pi}_{\beta} = \max_{\mu\in\mathfrak{M}_\beta} F^{\pi,\mu}, \qquad F^{*}_{\beta} = \min_{\pi} F^{\pi}_{\beta}. \tag{39}$$

Proof. By Proposition 3 and the Banach fixed-point theorem, let $G^{\pi,\mu}$, G^{π}_{β}, and G^{*}_{β} denote the unique fixed points of the three operators, respectively. The first-violation-time decomposition gives

$$F^{\pi,\mu} = \mathcal{R}^{\pi,\mu}F^{\pi,\mu}. \tag{40}$$

Uniqueness of the pairwise operator fixed point therefore gives $F^{\pi,\mu} = G^{\pi,\mu}$.

Fix a protagonist policy π. For every $\mu \in \mathfrak{M}_\beta$, pointwise maximization gives

$$\mathcal{R}^{\pi,\mu}G^{\pi}_{\beta} \le \mathcal{R}^{\pi}_{\beta}G^{\pi}_{\beta} = G^{\pi}_{\beta}. \tag{41}$$

Iterating $\mathcal{R}^{\pi,\mu}$ from the nonnegative zero function and using monotonicity gives

$$F^{\pi,\mu} \le G^{\pi}_{\beta}. \tag{42}$$

Let μ^{π}_{β} be an adversary policy attaining the maximum in $\mathcal{R}^{\pi}_{\beta}G^{\pi}_{\beta}$. Then

$$\mathcal{R}^{\pi,\mu^{\pi}_{\beta}}G^{\pi}_{\beta} = \mathcal{R}^{\pi}_{\beta}G^{\pi}_{\beta} = G^{\pi}_{\beta}. \tag{43}$$

Uniqueness of the pairwise operator fixed point gives $G^{\pi}_{\beta} = F^{\pi,\mu^{\pi}_{\beta}}$. Together with the preceding upper bound, this yields

$$G^{\pi}_{\beta} = \max_{\mu\in\mathfrak{M}_\beta} F^{\pi,\mu} = F^{\pi}_{\beta}. \tag{44}$$

For the protagonist minimization, the fixed point G^{*}_{β} satisfies

$$G^{*}_{\beta} = \mathcal{R}^{*}_{\beta}G^{*}_{\beta} \le \mathcal{R}^{\pi}_{\beta}G^{*}_{\beta}, \qquad \forall\pi. \tag{45}$$

Iterating this inequality under the monotone contraction $\mathcal{R}^{\pi}_{\beta}$ and taking the limit gives $G^{*}_{\beta} \le F^{\pi}_{\beta}$ for every π. Let $\pi^{\star}$ be a protagonist policy attaining the minimum in $\mathcal{R}^{*}_{\beta} G^{*}_{\beta}$. Then $\mathcal{R}^{\pi^{\star}}_{\beta} G^{*}_{\beta} = G^{*}_{\beta}$, so uniqueness gives $G^{*}_{\beta} = F^{\pi^{\star}}_{\beta}$. Consequently,

$$G^{*}_{\beta} = \min_{\pi} F^{\pi}_{\beta} = F^{*}_{\beta}. \tag{46}$$

The stated fixed-point equations follow from these identifications. □

The following comparison principle transfers a one-step supersolution inequality to the corresponding robust CDF fixed point.

Lemma 1 (Comparison bound for the robust risky self-consistency operator)**.** *Fix a protagonist policy π and let $H : \mathcal{S} \to \mathbb{R}$ be a test function. If*

$$\mathcal{R}^{\pi}_{\beta} H \le H \tag{47}$$

pointwise, then

$$F^{\pi}_{\beta} \le H. \tag{48}$$

Proof. By induction, using the pointwise monotonicity in Proposition 3, the assumed inequality gives

$$\left(\mathcal{R}^{\pi}_{\beta}\right)^{m} H \le H, \qquad \forall m \ge 1. \tag{49}$$

By Propositions 3 and 4, these iterates converge in the sup norm to the unique fixed point F^{π}_{β}. Taking the limit preserves the pointwise inequality and yields $F^{\pi}_{\beta} \le H$. □

A.3 Pairwise, Robust, and Optimal Feasible Regions

This section translates the pairwise, robust, and optimal CDFs into policy-dependent feasible regions and identifies the safety certificates used in Section 4.

A.3.1 Feasible regions induced by the three CDFs

Definition 3 (Pairwise, robust, and optimal feasible regions)**.** *For any threshold $\varepsilon \ge 0$, the pairwise CDF induces the pairwise ε-feasible and strict feasible regions*

$$\mathcal{S}^{\pi,\mu}_{\varepsilon} := \{s \in \mathcal{S} \mid F^{\pi,\mu}(s) \le \varepsilon\}, \qquad \mathcal{S}^{\pi,\mu} := \mathcal{S}^{\pi,\mu}_{0}. \tag{50}$$

The robust CDF induces the robust ε-feasible and strict feasible regions

$$\mathcal{S}^{\pi}_{\beta,\varepsilon} := \{s \in \mathcal{S} \mid F^{\pi}_{\beta}(s) \le \varepsilon\}, \qquad \mathcal{S}^{\pi}_{\beta} := \mathcal{S}^{\pi}_{\beta,0}, \tag{51}$$

and the optimal CDF induces the optimal ε-feasible and strict feasible regions

$$\mathcal{S}^{*}_{\beta,\varepsilon} := \{s \in \mathcal{S} \mid F^{*}_{\beta}(s) \le \varepsilon\}, \qquad \mathcal{S}^{*}_{\beta} := \mathcal{S}^{*}_{\beta,0}. \tag{52}$$

The policy class attaining the optimal CDF at strength β is

$$\Pi^{\max}_{\beta} := \{\pi \mid F^{\pi}_{\beta} = F^{*}_{\beta}\}. \tag{53}$$

The main text uses $\mathcal{S}^{\pi}_{\beta,\varepsilon}$ for robust ε-feasible regions and $\mathcal{S}^{\pi}_{\beta}$ for the robust strict feasible region that evolves along the protagonist policy sequence. By construction, for every protagonist policy π and threshold $\varepsilon \ge 0$,

$$F^{*}_{\beta} \le F^{\pi}_{\beta}, \qquad \mathcal{S}^{\pi}_{\beta,\varepsilon} \subseteq \mathcal{S}^{*}_{\beta,\varepsilon}. \tag{54}$$

Thus, at a fixed adversarial strength, the optimal ε-feasible region contains the robust ε-feasible region induced by any protagonist policy; at zero level, this gives $\mathcal{S}^{\pi}_{\beta} \subseteq \mathcal{S}^{*}_{\beta}$.

A.3.2 Safety certificates

Membership in a robust ε-feasible region yields the finite-horizon probability bound stated in Section 4. The zero-level case then gives the strict feasible region semantics used by policy iteration.

Safety certificate proof. This proves Proposition 1 stated in Section 4.

Proof. If $s \in \mathcal{S}^{\pi}_{\beta,\varepsilon}$, then $F^{\pi}_{\beta}(s) \le \varepsilon$. By definition,

$$F^{\pi}_{\beta}(s) = \max_{\mu \in \mathfrak{M}_{\beta}} \mathbb{E}_{\pi,\mu}\Big[\gamma^{N(\tau)} \mid s_0 = s\Big]. \tag{55}$$

Fix any adversary policy $\mu \in \mathfrak{M}_{\beta}$ and any horizon $H \ge 0$. The event $\{N(\tau) \le H\}$ implies $\gamma^{N(\tau)} \ge \gamma^{H}$, because $\gamma \in (0,1)$. Therefore,

$$\mathbb{E}_{\pi,\mu}\Big[\gamma^{N(\tau)} \mid s_0 = s\Big] \ge \gamma^{H} \mathbb{P}^{\pi,\mu}_{s}\big(N(\tau) \le H\big). \tag{56}$$

Since

$$F^{\pi}_{\beta}(s) \ge \mathbb{E}_{\pi,\mu}\Big[\gamma^{N(\tau)} \mid s_0 = s\Big], \tag{57}$$

we obtain

$$\mathbb{P}^{\pi,\mu}_{s}\big(N(\tau) \le H\big) \le \frac{F^{\pi}_{\beta}(s)}{\gamma^{H}} \le \varepsilon/\gamma^{H}. \tag{58}$$

Since μ was arbitrary, the bound holds uniformly over all adversaries in $\mathfrak{M}_{\beta}$. □

Strict safety at the zero level. The zero-level specialization of Proposition 1 gives the following corollary.

Corollary 1 (Infinite horizon safety at the zero level)**.** *If $s \in \mathcal{S}^{\pi}_{\beta}$, then for every adversary policy $\mu \in \mathfrak{M}_{\beta}$,*

$$\mathbb{P}^{\pi,\mu}_{s}(N(\tau) < \infty) = 0. \tag{59}$$

Proof. If $s \in \mathcal{S}^{\pi}_{\beta}$, then $F^{\pi}_{\beta}(s) = 0$, so Proposition 1 with $\varepsilon = 0$ yields

$$\max_{\mu \in \mathfrak{M}_{\beta}} \mathbb{P}^{\pi,\mu}_{s}\big(N(\tau) \le H\big) = 0, \qquad \forall H \ge 0. \tag{60}$$

Hence, for every adversary policy $\mu \in \mathfrak{M}_{\beta}$,

$$\mathbb{P}^{\pi,\mu}_{s}\big(N(\tau) \le H\big) = 0, \qquad \forall H \ge 0. \tag{61}$$

Because

$$\{N(\tau) < \infty\} = \bigcup_{H=0}^{\infty} \{N(\tau) \le H\}, \tag{62}$$

countable subadditivity implies

$$\mathbb{P}^{\pi,\mu}_{s}(N(\tau) < \infty) = 0. \tag{63}$$

Thus every adversary in $\mathfrak{M}_{\beta}$ assigns zero probability to ever reaching the unsafe set. □

A.4 Operator Analysis of the Region-wise Protagonist Optimization

This section provides an operator analysis of the region-wise protagonist optimization as a policy improvement step with separate objectives inside and outside the current feasible region. The β-adversarial feasibility evaluation first induces the current robust CDF by maximizing over $\mathfrak{M}_{\beta}$. Policy improvement then uses that CDF to improve reward inside the current feasible region and reduce the CDF outside it. The definitions below provide the two ingredients underlying the fixed-β theoretical results: safe actions preserve the current region, and the task operator under μ^{π}_{β} supports policy improvement. The update characterization and its two monotonicity consequences are developed below.

A.4.1 Safe action set and constrained policy class

Definition 4 (Safe action set and constrained policy class)**.** *The protagonist constraint in Section 3 is statewise, but the update itself acts through actions. Given a protagonist policy π and its robust CDF F^π_β, the safe action set is*

$$\mathcal{A}^\pi_\beta(s) := \left\{ a \in \mathcal{A} \;\middle|\; \max_{\mu \in \mathfrak{M}_\beta(s)} \mathbb{E}_{u\sim\mu(\cdot|s)} \mathbb{E}_{s'\sim P(\cdot|s,a,u)} \left[F^\pi_\beta(s') \right] = 0 \right\}. \tag{64}$$

Thus $\mathcal{A}^\pi_\beta(s)$ collects actions whose worst one-step continuation over $\mathfrak{M}_\beta(s)$ preserves the current strict feasible region. The corresponding constrained policy class is

$$\Pi^\pi_\beta := \left\{ \tilde{\pi} \;\middle|\; \tilde{\pi}(a \mid s) = 0 \text{ if } a \notin \mathcal{A}^\pi_\beta(s),\ \forall s \in \mathcal{S}^\pi_\beta \right\}. \tag{65}$$

Once the robust CDF identifies where safety has already been secured, Π^π_β specifies which protagonist policies preserve that guarantee. For the policy iteration analysis, we assume $\mathcal{A}^\pi_\beta(s)$ is nonempty on $\mathcal{S}^\pi_\beta$ and that the current policy belongs to Π^π_β on this region. Because F^π_β is nonnegative, the defining equality for any $a \in \mathcal{A}^\pi_\beta(s)$ implies $F^\pi_\beta(s') = 0$ almost surely under every one-step continuation induced by a distribution in $\mathfrak{M}_\beta(s)$. Thus every policy in Π^π_β keeps the current strict feasible region forward invariant, and the restricted task operators are well-defined on $\mathcal{S}^\pi_\beta$.

A.4.2 Restricted Task Operators

Definition 5 (Worst adversary and restricted task operators)**.** *For a current protagonist policy π, let $\mu^\pi_\beta \in \mathfrak{M}_\beta$ be a stationary adversary policy satisfying, for each $s \in \mathcal{S}$,*

$$\mu^\pi_\beta(\cdot \mid s) \in \arg\max_{\mu \in \mathfrak{M}_\beta(s)} \mathbb{E}_{u\sim\mu(\cdot|s)} \mathbb{E}_{a\sim\pi(\cdot|s)} \mathbb{E}_{s'\sim P(\cdot|s,a,u)} [F^\pi_\beta(s')]. \tag{66}$$

For a candidate policy $\tilde{\pi} \in \Pi^\pi_\beta$, define the task operator on the current region

$$(\mathcal{T}^{\tilde{\pi}|\pi}_\beta V)(s) := \mathbb{E}_{a\sim\tilde{\pi}(\cdot|s)} \mathbb{E}_{u\sim\mu^\pi_\beta(\cdot|s)} \left[r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a,u)} [V(s')] \right]. \tag{67}$$

The associated region-wise optimal task operator is

$$(\mathcal{T}^{*|\pi}_\beta V)(s) := \begin{cases} \max_{\tilde{\pi}\in\Pi^\pi_\beta} (\mathcal{T}^{\tilde{\pi}|\pi}_\beta V)(s), & s \in \mathcal{S}^\pi_\beta, \\ V(s), & s \notin \mathcal{S}^\pi_\beta. \end{cases} \tag{68}$$

Let $V^{\tilde{\pi}|\pi}_\beta$ and $V^{*|\pi}_\beta$ denote the fixed points of the corresponding operators restricted to the current feasible region $\mathcal{S}^\pi_\beta$.

Over the current feasible region, the restricted task operator therefore conditions on the same adversary that defines F^π_β. This is the operator object behind the reward-improvement statement on the current region in Section 4.

A.4.3 Monotonicity and fixed points of restricted task operators

The restricted task operators inherit the same dynamic-programming structure as the standard Bellman operator once μ^π_β is fixed.

Proposition 5 (Monotonicity and contraction of restricted task operators)**.** *Fix a protagonist policy π and a worst adversary $\mu^\pi_\beta \in \mathfrak{M}_\beta$. For every candidate policy $\tilde{\pi} \in \Pi^\pi_\beta$ and any bounded functions V_1, V_2 on $\mathcal{S}^\pi_\beta$ with $V_1 \le V_2$ pointwise, we have*

$$\mathcal{T}^{\tilde{\pi}|\pi}_\beta V_1 \le \mathcal{T}^{\tilde{\pi}|\pi}_\beta V_2, \qquad \mathcal{T}^{*|\pi}_\beta V_1 \le \mathcal{T}^{*|\pi}_\beta V_2. \tag{69}$$

Moreover,

$$\|\mathcal{T}^{\tilde{\pi}|\pi}_\beta V_1 - \mathcal{T}^{\tilde{\pi}|\pi}_\beta V_2\|_{\infty,\mathcal{S}^\pi_\beta} \le \gamma \|V_1 - V_2\|_{\infty,\mathcal{S}^\pi_\beta}, \tag{70}$$

$$\|\mathcal{T}^{*|\pi}_\beta V_1 - \mathcal{T}^{*|\pi}_\beta V_2\|_{\infty,\mathcal{S}^\pi_\beta} \le \gamma \|V_1 - V_2\|_{\infty,\mathcal{S}^\pi_\beta}. \tag{71}$$

Hence both restricted operators admit unique fixed points on $\mathcal{S}_\beta^\pi$, namely $V_\beta^{\tilde{\pi}|\pi}$ and $V_\beta^{*|\pi}$.

Proof. For a fixed candidate policy, the safe-action condition keeps all successor states in $\mathcal{S}_\beta^\pi$. The reward term cancels when two operator values are subtracted, and therefore, for every $s \in \mathcal{S}_\beta^\pi$,

$$\left|(\mathcal{T}_\beta^{\tilde{\pi}|\pi} V_1)(s) - (\mathcal{T}_\beta^{\tilde{\pi}|\pi} V_2)(s)\right| \le \gamma \|V_1 - V_2\|_{\infty, \mathcal{S}_\beta^\pi}. \tag{72}$$

Pointwise order preservation gives monotonicity for a fixed candidate policy. For the region-wise optimal task operator, pointwise maximization preserves monotonicity and satisfies

$$\left|\max_{\tilde{\pi} \in \Pi_\beta^\pi} x_{\tilde{\pi}} - \max_{\tilde{\pi} \in \Pi_\beta^\pi} y_{\tilde{\pi}}\right| \le \sup_{\tilde{\pi} \in \Pi_\beta^\pi} |x_{\tilde{\pi}} - y_{\tilde{\pi}}|, \tag{73}$$

which gives the same contraction bound on $\mathcal{S}_\beta^\pi$. The Banach fixed-point theorem then gives the two unique fixed points on the current region. □

A.4.4 Operator View of Co-SPAR Policy Iteration

With the restricted task operator and its fixed-point properties in place, we next state the complete region-wise policy improvement step. Co-SPAR policy iteration evolves a protagonist policy sequence, while the adversary enters each round through β-adversarial feasibility evaluation of the current protagonist policy. Under the assumptions above, Co-SPAR policy iteration at fixed β is summarized by the following operator view.

Proposition 6 (Operator view of Co-SPAR policy iteration)**.** *Fix a current protagonist policy π_k and let π_{k+1} denote the policy iteration update. The current robust CDF satisfies*

$$F_k = F_\beta^{\pi_k}, \qquad F_k = \mathcal{R}_\beta^{\pi_k} F_k, \tag{74}$$

which induces the strict feasible region

$$\mathcal{S}_k := \mathcal{S}_\beta^{\pi_k} = \{s \in \mathcal{S} \mid F_k(s) = 0\}. \tag{75}$$

The corresponding adversary realizes the worst continuation over $\mathfrak{M}_\beta$ encoded by $\mathcal{R}_\beta^{\pi_k}$.
Given this robust CDF, region-wise policy improvement splits by region. Outside $\mathcal{S}_k$, the protagonist update satisfies

$$\mathcal{R}_\beta^{\pi_{k+1}} F_k = \mathcal{R}_\beta^* F_k \le \mathcal{R}_\beta^{\pi_k} F_k = F_k, \qquad \forall s \notin \mathcal{S}_k. \tag{76}$$

Inside $\mathcal{S}_k$, the protagonist update is supported on the safe action set and satisfies the task improvement equation under $\mu_\beta^{\pi_k}$:

$$\pi_{k+1}(a \mid s) = 0 \quad \text{if } a \notin \mathcal{A}_\beta^{\pi_k}(s), \qquad \forall s \in \mathcal{S}_k, \tag{77}$$

$$\mathcal{T}_\beta^{\pi_{k+1}|\pi_k} V_\beta^{\pi_k|\pi_k} = \mathcal{T}_\beta^{*|\pi_k} V_\beta^{\pi_k|\pi_k}, \qquad \forall s \in \mathcal{S}_k. \tag{78}$$

Hence Co-SPAR policy iteration at fixed β maps π_k to π_{k+1} through the chain

$$\pi_k \;\longrightarrow\; F_\beta^{\pi_k} \;\longrightarrow\; \mathcal{S}_\beta^{\pi_k} \;\longrightarrow\; \left(\mu_\beta^{\pi_k}, \mathcal{T}_\beta^{*|\pi_k}\right) \;\longrightarrow\; \pi_{k+1}. \tag{79}$$

Proof. The fixed-point equation for F_k follows from Proposition 4. Outside the current region, the protagonist update selects a state-wise minimizer of the robust risky Bellman operator, and the current policy is one admissible candidate, which gives the stated inequality. Inside the current region, the protagonist update selects a policy from Π_β^π that maximizes the restricted task operator evaluated at the current policy value, giving the two stated conditions. □

This proposition is the operator characterization used by the proofs in Section 4: the adversary determines the safety envelope, and region-wise policy improvement reduces the robust CDF outside the current region and performs constrained task improvement within it.

A.4.5 MONOTONE EXPANSION OF THE ROBUST FEASIBLE REGION.

This subsection proves Theorem 1. The robust CDF sequence is nonincreasing, and the induced strict feasible region expands monotonically.

Proof. Let $F_k = F_\beta^{\pi_k}$ and $\mathcal{S}_k = \{s \mid F_k(s) = 0\}$. By Lemma 1, applied to π_{k+1} and $H = F_k$, it suffices to establish

$$\mathcal{R}_\beta^{\pi_{k+1}} F_k \le F_k. \tag{80}$$

First consider a state $s \in \mathcal{S}_k$. Proposition 6 ensures that every action in the support of $\pi_{k+1}(\cdot \mid s)$ belongs to $\mathcal{A}_\beta^{\pi_k}(s)$. Since $F_k(s) = 0$ and $F_k = \mathcal{R}_\beta^{\pi_k} F_k$ with a nonnegative continuation term, we also have $s \in \mathcal{C}$. Therefore, for every action a in the support of $\pi_{k+1}(\cdot \mid s)$,

$$\max_{\mu \in \mathfrak{M}_\beta(s)} \mathbb{E}_{u\sim\mu(\cdot|s)} \mathbb{E}_{s'\sim P(\cdot|s,a,u)} \left[F_k(s')\right] = 0. \tag{81}$$

By definition of $\mathcal{A}_\beta^{\pi_k}(s)$, this implies

$$(\mathcal{R}_\beta^{\pi_{k+1}} F_k)(s) = 0 = F_k(s). \tag{82}$$

Now consider $s \notin \mathcal{S}_k$. By the policy iteration update in Proposition 6,

$$(\mathcal{R}_\beta^{\pi_{k+1}} F_k)(s) = (\mathcal{R}_\beta^* F_k)(s) \le (\mathcal{R}_\beta^{\pi_k} F_k)(s) = F_k(s). \tag{83}$$

The two cases establish the required inequality. Applying Lemma 1 gives

$$F_{k+1} \le F_k. \tag{84}$$

For every $s \in \mathcal{S}_k$, CDF nonnegativity now gives

$$0 \le F_{k+1}(s) \le F_k(s) = 0. \tag{85}$$

Therefore

$$\mathcal{S}_k \subseteq \mathcal{S}_{k+1}. \tag{86}$$

This proves that the feasible region expands monotonically along the protagonist policy sequence generated by Co-SPAR policy iteration. □

The same operator relations also identify the limit of the robust CDF sequence.

Corollary 2 (Convergence to the optimal CDF)**.** *Fix an adversarial strength β, let $\{\pi_k\}_{k\ge0}$ be generated by the Co-SPAR policy iteration in Proposition 6, and let $F_k := F_\beta^{\pi_k}$. Then*

$$\mathcal{R}_\beta^{\pi_{k+1}} F_k = \mathcal{R}_\beta^* F_k, \tag{87}$$

and

$$F_\beta^* \le F_{k+1} \le \mathcal{R}_\beta^* F_k \le F_k. \tag{88}$$

Moreover,

$$\|F_{k+1} - F_\beta^*\|_\infty \le \gamma \|F_k - F_\beta^*\|_\infty, \tag{89}$$

so F_k converges uniformly to F_β^*.

Proof. Outside $\mathcal{S}_k$, Proposition 6 gives

$$(\mathcal{R}_\beta^{\pi_{k+1}} F_k)(s) = (\mathcal{R}_\beta^* F_k)(s). \tag{90}$$

For $s \in \mathcal{S}_k$, Equation 82 gives

$$(\mathcal{R}_\beta^{\pi_{k+1}} F_k)(s) = 0. \tag{91}$$

The robust risky Bellman operator is nonnegative on F_k, and the current policy π_k is a candidate in its protagonist minimization, so

$$0 \le (\mathcal{R}_\beta^* F_k)(s) \le (\mathcal{R}_\beta^{\pi_k} F_k)(s) = F_k(s) = 0. \tag{92}$$

Thus $\mathcal{R}_\beta^{\pi_{k+1}} F_k = \mathcal{R}_\beta^* F_k$ everywhere. Because the current policy is also a candidate in the minimization, this common function satisfies

$$\mathcal{R}_\beta^* F_k \le \mathcal{R}_\beta^{\pi_k} F_k = F_k. \tag{93}$$

Let $H := \mathcal{R}_\beta^* F_k$. Monotonicity gives

$$\mathcal{R}_\beta^{\pi_{k+1}} H \le \mathcal{R}_\beta^{\pi_{k+1}} F_k = H. \tag{94}$$

Lemma 1 therefore yields

$$F_{k+1} \le H = \mathcal{R}_\beta^* F_k. \tag{95}$$

By definition, $F_\beta^* \le F_{k+1}$. Using the fixed-point equation $F_\beta^* = \mathcal{R}_\beta^* F_\beta^*$ from Proposition 4, we obtain

$$0 \le F_{k+1} - F_\beta^* \le \mathcal{R}_\beta^* F_k - \mathcal{R}_\beta^* F_\beta^*. \tag{96}$$

The contraction property in Proposition 3 then gives

$$\|F_{k+1} - F_\beta^*\|_\infty \le \gamma \|F_k - F_\beta^*\|_\infty. \tag{97}$$

Iterating this bound proves uniform convergence to F_β^*. □

A.4.6 Monotone task improvement within the current feasible region.

We now establish task improvement on the current feasible region $\mathcal{S}_k$.

Theorem 3 (Monotone task improvement within the robust feasible region.). *Let $\mu_\beta^{\pi_k} \in \mathfrak{M}_\beta$ denote a stationary worst adversary associated with $F_\beta^{\pi_k}$, and let*

$$V_k^\pi(s) := \mathbb{E}_{\pi, \mu_\beta^{\pi_k}} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s \right]. \tag{98}$$

If π_{k+1} is obtained by the region-wise protagonist optimization in Proposition 6, then

$$V_k^{\pi_{k+1}}(s) \ge V_k^{\pi_k}(s), \qquad \forall s \in \mathcal{S}_k. \tag{99}$$

Proof. Fix an iteration k. Inside the current feasible region $\mathcal{S}_k$, the task objective is evaluated under the adversary $\mu_\beta^{\pi_k}$. This adversary also induces the current robust CDF $F_k = F_\beta^{\pi_k}$. The protagonist update therefore reduces to a standard discounted control problem with constrained policy class $\Pi_\beta^{\pi_k}$.

For any candidate policy $\pi \in \Pi_\beta^{\pi_k}$, let

$$V_k^\pi(s) := \mathbb{E}_{\pi, \mu_\beta^{\pi_k}} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s \right]. \tag{100}$$

On $\mathcal{S}_k$, this is exactly the value represented by the fixed point of the restricted task operator under $\mu_\beta^{\pi_k}$, so it is convenient to write

$$V_\beta^{\pi|\pi_k}(s) = V_k^\pi(s), \qquad \forall s \in \mathcal{S}_k. \tag{101}$$

The current policy belongs to Π_β^π on $\mathcal{S}_k$ by the safe-policy-class assumption above. The task improvement condition in Proposition 6 therefore gives, for every $s \in \mathcal{S}_k$,

$$\mathcal{T}_\beta^{\pi_{k+1}|\pi_k} V_\beta^{\pi_k|\pi_k}(s) \ge \mathcal{T}_\beta^{\pi_k|\pi_k} V_\beta^{\pi_k|\pi_k}(s). \tag{102}$$

Because $V_\beta^{\pi_k|\pi_k}$ is the fixed point of $\mathcal{T}_\beta^{\pi_k|\pi_k}$, the right-hand side equals

$$V_\beta^{\pi_k|\pi_k}(s). \tag{103}$$

Hence

$$\mathcal{T}_\beta^{\pi_{k+1}|\pi_k} V_\beta^{\pi_k|\pi_k}(s) \ge V_\beta^{\pi_k|\pi_k}(s), \qquad \forall s \in \mathcal{S}_k. \tag{104}$$

Now repeatedly apply the monotone contraction

$$\mathcal{T}_\beta^{\pi_{k+1}|\pi_k} \tag{105}$$

from Proposition 5 to both sides. Monotonicity gives the pointwise chain

$$V_\beta^{\pi_k|\pi_k} \le \mathcal{T}_\beta^{\pi_{k+1}|\pi_k} V_\beta^{\pi_k|\pi_k} \le \left(\mathcal{T}_\beta^{\pi_{k+1}|\pi_k}\right)^2 V_\beta^{\pi_k|\pi_k} \le \cdots. \tag{106}$$

Repeated application converges to the fixed point

$$V_\beta^{\pi_{k+1}|\pi_k}, \tag{107}$$

which therefore satisfies

$$V_\beta^{\pi_{k+1}|\pi_k}(s) \ge V_\beta^{\pi_k|\pi_k}(s), \qquad \forall s \in \mathcal{S}_k. \tag{108}$$

Replacing the operator shorthand by the value notation above gives

$$V_k^{\pi_{k+1}}(s) \ge V_k^{\pi_k}(s), \qquad \forall s \in \mathcal{S}_k. \tag{109}$$

This proves monotone task improvement within the current feasible region. □

B Robust Feasible Region Spectrum and Safety Certificate Transfer

The analysis in Appendix A holds adversarial strength β fixed. Here, nested adversary classes order robust CDFs and feasible regions across strengths. A protagonist optimal at a stronger strength also retains its safety certificates against weaker classes.

B.1 Robust feasible region spectrum

This subsection proves Proposition 2. Increasing the adversarial strength increases the robust and optimal CDFs and shrinks the corresponding robust and optimal ε-feasible regions.

Proof. Let $\beta_1 < \beta_2$. By nestedness,

$$\mathfrak{M}_{\beta_1}(s) \subseteq \mathfrak{M}_{\beta_2}(s) \qquad \forall s, \tag{110}$$

so for every candidate F and protagonist policy π,

$$\mathcal{R}_{\beta_1}^\pi F \le \mathcal{R}_{\beta_2}^\pi F. \tag{111}$$

Iterating the two contraction mappings from a common initialization and taking limits gives

$$F_{\beta_1}^\pi \le F_{\beta_2}^\pi. \tag{112}$$

Taking the minimum over protagonist policies yields

$$F_{\beta_1}^* \le F_{\beta_2}^*. \tag{113}$$

The region inclusions then follow immediately by taking sublevel sets of the two CDFs. □

B.2 Safety certificate transfer

This subsection proves Theorem 2. The proof places the robust CDF and robust feasible region of a protagonist optimal at the stronger strength between their optimal counterparts at the two strengths.

Proof. Let $\beta_1 < \beta_2$ and let $\pi_{\beta_2}^\star$ satisfy

$$F_{\beta_2}^{\pi_{\beta_2}^\star} = F_{\beta_2}^*. \tag{114}$$

By the nestedness $\mathfrak{M}_{\beta_1}(s) \subseteq \mathfrak{M}_{\beta_2}(s)$ and the definition of the robust risky self-consistency operator in Definition 2, for every bounded candidate F,

$$\mathcal{R}_{\beta_1}^{\pi_{\beta_2}^\star} F \le \mathcal{R}_{\beta_2}^{\pi_{\beta_2}^\star} F. \tag{115}$$

Applying this to $F^*_{\beta_2}$ gives

$$\mathcal{R}^{\pi^\star_{\beta_2}}_{\beta_1} F^*_{\beta_2} \le \mathcal{R}^{\pi^\star_{\beta_2}}_{\beta_2} F^*_{\beta_2} = F^*_{\beta_2}. \tag{116}$$

Thus $F^*_{\beta_2}$ is a supersolution for the robust risky self-consistency operator $\mathcal{R}^{\pi^\star_{\beta_2}}_{\beta_1}$. Iterating $\mathcal{R}^{\pi^\star_{\beta_2}}_{\beta_1}$ from this supersolution and using contraction convergence to the fixed point yields

$$F^{\pi^\star_{\beta_2}}_{\beta_1} \le F^*_{\beta_2}. \tag{117}$$

On the other hand, by definition of the optimal CDF at strength β_1,

$$F^*_{\beta_1} \le F^{\pi^\star_{\beta_2}}_{\beta_1}. \tag{118}$$

Combining the two inequalities gives the sandwich

$$F^*_{\beta_1} \le F^{\pi^\star_{\beta_2}}_{\beta_1} \le F^*_{\beta_2}. \tag{119}$$

The corresponding ε-feasible region inclusions follow immediately from taking sublevel sets:

$$\mathcal{S}^*_{\beta_2,\varepsilon} \subseteq \mathcal{S}^{\pi^\star_{\beta_2}}_{\beta_1,\varepsilon} \subseteq \mathcal{S}^*_{\beta_1,\varepsilon}. \tag{120}$$

This is exactly the claimed transfer relation across β. □

The transfer theorem can also be summarized by a certified adversarial-strength radius.

B.3 Certified adversarial-strength radius

Definition 6 (Statewise adversarial-strength radius)**.** *The nested feasible regions also induce a statewise summary of how much adversarial strength a policy can certify at a given state. For any protagonist policy π, define*

$$\rho^\pi(s) := \sup\{\beta \ge 0 \mid F^\pi_\beta(s) = 0\}. \tag{121}$$

This quantity is the largest adversarial-strength level for which state s remains in the strict feasible region of π.

It compresses the nested feasible regions across adversarial strengths into a single statewise quantity and is useful when comparing certified robustness across states.

The following remark applies this summary to a policy optimal at the stronger strength.

Remark 2 (Certified radius under policy transfer)**.** *Let $\beta_2 > 0$, let π satisfy $F^\pi_{\beta_2} = F^*_{\beta_2}$, and let $s \in \mathcal{S}^*_{\beta_2}$. Then*

$$\rho^\pi(s) \ge \beta_2. \tag{122}$$

Moreover, the transfer inclusion in Theorem 2 gives

$$s \in \mathcal{S}^\pi_{\beta_1}, \qquad \forall\, \beta_1 < \beta_2. \tag{123}$$

Thus, the stronger-strength optimum certifies s up to at least β_2, and the same certificate remains valid at every weaker adversarial strength.

Proof. Since π is optimal at β_2, $s \in \mathcal{S}^*_{\beta_2}$ implies $F^\pi_{\beta_2}(s) = 0$, and the definition of ρ^π gives $\rho^\pi(s) \ge \beta_2$. The second inclusion is Theorem 2 specialized to $\varepsilon = 0$. □

C Implementation Details

This appendix turns the operator view into trainable optimization problems and scalable approximations.

Algorithm 1 Constrained Self-Play with Adversarial Rationality (Co-SPAR)

1: **Input:** protagonist $\pi_{\theta_{\text{old}}}$, adversary $\mu_{\xi_{\text{old}}}$, critics V_ψ, F_{ϕ^π}, and F_{ϕ^μ}, behavioral prior μ_{prior}, budget β, batch $\mathcal{D}$.
2: Collect rollouts under $(\pi_{\theta_{\text{old}}}, \mu_{\xi_{\text{old}}})$ and compute reward, CDF, and continuation targets.
3: Update F_{ϕ^π} and F_{ϕ^μ} and form $\widehat{\mathcal{S}} = \{s \mid F_{\phi^\pi}(s) \le \epsilon\}$.
4: Estimate $A^{V,\pi}$, $A^{F,\pi}$, and $A^{F,\mu}$ by GAE, and build $\bar{A}_t^\pi$.
5: Update the protagonist by maximizing $L_\pi(\theta)$ in equation 11.
6: Update μ_ξ by maximizing $L_\mu(\xi)$ in equation 11.
7: Update the adversary KL multiplier by equation 12.
8: **Output:** updated protagonist π_θ, adversary μ_ξ, and critics $(V_\psi, F_{\phi^\pi}, F_{\phi^\mu})$.

C.1 Overall training loop

C.2 Region-Wise Protagonist Optimization

We approximate the region-wise protagonist optimization from Section 3.2 with the following surrogate. Let θ^π denote the protagonist parameters, let V_ψ be a task critic, and let F_{ϕ^π} be the protagonist-side CDF critic. Using the estimated feasible region $\widehat{\mathcal{S}} = \{s \mid F_{\phi^\pi}(s) \le \epsilon\}$, the actor is trained with the shared clipped surrogate equation 11, instantiated by the region-wise advantage below:

$$\bar{A}_t^\pi := \mathbf{1}[F_{\phi^\pi}(s_t) \le \epsilon] \left(A_t^{V,\pi} - w(s_t) A_t^{F,\pi}\right) - \mathbf{1}[F_{\phi^\pi}(s_t) > \epsilon]\, w_{\max} A_t^{F,\pi}. \tag{124}$$

For a rollout segment $(s_t, a_t, u_t, s_{t+1})_{t=0}^{T-1}$, let $c_t := c(s_t)$. We use the temporal-difference residuals

$$\delta_t^{V,\pi} := r(s_t, a_t) + \gamma V_\psi(s_{t+1}) - V_\psi(s_t), \tag{125}$$

$$\delta_t^{F,\pi} := c_t + (1 - c_t)\gamma F_{\phi^\pi}(s_{t+1}) - F_{\phi^\pi}(s_t), \tag{126}$$

and define

$$A_t^{V,\pi} := \sum_{l=0}^{T-t-1} (\gamma\lambda_V)^l \delta_{t+l}^{V,\pi}, \tag{127}$$

$$A_t^{F,\pi} := \sum_{l=0}^{T-t-1} (\gamma\lambda_F)^l \left(\prod_{j=0}^{l-1} (1 - c_{t+j})\right) \delta_{t+l}^{F,\pi}, \tag{128}$$

where the empty product for $l = 0$ is one. The trace parameters satisfy $\lambda_V, \lambda_F \in [0, 1]$. Thus $A_t^{V,\pi}$ is the standard task GAE (Schulman et al., 2017), while $A_t^{F,\pi}$ is the protagonist CDF GAE induced by the robust risky self-consistency equation and truncated after the first realized violation. The state-dependent weight $w(s) := C(1/(\alpha - r(s)) - 1/\alpha)$, where $r(s) := \min(F_{\phi^\pi}(s)/\epsilon, 1)$ and $C := w_{\max}\alpha(\alpha - 1)$, increases the safety penalty as the state approaches the estimated boundary while remaining bounded by $w_{\max}$. In practice, $\widehat{\mathcal{S}}$ is formed pointwise on the sampled batch using the current protagonist critic and a fixed tolerance ϵ. This surrogate applies the task objective within the estimated feasible region and the CDF-reduction objective outside it.

C.3 Prior-constrained Adversary Optimization

We implement the prior-constrained adversary optimization from Section 3.1 through a KL budget and its associated dual update. The adversary policy is parameterized as

$$\mu_\xi(u \mid s), \tag{129}$$

with an adversary-side CDF critic F_{ϕ^μ} that evaluates first-violation outcomes under the evolving adversarial policy. The resulting CDF advantage provides the adversary's ascent signal for increasing estimated violation risk. In practice, the adversary update uses a clipped surrogate objective with the adversary ratio ρ_μ and the CDF advantage $A^{F,\mu}$. Using the rollout notation above, we define

$$\delta_t^{F,\mu} := c_t + (1 - c_t)\gamma F_{\phi^\mu}(s_{t+1}) - F_{\phi^\mu}(s_t), \tag{130}$$

$$A_t^{F,\mu} := \sum_{l=0}^{T-t-1} (\gamma\lambda_F)^l \left(\prod_{j=0}^{l-1} (1 - c_{t+j}) \right) \delta_{t+l}^{F,\mu}. \tag{131}$$

Equivalently, the adversary objective in Section 3 can be written as

$$L_\mu(\xi) := L_{\text{clip}}^\mu - \eta_\beta D_{\text{KL}}(\mu_\xi(\cdot \mid s) \parallel \mu_{\text{prior}}(\cdot \mid s)). \tag{132}$$

The CDF advantage provides the adversary optimization signal, while the dual variable regulates the KL budget that defines $\mathfrak{M}_\beta$.

To preserve the interpretation of β as an adversarial-strength index, we treat β as a target KL budget and adapt the positive multiplier η_β for that constraint. This preserves the theoretical role of β as an index that expands $\mathfrak{M}_\beta$ without altering the safety object. Define

$$\widehat{\text{KL}}_\beta(\xi) := \mathbb{E}_{\text{batch}} \left[D_{\text{KL}}\big(\mu_\xi(\cdot \mid s) \,\big\|\, \mu_{\text{prior}}(\cdot \mid s)\big) \right]. \tag{133}$$

The multiplier follows Eq. 12, using $\widehat{\text{KL}}_\beta(\xi)$ as the empirical KL. When the realized KL is larger than the target budget, η_β increases and the KL term in the adversary objective becomes stronger; when it is smaller, the penalty relaxes. The budget parameter therefore continues to index adversarial strength in the implementation.

C.4 Self-play with interacting policy populations

The main theory uses one protagonist policy π and one adversary policy μ as a two-player abstraction. These policies represent the two optimization roles and may be interpreted as role-level joint policies. In our self-play setting, these role policies are instantiated by the current interacting policy collections $\pi_{1:n_\pi}$ and $\mu_{1:n_\mu}$. The robust safety objective uses violations from the protagonist population under the joint adversary policy. At the population level, the objective is written for policy collections, whereas the implementation updates individual policies and uses a local signal for each adversary. We therefore introduce two assumptions specifically to connect the population objective to the implemented decentralized updates.

Population decomposition. Let $\mathcal{I}^\pi$ and $\mathcal{I}^\mu$ denote the protagonist and adversary agent sets, and at the population level write $\pi \equiv \pi_{1:n_\pi} := (\pi_i)_{i\in\mathcal{I}^\pi}$ and $\mu \equiv \mu_{1:n_\mu} := (\mu_j)_{j\in\mathcal{I}^\mu}$. For protagonist i, let $c_i(s) \in \{0, 1\}$ denote its violation indicator and let $N_i(\tau)$ be the first time at which $c_i(s_t) = 1$. The corresponding individual CDF, evaluated under the full policy populations, is

$$F_i^{\pi_{1:n_\pi},\mu_{1:n_\mu}}(s) := \mathbb{E}_{\pi_{1:n_\pi},\mu_{1:n_\mu}} \left[\gamma^{N_i(\tau)} \mid s_0 = s \right]. \tag{134}$$

Let $\mathfrak{M}_{\beta,\mathcal{I}^\mu}$ denote the population-level adversary class induced by the KL budget at strength β. For the joint protagonist policy, the violation indicator in the two-role formulation is $c_{\mathcal{I}^\pi}(s) := \max_{i\in\mathcal{I}^\pi} c_i(s)$, with first violation time $N_{\mathcal{I}^\pi}(\tau) := \inf\{t \geq 0 \mid c_{\mathcal{I}^\pi}(s_t) = 1\}$. Let $F^{\pi_{1:n_\pi},\mu_{1:n_\mu}}$ denote the CDF induced by this indicator. Its robust form is

$$F_\beta^\pi(s) = \max_{\mu_{1:n_\mu}\in\mathfrak{M}_{\beta,\mathcal{I}^\mu}} \mathbb{E}_{\pi_{1:n_\pi},\mu_{1:n_\mu}} \left[\gamma^{N_{\mathcal{I}^\pi}(\tau)} \mid s_0 = s \right]. \tag{135}$$

The indicator records a violation when any protagonist violates, while the maximization selects the worst admissible joint adversary policy.

Assumptions bridging the formulation and updates. The population objective is defined over policy collections, while the implementation updates each policy separately. We first state the assumption that connects the joint population objective to these individual updates.

Assumption 1 (Independent update approximation)**.** *At the population-objective level, the effect of a joint policy update is assumed to be approximated by the collection of individual policy updates. The protagonist component is represented by the individual CDF objectives $\{F_i^\pi\}_{i\in\mathcal{I}^\pi}$, and the adversary component by the corresponding adversary-side CDF targets $\{F_j^\mu\}_{j\in\mathcal{I}^\mu}$. This approximation is evaluated under the rollout distribution induced by the current protagonist and adversary policy populations.*

Assumption 1 specifies how the policies are updated. If every adversary were trained from violations aggregated over the entire protagonist population, the resulting signal mixes violations associated with different adversaries and provides less localized credit for each adversary update (Foerster et al., 2018). The following assumption introduces the locality condition used to obtain a more focused adversary-side signal.

Assumption 2 (Neighborhood sufficiency)**.** *For each adversary j, let $\mathcal{N}_j \subseteq \mathcal{I}^\pi$ denote its interaction neighborhood. The neighborhood is assumed to be sufficient for the adversary-side CDF update: replacing population-wide protagonist violations by those from $\mathcal{N}_j$ preserves the relevant update for μ_j. Protagonists outside $\mathcal{N}_j$ can therefore be omitted from the CDF signal used by adversary j.*

From assumptions to implemented updates. For the safety component, the population objective in Eq. 135 is decomposed into role-specific individual objectives as follows:

$$\begin{aligned} &\text{Protagonist objective:} \quad \min_{\pi} F_\beta^\pi \xrightarrow{\text{Assumption 1}} \min_{\pi_i} F_i^\pi, \quad i \in \mathcal{I}^\pi, \\ &\text{Adversary objective:} \quad \max_{\mu_{1:n_\mu}} F^{\pi_{1:n_\pi}, \mu_{1:n_\mu}} \xrightarrow{\text{Assumption 1}} \max_{\mu_j} F_j^\mu, \quad j \in \mathcal{I}^\mu, \end{aligned} \tag{136}$$

These objectives describe the safety component of the updates; the protagonist continues to use the task objective within the current region specified in the main formulation.

For adversary j, Assumption 2 allows the population-wide violation signal to be replaced by the neighborhood signal

$$c_j(s) := \max_{i \in \mathcal{N}_j} c_i(s). \tag{137}$$

Accordingly, the individual CDF target for μ_j is constructed from c_j. At each self-play round, π_i is updated from F_i^π formed using c_i and the full adversary population, while μ_j is updated from F_j^μ formed using c_j. All individual policies use the same rollout batch and are updated in parallel. The trajectory still evolves under the full policy populations; only the CDF signal for adversary j is restricted to $\mathcal{N}_j$.

Together, Assumptions 1 and 2 connect the joint objective to its per-agent implementation: the former specifies how the update is decomposed, while the latter specifies the adversary-side signal used by each individual update.

The practical algorithm provides a scalable approximation to exact policy iteration through learned CDF critics, clipped policy updates, empirical KL control, and decentralized population updates.

D Additional Experimental Results

D.1 Training and Evaluation Protocol

Training and role assignment. We train on 10,000 WOMD scenarios in PufferDrive. During training and evaluation, each vehicle independently joins the protagonist population with probability $p_{\text{prota}} = 0.8$ and the adversary population otherwise. If a scene has no protagonist, one vehicle is randomly reassigned. Tables 3 and 4 give the optimization settings and configurations for each method.

Held-out evaluation. A separate set of WOMD scenarios is used for validation and protagonist configuration selection under Log Replay. We use the same 1,000 WOMD evaluation scenarios, disjoint from both the training and validation sets, for every method comparison. For every held-out initial state, each of the three independently trained seeds is evaluated with 32 rollouts using sampled actions. We report the mean and standard deviation across the three seeds. Under Log Replay, evaluated protagonists interact with replayed adversaries, while evaluated adversaries interact with replayed protagonists under the same population split.

Baseline training protocols. RARL and ASP train their interacting roles jointly using their respective minimax and teacher–student objectives. SAGE first trains its endpoint adversaries against a fixed protagonist and then trains the protagonist with the resulting adversary curriculum. Intermediate strengths are obtained by interpolating the endpoint policies. The protagonist curriculum used in the main comparison runs from $\lambda = 0$ to 0.2, starting at the $\eta = 0.02$ endpoint. FREA follows a three-stage protocol: protagonist feasibility pretraining, adversary training against the fixed protagonist, and protagonist training against the fixed adversary.

Table 3: Evaluation protocol and policy optimization settings.

Setting	Value
Environment and evaluation	
Training scenarios	10,000 WOMD scenarios
Validation scenarios	Separate held-out WOMD set
Evaluation scenarios	1,000 held-out WOMD scenarios
Protagonist assignment probability p_{prota}	0.8
Minimum protagonists per scene	1
Training seeds	$1, 2, 3$
Evaluation rollouts per initial state	32 (sampled actions)
Policy optimization	
Total training steps	800M
Learning rate	0.003
Rollout batch size	524,288
Minibatch size	32,768
BPTT horizon	32
Update epochs	1
Discount factor γ	0.98
GAE trace parameters	$\lambda_V = \lambda_F = 0.95$
Clipping parameter δ	0.2

Table 4: Configurations for each method.

Parameter	Value
Co-SPAR	
Adversary KL budget β	$\{0.3, 1, 3, 10\}$
Training endpoints	BC Trained ($\beta = 0$), Adv Trained ($\beta \to \infty$)
Selected protagonist setting	$\beta = 0.3$
Protagonist feasibility threshold ϵ	0.1
Protagonist barrier parameter α	1.2
Protagonist barrier cap $w_{\max}$	1.0
Initial KL multiplier η_β	1.0
KL multiplier adjustment factor λ_η	1.2
SAGE	
Endpoint adversary KL coefficients η	$\{0.02, 0\}$
Interpolation coefficient λ	$\{0, 0.2, 0.4, 0.6, 0.8, 1\}$
Selected protagonist setting	Curriculum, $\lambda : 0 \to 0.2$
ASP	
Adversary KL coefficient η	$\{0.02, 6 \times 10^{-3}, 2 \times 10^{-3}, 6 \times 10^{-4}, 0\}$
Selected protagonist setting	$\eta = 0.02$
RARL	
Adversary KL coefficient η	$\{0.02, 6 \times 10^{-3}, 2 \times 10^{-3}, 6 \times 10^{-4}, 0\}$
Selected protagonist setting	$\eta = 0.02$
FREA	
Adversary feasibility threshold ϵ	0.1
BC Prior	
Training epochs	100
BC batch size	8,192
Weight decay	10^{-3}

D.2 Adversary Evaluation

Adversary metrics.

- **Attack effectiveness.** Attack success rate (ASR) is the fraction of adversarial vehicles for which a protagonist in the interaction neighborhood incurs a collision or off-road violation. Protagonist collision rate (PCR) and protagonist off-road rate (POR) separate the two violation types.
- **Attack realism.** The WOSAC realism score aggregates kinematic, interactive, and map-based components (Montali et al., 2023). Adversary off-road rate (AOR) measures road departure, while minimum average displacement error (minADE) reports the closest sampled trajectory to the logged trajectory.

Attack–realism across protagonists. BC Prior is the behavior-cloned policy. BC Trained is a Co-SPAR protagonist trained against a fixed BC Prior policy and corresponds to $\beta = 0$. Across all four protagonist conditions, Co-SPAR attains higher attack success within the realism range covered by competing methods and traces a consistent progression toward stronger attacks as β increases. Figures 5 and 6 show the per-protagonist trade-offs and strength sweeps, respectively.

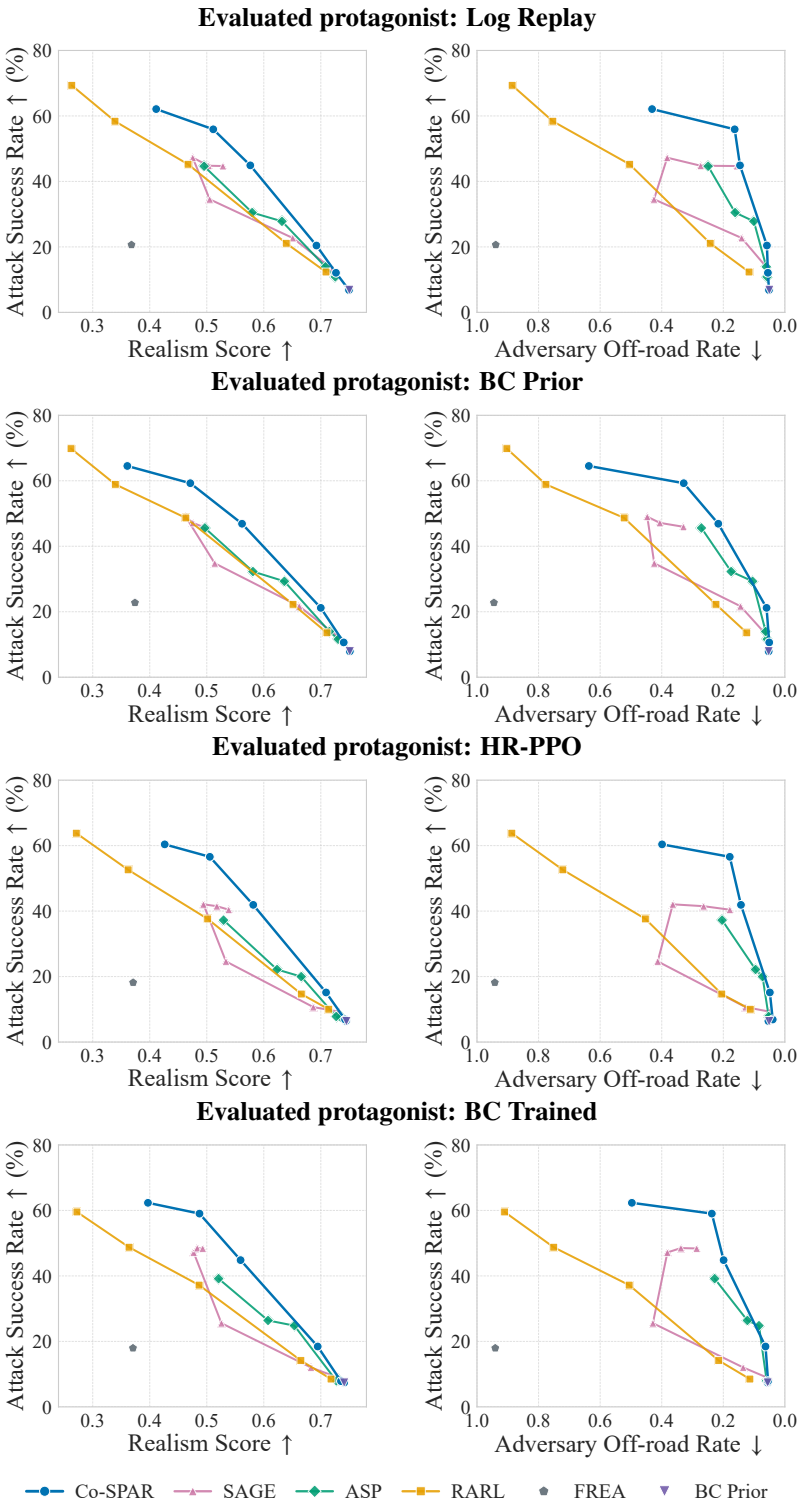


Figure 5: Adversarial trade-offs across the four evaluated protagonist conditions. The left column compares attack success with the aggregate realism score, and the right column compares attack success with adversary off-road rate.

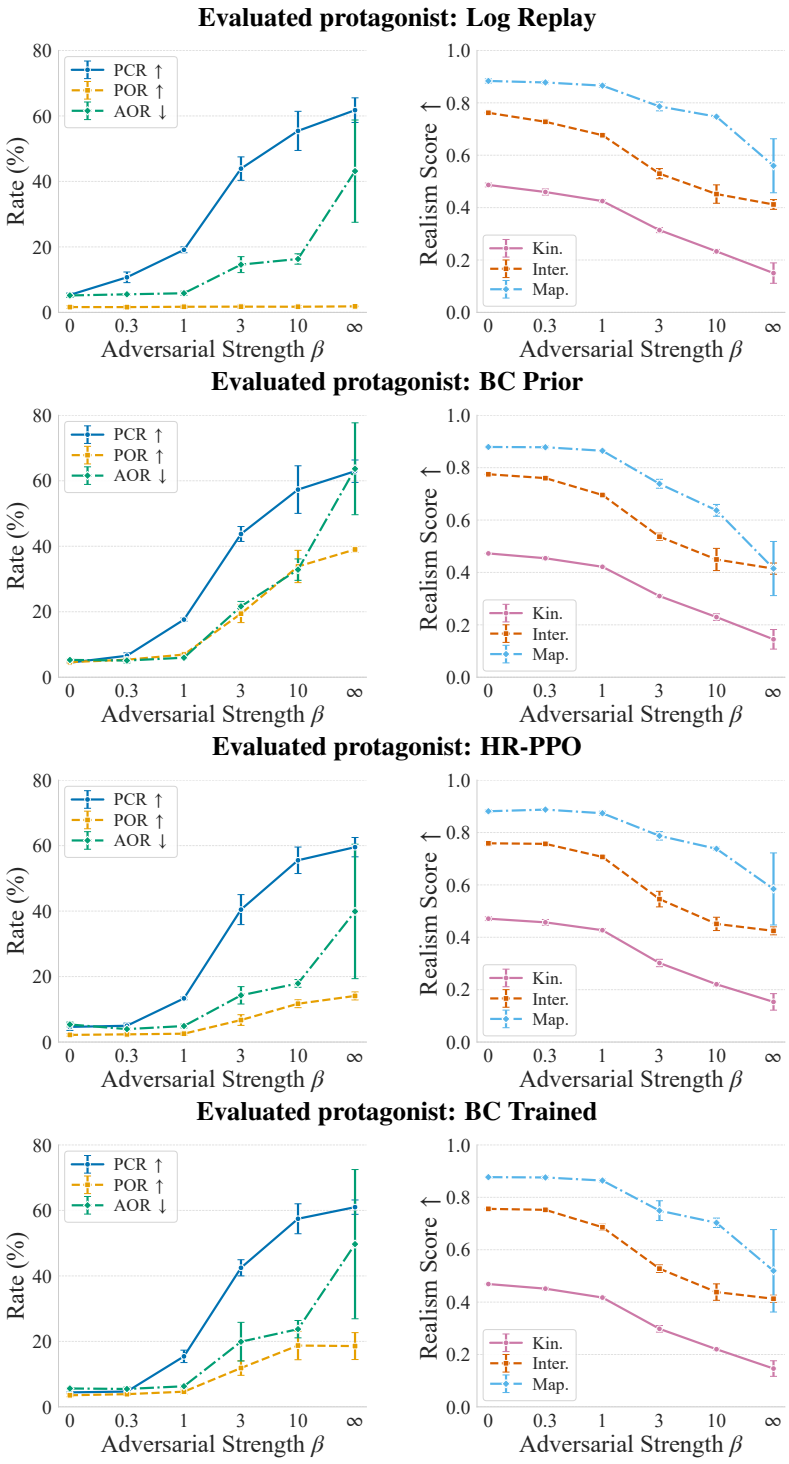


Figure 6: Co-SPAR adversary spectra across the four evaluated protagonist conditions. The left column reports attack outcomes, and the right column reports the components of adversary realism as the adversarial strength β increases.

Detailed adversary metrics across protagonists. The complete metric breakdown confirms the same controllable attack–realism spectrum under each evaluated protagonist (Table 5).

Table 5: Adversary performance across protagonist conditions and method-specific configurations. Metric definitions are provided in Appendix D.2. All rates are percentages, the WOSAC realism score and its components are unitless, and minADE is measured in meters. Each cell reports mean ± standard deviation across three independently trained seeds.

Evaluated Protagonist: Log Replay	Attack Effectiveness			Attack Realism					
Adversary	ASR ↑	PCR ↑	POR ↑	Realism ↑	Kin. ↑	Inter. ↑	Map. ↑	AOR ↓	minADE ↓
BC Prior	6.8±0.3	5.3±0.5	1.6±0.4	0.750±0.003	0.487±0.007	0.762±0.001	0.885±0.006	5.1±0.4	1.92±0.02
FREA	20.6±4.0	19.2±4.3	1.8±0.4	0.368±0.011	0.078±0.008	0.655±0.024	0.165±0.010	93.9±0.9	15.26±4.77
RARL									
$\eta = 0.02$	12.3±0.3	10.8±0.4	1.7±0.4	0.709±0.000	0.453±0.006	0.725±0.003	0.835±0.006	11.6±0.4	2.79±0.10
$\eta = 6 \times 10^{-3}$	21.1±2.6	19.5±2.9	1.8±0.5	0.639±0.022	0.392±0.021	0.673±0.013	0.738±0.034	24.2±4.4	3.59±0.18
$\eta = 2 \times 10^{-3}$	45.2±3.8	44.3±4.0	1.7±0.3	0.467±0.025	0.237±0.032	0.524±0.024	0.525±0.067	50.4±9.5	10.10±3.87
$\eta = 6 \times 10^{-4}$	58.4±1.0	57.8±1.3	1.6±0.4	0.339±0.007	0.115±0.008	0.441±0.008	0.336±0.034	75.4±3.2	17.01±1.38
$\eta = 0$	69.3±4.5	68.7±4.5	1.8±0.1	0.263±0.015	0.060±0.005	0.377±0.024	0.233±0.026	88.6±3.2	18.36±1.57
ASP									
$\eta = 0.02$	10.9±0.4	9.5±0.9	1.7±0.6	0.725±0.004	0.457±0.013	0.732±0.002	0.870±0.004	5.8±1.0	3.47±0.31
$\eta = 6 \times 10^{-3}$	13.9±3.4	12.5±3.0	1.7±0.6	0.708±0.009	0.428±0.015	0.714±0.018	0.861±0.012	6.0±2.1	4.93±1.00
$\eta = 2 \times 10^{-3}$	27.8±2.4	26.6±2.0	1.5±0.5	0.632±0.005	0.326±0.018	0.626±0.008	0.814±0.012	10.1±1.6	13.86±1.29
$\eta = 6 \times 10^{-4}$	30.5±4.4	29.1±4.1	1.5±0.5	0.580±0.043	0.199±0.050	0.609±0.022	0.759±0.075	16.2±9.8	20.41±2.76
$\eta = 0$	44.6±5.0	43.7±5.2	1.4±0.3	0.495±0.020	0.094±0.008	0.525±0.030	0.686±0.039	24.9±5.2	21.17±2.79
SAGE									
$\lambda = 0$	13.2±1.5	11.9±1.4	1.6±0.6	0.720±0.006	0.451±0.006	0.719±0.009	0.875±0.006	5.5±0.2	3.18±0.08
$\lambda = 0.2$	22.7±3.0	21.7±3.0	1.5±0.4	0.651±0.011	0.373±0.004	0.656±0.018	0.802±0.008	14.1±1.2	6.31±0.10
$\lambda = 0.4$	34.5±1.8	33.7±1.3	1.6±0.5	0.505±0.002	0.191±0.003	0.590±0.009	0.575±0.015	42.5±1.8	6.20±0.17
$\lambda = 0.6$	47.2±0.6	46.3±1.1	1.9±0.5	0.475±0.003	0.183±0.003	0.510±0.006	0.598±0.007	38.2±0.8	8.83±0.26
$\lambda = 0.8$	44.8±1.3	43.8±1.7	1.8±0.4	0.503±0.004	0.180±0.003	0.521±0.009	0.666±0.003	27.4±0.5	10.37±0.34
$\lambda = 1$	44.7±0.9	43.7±1.3	1.8±0.3	0.528±0.007	0.171±0.002	0.520±0.008	0.744±0.015	15.6±2.2	12.96±0.39
Co-SPAR (ours)									
$\beta = 0.3$	12.1±1.6	10.7±1.6	1.6±0.6	0.727±0.006	0.460±0.012	0.728±0.009	0.878±0.005	5.5±0.3	3.23±0.15
$\beta = 1$	20.4±1.1	19.1±0.9	1.7±0.5	0.692±0.004	0.425±0.006	0.676±0.006	0.866±0.007	5.8±0.6	2.92±0.14
$\beta = 3$	44.9±3.9	43.9±3.6	1.8±0.5	0.576±0.008	0.314±0.010	0.530±0.019	0.786±0.017	14.6±2.4	5.62±0.48
$\beta = 10$	55.9±6.0	55.4±6.0	1.7±0.4	0.511±0.019	0.233±0.007	0.452±0.035	0.747±0.005	16.3±1.6	12.84±2.09
$\beta \to \infty$	62.1±3.8	61.8±3.8	1.8±0.1	0.411±0.044	0.150±0.039	0.412±0.019	0.560±0.103	43.1±15.6	14.38±1.25

Adversary metrics (continued)

Adversary	Attack Effectiveness			Attack Realism					
	ASR ↑	PCR ↑	POR ↑	Realism ↑	Kin. ↑	Inter. ↑	Map. ↑	AOR ↓	minADE ↓
Evaluated Protagonist: BC Prior									
BC Prior	7.9±0.2	4.4±0.1	4.4±0.2	0.751±0.000	0.473±0.000	0.775±0.000	0.879±0.001	5.3±0.1	2.40±0.02
FREA	22.8±4.2	18.5±4.2	7.0±1.0	0.374±0.011	0.075±0.005	0.667±0.024	0.168±0.014	94.5±0.5	15.11±4.20
RARL									
$\eta = 0.02$	13.6±0.9	9.8±0.8	5.9±0.5	0.711±0.003	0.446±0.005	0.737±0.005	0.828±0.002	12.4±0.4	2.86±0.08
$\eta = 6 \times 10^{-3}$	22.3±2.3	18.4±2.5	7.2±0.4	0.651±0.022	0.392±0.020	0.689±0.015	0.751±0.032	22.4±4.3	3.65±0.21
$\eta = 2 \times 10^{-3}$	48.6±4.8	45.8±4.9	14.3±2.3	0.463±0.033	0.235±0.036	0.522±0.029	0.519±0.072	52.2±9.6	10.77±4.43
$\eta = 6 \times 10^{-4}$	58.9±0.6	57.2±0.7	21.1±2.6	0.340±0.011	0.114±0.007	0.451±0.004	0.326±0.034	77.6±4.0	18.17±0.91
$\eta = 0$	69.8±2.4	68.3±2.7	24.8±7.2	0.262±0.014	0.057±0.003	0.380±0.016	0.227±0.024	90.3±2.7	19.10±1.10
ASP									
$\eta = 0.02$	11.6±0.7	7.9±0.4	4.7±0.2	0.730±0.005	0.449±0.012	0.749±0.002	0.868±0.008	5.7±0.9	3.78±0.40
$\eta = 6 \times 10^{-3}$	14.0±0.8	10.5±0.8	5.0±0.5	0.715±0.004	0.426±0.009	0.733±0.005	0.857±0.007	6.2±1.9	5.09±1.00
$\eta = 2 \times 10^{-3}$	29.3±1.8	26.1±1.9	9.2±0.9	0.636±0.009	0.330±0.018	0.637±0.011	0.810±0.008	10.5±1.6	14.04±1.52
$\eta = 6 \times 10^{-4}$	32.3±4.3	28.9±4.8	8.7±2.7	0.581±0.046	0.200±0.053	0.618±0.027	0.751±0.068	17.5±9.4	20.54±2.93
$\eta = 0$	45.6±3.0	42.8±3.3	10.7±1.7	0.497±0.019	0.094±0.009	0.537±0.019	0.674±0.045	27.1±6.4	21.16±2.08
SAGE									
$\lambda = 0$	12.3±0.1	8.8±0.1	4.8±0.0	0.731±0.001	0.450±0.000	0.745±0.001	0.872±0.001	5.4±0.1	3.51±0.06
$\lambda = 0.2$	21.7±0.0	18.5±0.2	6.1±0.1	0.662±0.001	0.374±0.000	0.681±0.001	0.802±0.001	14.4±0.1	6.68±0.05
$\lambda = 0.4$	34.8±0.3	31.6±0.3	9.5±0.1	0.514±0.002	0.194±0.001	0.608±0.002	0.577±0.003	42.4±0.5	6.63±0.04
$\lambda = 0.6$	49.0±0.7	46.4±0.7	19.2±0.1	0.463±0.003	0.180±0.000	0.516±0.004	0.557±0.004	44.7±0.6	9.87±0.20
$\lambda = 0.8$	47.1±0.6	44.3±0.6	21.7±0.4	0.476±0.001	0.178±0.000	0.529±0.003	0.577±0.001	40.6±0.1	11.72±0.16
$\lambda = 1$	45.9±0.2	43.3±0.2	21.8±0.5	0.493±0.001	0.172±0.000	0.534±0.001	0.625±0.002	32.9±0.3	14.10±0.05
Co-SPAR (ours)									
$\beta = 0.3$	10.6±0.9	6.6±0.9	5.3±0.1	0.740±0.002	0.454±0.005	0.760±0.005	0.878±0.005	5.1±0.7	3.96±0.11
$\beta = 1$	21.2±0.7	17.6±0.5	6.9±0.3	0.700±0.001	0.422±0.001	0.696±0.002	0.865±0.003	6.0±0.3	3.20±0.20
$\beta = 3$	46.9±2.3	43.7±2.3	19.4±2.7	0.562±0.011	0.310±0.005	0.537±0.013	0.739±0.017	21.6±1.5	6.03±0.39
$\beta = 10$	59.3±7.0	57.3±7.3	33.8±4.9	0.471±0.029	0.230±0.013	0.450±0.042	0.637±0.023	32.9±3.3	12.79±1.09
$\beta \to \infty$	64.5±3.1	62.9±3.4	39.0±0.9	0.361±0.039	0.145±0.037	0.415±0.021	0.415±0.103	63.7±14.0	13.89±0.20

Adversary metrics (continued)

Adversary	Attack Effectiveness			Attack Realism					
	ASR ↑	PCR ↑	POR ↑	Realism ↑	Kin. ↑	Inter. ↑	Map. ↑	AOR ↓	minADE ↓
Evaluated Protagonist: HR-PPO									
BC Prior	6.3±1.1	4.5±0.8	2.2±0.4	0.745±0.004	0.471±0.007	0.760±0.008	0.882±0.005	5.1±0.8	2.64±0.08
FREA	18.2±3.1	15.9±3.4	3.3±0.7	0.371±0.005	0.075±0.004	0.665±0.019	0.162±0.015	94.2±0.8	15.70±5.07
RARL									
$\eta = 0.02$	10.0±0.8	8.2±0.5	2.4±0.3	0.714±0.001	0.451±0.006	0.734±0.004	0.837±0.007	11.2±0.2	2.82±0.18
$\eta = 6 \times 10^{-3}$	14.7±2.0	12.8±2.0	2.6±0.7	0.666±0.015	0.398±0.017	0.708±0.011	0.765±0.020	20.5±2.9	3.57±0.16
$\eta = 2 \times 10^{-3}$	37.6±5.3	36.1±5.4	4.2±1.2	0.502±0.040	0.245±0.040	0.567±0.037	0.564±0.081	45.3±11.4	10.64±4.62
$\eta = 6 \times 10^{-4}$	52.6±1.5	51.5±1.6	6.1±0.7	0.362±0.011	0.113±0.006	0.474±0.008	0.361±0.025	72.1±2.1	17.80±0.79
$\eta = 0$	63.8±4.6	62.9±4.8	7.7±2.0	0.271±0.011	0.059±0.006	0.396±0.019	0.232±0.021	88.7±2.4	19.92±4.06
ASP									
$\eta = 0.02$	7.6±0.7	5.9±0.7	2.2±0.5	0.735±0.007	0.453±0.013	0.750±0.006	0.876±0.006	4.8±1.3	3.84±0.49
$\eta = 6 \times 10^{-3}$	7.8±0.4	6.0±0.5	2.3±0.3	0.727±0.006	0.429±0.014	0.751±0.002	0.867±0.011	5.3±1.7	5.29±1.07
$\eta = 2 \times 10^{-3}$	20.0±2.0	18.3±1.8	2.8±0.7	0.666±0.009	0.338±0.010	0.678±0.010	0.837±0.013	7.3±2.0	15.89±0.86
$\eta = 6 \times 10^{-4}$	22.2±7.3	20.6±7.3	2.7±0.7	0.623±0.038	0.207±0.052	0.664±0.043	0.809±0.045	9.5±5.9	22.19±1.38
$\eta = 0$	37.2±8.0	35.8±8.1	3.4±1.1	0.529±0.033	0.098±0.008	0.575±0.050	0.718±0.037	20.4±4.9	21.48±3.86
SAGE									
$\lambda = 0$	9.3±0.4	7.5±0.5	2.4±0.3	0.731±0.002	0.454±0.006	0.739±0.002	0.879±0.005	5.1±0.2	3.72±0.16
$\lambda = 0.2$	10.7±1.2	9.1±1.3	2.2±0.3	0.687±0.003	0.380±0.003	0.726±0.005	0.812±0.002	12.8±0.6	6.71±0.25
$\lambda = 0.4$	24.6±1.9	22.9±1.9	3.1±0.6	0.534±0.004	0.196±0.003	0.643±0.013	0.585±0.013	41.3±1.4	6.37±0.04
$\lambda = 0.6$	42.1±2.3	40.8±2.5	5.1±0.9	0.494±0.010	0.184±0.002	0.540±0.019	0.612±0.004	36.4±0.6	9.91±0.93
$\lambda = 0.8$	41.5±3.4	40.2±3.5	6.2±1.4	0.518±0.012	0.180±0.000	0.547±0.022	0.673±0.006	26.4±0.6	12.25±1.14
$\lambda = 1$	40.4±4.8	39.1±4.9	6.5±2.0	0.538±0.016	0.170±0.002	0.552±0.028	0.732±0.014	17.9±2.0	14.76±1.02
Co-SPAR (ours)									
$\beta = 0.3$	6.9±0.9	5.0±0.8	2.4±0.3	0.743±0.001	0.457±0.011	0.757±0.008	0.887±0.003	4.0±0.2	4.32±0.28
$\beta = 1$	15.1±0.5	13.3±0.4	2.6±0.5	0.709±0.002	0.427±0.006	0.707±0.004	0.874±0.008	4.9±0.6	3.17±0.15
$\beta = 3$	41.9±4.5	40.5±4.6	6.7±1.6	0.582±0.014	0.302±0.014	0.546±0.030	0.787±0.016	14.3±2.7	6.15±0.20
$\beta = 10$	56.6±4.1	55.5±4.0	11.7±1.2	0.505±0.013	0.221±0.006	0.451±0.025	0.738±0.006	17.9±1.2	13.81±0.42
$\beta \to \infty$	60.4±3.0	59.5±3.0	14.1±1.2	0.426±0.054	0.153±0.032	0.425±0.015	0.584±0.138	39.9±20.5	13.19±1.35

Adversary metrics (continued)

Evaluated Protagonist: BC Trained		Attack Effectiveness			Attack Realism					
	Adversary	ASR ↑	PCR ↑	POR ↑	Realism ↑	Kin. ↑	Inter. ↑	Map. ↑	AOR ↓	minADE ↓
	BC Prior	7.4±0.5	4.4±0.2	3.6±0.4	0.741±0.001	0.469±0.001	0.756±0.001	0.878±0.000	5.6±0.1	2.15±0.04
	FREA	18.0±5.2	14.7±5.1	4.5±0.5	0.371±0.016	0.077±0.006	0.654±0.031	0.173±0.012	94.1±1.0	15.25±5.60
	RARL									
	$\eta = 0.02$	8.5±0.4	5.5±0.7	3.7±0.3	0.718±0.002	0.447±0.004	0.747±0.004	0.836±0.002	11.4±0.2	2.83±0.08
	$\eta = 6 \times 10^{-3}$	14.2±1.7	11.0±1.6	4.2±0.3	0.665±0.016	0.395±0.016	0.713±0.009	0.758±0.026	21.5±3.7	3.48±0.18
	$\eta = 2 \times 10^{-3}$	37.2±5.3	34.6±5.4	7.7±1.5	0.487±0.032	0.236±0.037	0.564±0.033	0.530±0.069	50.6±9.8	10.29±4.69
	$\eta = 6 \times 10^{-4}$	48.8±1.0	46.8±1.0	10.6±2.9	0.364±0.006	0.113±0.005	0.488±0.009	0.347±0.030	75.1±3.7	17.95±0.84
	$\eta = 0$	59.5±3.7	58.2±4.0	10.1±3.3	0.272±0.013	0.059±0.002	0.406±0.019	0.223±0.022	91.0±2.4	19.00±1.81
	ASP									
	$\eta = 0.02$	8.1±0.3	4.9±0.3	3.8±0.2	0.733±0.005	0.449±0.011	0.753±0.001	0.869±0.009	5.5±1.0	3.64±0.39
	$\eta = 6 \times 10^{-3}$	8.0±0.6	4.9±0.6	3.7±0.4	0.727±0.004	0.427±0.010	0.759±0.002	0.858±0.007	5.9±1.9	5.06±1.15
	$\eta = 2 \times 10^{-3}$	24.8±1.7	21.6±1.5	4.8±0.3	0.654±0.009	0.340±0.011	0.657±0.009	0.828±0.012	8.4±2.1	14.59±2.32
	$\eta = 6 \times 10^{-4}$	26.4±6.9	23.1±6.9	5.1±0.6	0.607±0.046	0.206±0.054	0.646±0.040	0.788±0.061	12.2±8.2	22.35±0.91
	$\eta = 0$	39.2±2.3	36.6±2.7	7.9±0.7	0.521±0.020	0.095±0.008	0.566±0.019	0.706±0.037	22.8±5.4	22.42±2.36
	SAGE									
	$\lambda = 0$	8.9±0.1	5.9±0.2	3.8±0.2	0.730±0.000	0.450±0.000	0.745±0.001	0.871±0.001	5.7±0.3	3.26±0.04
	$\lambda = 0.2$	12.0±0.2	9.1±0.3	3.9±0.3	0.683±0.000	0.377±0.000	0.721±0.002	0.809±0.002	13.6±0.3	6.59±0.11
	$\lambda = 0.4$	25.6±1.8	22.5±1.6	5.7±0.8	0.526±0.006	0.191±0.001	0.638±0.009	0.572±0.007	42.9±1.0	6.25±0.07
	$\lambda = 0.6$	47.2±1.2	45.1±1.2	11.6±3.3	0.477±0.008	0.177±0.003	0.511±0.009	0.605±0.014	38.2±2.1	10.01±0.64
	$\lambda = 0.8$	48.5±1.8	46.5±1.6	13.9±4.0	0.483±0.015	0.168±0.005	0.509±0.010	0.631±0.045	33.8±6.4	13.23±1.71
	$\lambda = 1$	48.4±1.7	46.3±1.7	14.5±3.9	0.493±0.016	0.160±0.005	0.511±0.009	0.660±0.051	28.6±7.1	15.89±2.41
	Co-SPAR (ours)									
	$\beta = 0.3$	7.9±0.7	4.7±0.4	3.9±0.4	0.735±0.001	0.451±0.005	0.752±0.001	0.876±0.005	5.5±0.6	3.30±0.06
	$\beta = 1$	18.4±2.2	15.4±1.9	4.6±0.8	0.695±0.007	0.417±0.004	0.686±0.012	0.864±0.005	6.3±0.4	3.08±0.12
	$\beta = 3$	44.8±2.4	42.5±2.5	11.9±2.2	0.559±0.013	0.298±0.013	0.528±0.014	0.749±0.038	19.9±5.9	6.20±1.00
	$\beta = 10$	59.0±4.5	57.5±4.6	18.7±4.3	0.487±0.019	0.220±0.004	0.438±0.031	0.703±0.018	23.7±2.7	15.40±4.26
	$\beta \to \infty$	62.3±2.0	61.0±2.2	18.6±4.1	0.397±0.054	0.146±0.030	0.413±0.014	0.520±0.157	49.7±22.8	15.34±3.48

Adversary metrics (continued)

Evaluated Protagonist: Average	Attack Effectiveness			Attack Realism					
Adversary	ASR ↑	PCR ↑	POR ↑	Realism ↑	Kin. ↑	Inter. ↑	Map. ↑	AOR ↓	minADE ↓
BC Prior	7.1±0.3	4.7±0.1	2.9±0.2	0.747±0.001	0.475±0.003	0.763±0.002	0.881±0.002	5.3±0.3	2.28±0.01
FREA	19.9±4.1	17.1±4.2	4.2±0.5	0.371±0.011	0.076±0.005	0.660±0.024	0.167±0.012	94.2±0.7	15.33±4.89
RARL									
$\eta = 0.02$	11.1±0.2	8.6±0.2	3.4±0.2	0.713±0.001	0.449±0.004	0.736±0.002	0.834±0.003	11.7±0.2	2.83±0.10
$\eta = 6 \times 10^{-3}$	18.0±2.1	15.4±2.2	4.0±0.3	0.655±0.019	0.394±0.019	0.696±0.011	0.753±0.028	22.2±3.8	3.57±0.17
$\eta = 2 \times 10^{-3}$	42.2±4.7	40.2±4.9	7.0±1.3	0.480±0.032	0.238±0.036	0.544±0.031	0.535±0.072	49.6±10.0	10.45±4.39
$\eta = 6 \times 10^{-4}$	54.7±0.7	53.3±0.8	9.9±0.2	0.351±0.007	0.114±0.006	0.463±0.004	0.343±0.029	75.1±2.9	17.73±0.66
$\eta = 0$	65.6±3.8	64.5±3.9	11.1±3.0	0.267±0.013	0.059±0.004	0.390±0.020	0.229±0.022	89.7±2.6	19.09±1.88
ASP									
$\eta = 0.02$	9.6±0.4	7.0±0.3	3.1±0.2	0.731±0.005	0.452±0.012	0.746±0.002	0.871±0.006	5.5±1.1	3.68±0.39
$\eta = 6 \times 10^{-3}$	10.9±0.8	8.5±0.7	3.1±0.3	0.719±0.005	0.428±0.012	0.739±0.005	0.861±0.009	5.9±1.8	5.09±1.06
$\eta = 2 \times 10^{-3}$	25.5±1.8	23.2±1.6	4.6±0.2	0.647±0.008	0.333±0.014	0.649±0.008	0.822±0.011	9.1±1.8	14.59±1.48
$\eta = 6 \times 10^{-4}$	27.8±5.6	25.4±5.7	4.5±0.9	0.598±0.043	0.203±0.052	0.634±0.032	0.777±0.061	13.8±8.2	21.37±1.82
$\eta = 0$	41.7±4.6	39.7±4.8	5.9±0.7	0.511±0.023	0.095±0.008	0.551±0.029	0.696±0.038	23.8±5.2	21.56±2.76
SAGE									
$\lambda = 0$	10.9±0.5	8.5±0.4	3.1±0.2	0.728±0.002	0.451±0.003	0.737±0.002	0.874±0.003	5.4±0.2	3.42±0.04
$\lambda = 0.2$	16.8±1.0	14.6±1.0	3.4±0.2	0.671±0.004	0.376±0.002	0.696±0.005	0.806±0.003	13.7±0.5	6.57±0.08
$\lambda = 0.4$	29.9±0.9	27.7±0.7	5.0±0.0	0.520±0.001	0.193±0.002	0.620±0.004	0.577±0.008	42.3±0.9	6.36±0.06
$\lambda = 0.6$	46.4±0.3	44.6±0.5	9.4±0.8	0.477±0.002	0.181±0.001	0.519±0.004	0.593±0.001	39.4±0.1	9.66±0.20
$\lambda = 0.8$	45.5±0.9	43.7±1.0	10.9±1.1	0.495±0.006	0.176±0.002	0.526±0.006	0.637±0.012	32.0±1.6	11.89±0.32
$\lambda = 1$	44.9±1.0	43.1±1.1	11.2±1.2	0.513±0.008	0.168±0.001	0.529±0.006	0.690±0.019	23.8±2.6	14.43±0.42
Co-SPAR (ours)									
$\beta = 0.3$	9.4±0.6	6.7±0.5	3.3±0.2	0.736±0.002	0.456±0.008	0.749±0.002	0.880±0.002	5.0±0.3	3.70±0.12
$\beta = 1$	18.8±0.7	16.4±0.7	4.0±0.1	0.699±0.003	0.423±0.004	0.691±0.003	0.867±0.005	5.7±0.4	3.09±0.14
$\beta = 3$	44.6±3.3	42.6±3.2	9.9±0.6	0.570±0.007	0.306±0.009	0.535±0.019	0.765±0.014	17.6±2.4	6.00±0.51
$\beta = 10$	57.7±5.2	56.4±5.3	16.5±2.2	0.494±0.020	0.226±0.007	0.448±0.033	0.706±0.012	22.7±2.0	13.71±1.92
$\beta \to \infty$	62.3±2.6	61.3±2.6	18.4±1.3	0.399±0.047	0.149±0.034	0.416±0.015	0.520±0.124	49.1±18.0	14.20±1.37

D.3 PROTAGONIST EVALUATION

Protagonist metrics.

- **Overall driving performance.** Success rate (SR) is the fraction of protagonists that reach the goal without a collision or off-road violation.
- **Task progress.** Goal-reaching rate (GRR) is the fraction of protagonists that reach the goal. Safe route progress (SRP) is the maximum normalized route progress before the first safety violation, with a value of 1 when the goal is reached.
- **Safety Violations.** Safety violation rate (SVR) is the fraction of protagonists that incur a collision or off-road violation. Collision rate (CR) and off-road rate (OR) report the two violation types separately.

Protagonist performance against individual adversaries. The strong adversary average equally weights the strongest evaluated configuration from each method: FREA, RARL ($\eta = 0$), ASP ($\eta = 0$), SAGE ($\lambda = 1$), and Co-SPAR ($\beta \to \infty$). Co-SPAR leads in SR and SRP against each individual strong adversary as well as their equal-weight average (Table 6).

Table 6: Protagonist performance across evaluated adversaries. Metric definitions are provided in Appendix D.3. Strong Adv. Avg. equally averages the strongest evaluated adversary configuration from each method. Overall Average additionally includes Log Replay and BC Prior. All metrics are percentages. Each cell reports mean ± standard deviation across three independently trained seeds. **Bold** and underlined mark the best and second-best displayed means within each adversary block.

Eval. Adv.	Protagonist	Overall SR ↑	Task Progress GRR ↑	Task Progress SRP ↑	Safety Violations SVR ↓	Safety Violations CR ↓	Safety Violations OR ↓
Log Replay	HR-PPO	60.8 ± 6.6	65.3 ± 7.8	83.7 ± 3.8	$\mathbf{6.7} \pm 0.9$	3.4 ± 0.5	$\mathbf{3.6} \pm 0.4$
	FREA	66.1 ± 2.4	73.8 ± 2.9	86.7 ± 1.9	9.7 ± 0.7	5.1 ± 0.4	5.0 ± 0.4
	RARL	69.7 ± 7.3	76.3 ± 8.4	88.2 ± 2.8	8.6 ± 0.9	4.3 ± 0.5	4.7 ± 0.5
	ASP	69.6 ± 3.0	75.6 ± 3.7	88.7 ± 2.6	8.5 ± 0.9	4.2 ± 0.6	4.6 ± 0.3
	SAGE	68.3 ± 6.0	74.5 ± 6.3	86.7 ± 2.8	8.4 ± 0.5	4.1 ± 0.6	4.7 ± 0.2
	Co-SPAR	$\mathbf{82.2} \pm 1.7$	$\mathbf{88.3} \pm 1.3$	$\mathbf{93.4} \pm 1.1$	7.5 ± 0.8	$\mathbf{3.2} \pm 0.4$	4.6 ± 0.5
BC Prior	HR-PPO	60.5 ± 6.5	64.8 ± 7.6	83.5 ± 3.9	$\mathbf{6.0} \pm 1.0$	$\mathbf{2.6} \pm 0.8$	$\mathbf{3.6} \pm 0.3$
	FREA	66.0 ± 2.7	73.5 ± 3.1	86.7 ± 2.0	8.8 ± 0.7	4.2 ± 0.4	4.9 ± 0.3
	RARL	69.4 ± 7.0	75.7 ± 8.2	88.1 ± 2.8	8.0 ± 1.1	3.8 ± 0.7	4.5 ± 0.5
	ASP	69.3 ± 3.4	75.1 ± 4.0	88.6 ± 2.8	8.0 ± 1.0	3.8 ± 0.6	4.5 ± 0.4
	SAGE	68.1 ± 5.6	74.0 ± 5.9	86.7 ± 2.6	7.8 ± 0.6	3.6 ± 0.6	4.5 ± 0.2
	Co-SPAR	$\mathbf{82.3} \pm 1.4$	$\mathbf{88.0} \pm 1.3$	$\mathbf{93.6} \pm 1.0$	7.0 ± 0.7	2.8 ± 0.3	4.4 ± 0.5
FREA	HR-PPO	59.9 ± 7.1	66.1 ± 8.2	82.8 ± 4.4	$\mathbf{8.5} \pm 0.4$	$\mathbf{5.1} \pm 0.1$	$\mathbf{3.7} \pm 0.4$
	FREA	65.1 ± 2.6	74.0 ± 2.9	85.9 ± 1.9	10.6 ± 0.4	6.0 ± 0.1	5.1 ± 0.4
	RARL	68.5 ± 7.4	76.7 ± 8.2	87.1 ± 3.2	10.3 ± 0.2	6.1 ± 0.4	4.6 ± 0.5
	ASP	67.8 ± 3.1	75.6 ± 3.6	87.2 ± 2.4	10.7 ± 1.9	6.5 ± 1.6	4.6 ± 0.4
	SAGE	66.8 ± 5.8	74.7 ± 5.7	85.5 ± 2.5	10.4 ± 0.6	6.2 ± 0.7	4.6 ± 0.1
	Co-SPAR	$\mathbf{80.5} \pm 1.0$	$\mathbf{88.2} \pm 1.2$	$\mathbf{92.5} \pm 0.5$	9.3 ± 0.8	5.2 ± 0.6	4.4 ± 0.4
RARL ($\eta = 0$)	HR-PPO	53.5 ± 6.3	66.1 ± 8.6	76.7 ± 4.5	19.6 ± 1.5	16.4 ± 1.4	$\mathbf{4.5} \pm 0.1$
	FREA	58.1 ± 1.2	74.5 ± 2.3	79.4 ± 0.7	21.7 ± 1.8	17.5 ± 1.4	5.9 ± 0.8
	RARL	60.8 ± 5.9	76.5 ± 8.6	80.8 ± 3.2	21.0 ± 1.4	17.2 ± 1.1	5.5 ± 0.4
	ASP	60.3 ± 3.6	75.4 ± 3.2	81.1 ± 2.7	21.4 ± 1.2	17.5 ± 1.2	5.8 ± 0.7
	SAGE	59.6 ± 4.9	75.2 ± 5.2	79.5 ± 1.8	20.9 ± 1.1	17.0 ± 1.0	5.8 ± 0.4
	Co-SPAR	$\mathbf{72.2} \pm 1.0$	$\mathbf{87.0} \pm 1.6$	$\mathbf{86.4} \pm 0.6$	$\mathbf{19.4} \pm 0.3$	$\mathbf{15.6} \pm 0.8$	5.5 ± 0.5

Eval. Adv.	Protagonist	Overall SR ↑	Task Progress GRR ↑	Task Progress SRP ↑	Safety Violations SVR ↓	Safety Violations CR ↓	Safety Violations OR ↓
ASP ($\eta = 0$)	HR-PPO	57.6 ± 6.0	64.9 ± 8.8	79.8 ± 3.5	$\mathbf{11.9} \pm 2.8$	$\mathbf{8.6} \pm 2.6$	$\mathbf{3.8} \pm 0.5$
	FREA	62.5 ± 2.6	73.5 ± 2.3	82.6 ± 2.2	15.4 ± 0.7	10.9 ± 0.8	5.3 ± 0.6
	RARL	65.5 ± 6.0	75.7 ± 9.0	84.0 ± 2.3	14.1 ± 2.3	10.0 ± 2.0	4.8 ± 0.7
	ASP	65.2 ± 3.0	75.0 ± 4.3	84.1 ± 2.6	14.5 ± 1.1	10.5 ± 1.0	4.7 ± 0.4
	SAGE	64.3 ± 5.1	74.0 ± 5.5	82.5 ± 2.7	13.9 ± 0.7	9.7 ± 0.8	4.9 ± 0.0
	Co-SPAR	$\mathbf{77.7} \pm 1.9$	$\mathbf{87.3} \pm 1.6$	$\mathbf{89.4} \pm 1.6$	12.9 ± 1.7	8.7 ± 1.3	4.9 ± 0.5
SAGE ($\lambda = 1$)	HR-PPO	57.2 ± 6.4	64.2 ± 8.3	79.5 ± 4.1	$\mathbf{12.1} \pm 1.5$	$\mathbf{8.8} \pm 1.2$	$\mathbf{4.2} \pm 0.5$
	FREA	62.1 ± 2.4	72.7 ± 2.5	82.3 ± 1.7	15.4 ± 0.6	11.0 ± 0.2	5.8 ± 0.8
	RARL	65.1 ± 6.4	74.6 ± 8.3	83.8 ± 2.7	14.4 ± 1.2	10.3 ± 0.9	5.3 ± 0.9
	ASP	64.5 ± 3.3	73.5 ± 4.0	83.7 ± 2.8	15.0 ± 1.3	10.9 ± 1.0	5.3 ± 0.8
	SAGE	63.4 ± 5.0	72.6 ± 5.4	82.0 ± 2.3	14.6 ± 0.2	10.3 ± 0.2	5.8 ± 0.7
	Co-SPAR	$\mathbf{75.8} \pm 1.6$	$\mathbf{85.1} \pm 2.4$	$\mathbf{87.9} \pm 1.2$	15.0 ± 0.7	11.0 ± 0.3	5.7 ± 0.4
Co-SPAR ($\beta \to \infty$)	HR-PPO	55.0 ± 6.3	65.1 ± 7.5	77.5 ± 4.2	$\mathbf{16.8} \pm 1.1$	$\mathbf{13.6} \pm 1.1$	$\mathbf{5.6} \pm 0.2$
	FREA	59.6 ± 1.5	73.4 ± 2.7	80.3 ± 1.1	19.2 ± 1.7	15.0 ± 1.5	6.4 ± 0.4
	RARL	62.4 ± 6.1	75.2 ± 7.9	81.5 ± 2.7	18.7 ± 1.5	14.8 ± 1.2	6.5 ± 0.9
	ASP	61.8 ± 3.8	73.7 ± 3.4	81.6 ± 2.7	19.3 ± 1.4	15.3 ± 1.3	7.0 ± 0.8
	SAGE	60.8 ± 4.6	73.2 ± 5.2	80.0 ± 1.9	18.8 ± 1.3	14.8 ± 1.3	7.1 ± 0.7
	Co-SPAR	$\mathbf{72.6} \pm 1.3$	$\mathbf{84.6} \pm 2.7$	$\mathbf{85.8} \pm 1.0$	18.9 ± 0.6	15.1 ± 1.0	6.8 ± 0.3
Strong Adv. Avg.	HR-PPO	56.6 ± 6.4	65.2 ± 8.3	79.2 ± 4.1	$\mathbf{13.8} \pm 1.1$	$\mathbf{10.5} \pm 0.8$	$\mathbf{4.4} \pm 0.3$
	FREA	61.5 ± 2.0	73.6 ± 2.5	82.1 ± 1.5	16.5 ± 0.9	12.1 ± 0.5	5.7 ± 0.6
	RARL	64.4 ± 6.3	75.7 ± 8.4	83.4 ± 2.8	15.7 ± 1.1	11.7 ± 0.7	5.4 ± 0.7
	ASP	63.9 ± 3.3	74.6 ± 3.7	83.5 ± 2.6	16.2 ± 1.1	12.1 ± 0.8	5.5 ± 0.6
	SAGE	63.0 ± 5.0	74.0 ± 5.4	81.9 ± 2.2	15.7 ± 0.3	11.6 ± 0.3	5.7 ± 0.3
	Co-SPAR	$\mathbf{75.8} \pm 1.3$	$\mathbf{86.4} \pm 1.9$	$\mathbf{88.4} \pm 0.9$	15.1 ± 0.4	11.1 ± 0.1	5.4 ± 0.4
Overall Average	HR-PPO	57.8 ± 6.4	65.2 ± 8.1	80.5 ± 4.0	$\mathbf{11.7} \pm 1.0$	$\mathbf{8.3} \pm 0.7$	$\mathbf{4.1} \pm 0.3$
	FREA	62.8 ± 2.2	73.6 ± 2.7	83.4 ± 1.6	14.4 ± 0.8	9.9 ± 0.4	5.5 ± 0.5
	RARL	65.9 ± 6.6	75.8 ± 8.4	84.8 ± 2.8	13.6 ± 1.0	9.5 ± 0.6	5.1 ± 0.6
	ASP	65.5 ± 3.3	74.8 ± 3.7	85.0 ± 2.6	13.9 ± 1.0	9.8 ± 0.7	5.2 ± 0.5
	SAGE	64.5 ± 5.2	74.0 ± 5.6	83.3 ± 2.3	13.6 ± 0.1	9.4 ± 0.1	5.4 ± 0.2
	Co-SPAR	$\mathbf{77.6} \pm 1.3$	$\mathbf{86.9} \pm 1.7$	$\mathbf{89.8} \pm 0.9$	12.8 ± 0.5	8.8 ± 0.1	5.2 ± 0.4

Protagonist performance under different control settings. We further evaluate protagonist policies under single-protagonist log replay and all-agent policy control. In single-protagonist log replay, one vehicle follows the evaluated protagonist policy while all other vehicles replay their logged trajectories. In all-agent policy control, the evaluated protagonist policy controls every vehicle. Co-SPAR leads in SR and SRP in both settings. It also attains the lowest SVR under single-protagonist log replay and the second-lowest SVR, together with the lowest collision rate, under all-agent policy control (Tables 7 and 8).

Table 7: Protagonist performance under single-protagonist log replay. All metrics are percentages. Each cell reports mean ± standard deviation across three independently trained seeds. **Bold** and underlined mark the best and second-best displayed means, respectively.

Protagonist	Overall SR ↑	Task Progress GRR ↑	Task Progress SRP ↑	Safety Violations SVR ↓	Safety Violations CR ↓	Safety Violations OR ↓
HR-PPO	43.5 ± 1.6	58.0 ± 2.1	68.4 ± 0.9	27.0 ± 3.0	23.9 ± 2.0	4.6 ± 1.3
FREA	44.8 ± 4.1	64.1 ± 2.8	69.9 ± 2.7	26.2 ± 2.3	22.6 ± 2.7	5.5 ± 0.4
RARL	50.1 ± 8.7	66.0 ± 10.0	73.2 ± 5.9	24.2 ± 5.2	20.8 ± 5.5	4.9 ± 1.0
ASP	58.1 ± 5.5	72.1 ± 5.1	80.2 ± 4.5	20.0 ± 2.9	16.6 ± 3.1	4.3 ± 0.4
SAGE	51.6 ± 5.5	68.8 ± 4.7	73.1 ± 1.5	23.3 ± 2.4	19.5 ± 2.1	5.0 ± 0.5
Co-SPAR	$\mathbf{71.5} \pm 4.3$	$\mathbf{84.1} \pm 3.8$	$\mathbf{85.3} \pm 2.9$	$\mathbf{15.4} \pm 0.6$	$\mathbf{12.9} \pm 0.1$	$\mathbf{3.7} \pm 0.7$

Table 8: Protagonist performance under all-agent policy control. All metrics are percentages. Each cell reports mean ± standard deviation across three independently trained seeds. **Bold** and underlined mark the best and second-best displayed means, respectively.

Protagonist	Overall	Task Progress		Safety Violations		
	SR ↑	GRR ↑	SRP ↑	SVR ↓	CR ↓	OR ↓
HR-PPO	60.5 ± 6.5	64.6 ± 7.6	83.4 ± 3.9	$\mathbf{5.8} \pm 1.1$	2.4 ± 0.9	$\mathbf{3.5} \pm 0.3$
FREA	66.0 ± 2.5	73.3 ± 2.9	86.3 ± 1.8	8.8 ± 0.8	4.2 ± 0.4	4.9 ± 0.5
RARL	69.5 ± 7.3	75.7 ± 8.5	87.9 ± 2.9	7.8 ± 1.0	3.6 ± 0.7	4.6 ± 0.5
ASP	69.2 ± 3.6	75.0 ± 4.4	88.4 ± 2.9	8.0 ± 1.2	3.7 ± 0.8	4.5 ± 0.4
SAGE	68.0 ± 5.9	73.8 ± 6.1	86.3 ± 2.7	7.7 ± 0.3	3.5 ± 0.4	4.5 ± 0.2
Co-SPAR	$\mathbf{82.8} \pm 1.6$	$\mathbf{88.1} \pm 1.4$	$\mathbf{93.4} \pm 1.1$	6.5 ± 0.8	$\mathbf{2.3} \pm 0.3$	4.4 ± 0.6

D.4 Cross-Strength Analysis

CDF fields and robust feasible region spectrum. Figure 7 shows the CDF fields underlying the boundaries in Figure 4(a) for the car-following interaction. States below each black boundary are feasible at that strength. As β increases, higher CDF values extend toward lower relative velocities and the boundary moves downward, illustrating the nested regions in Proposition 2.

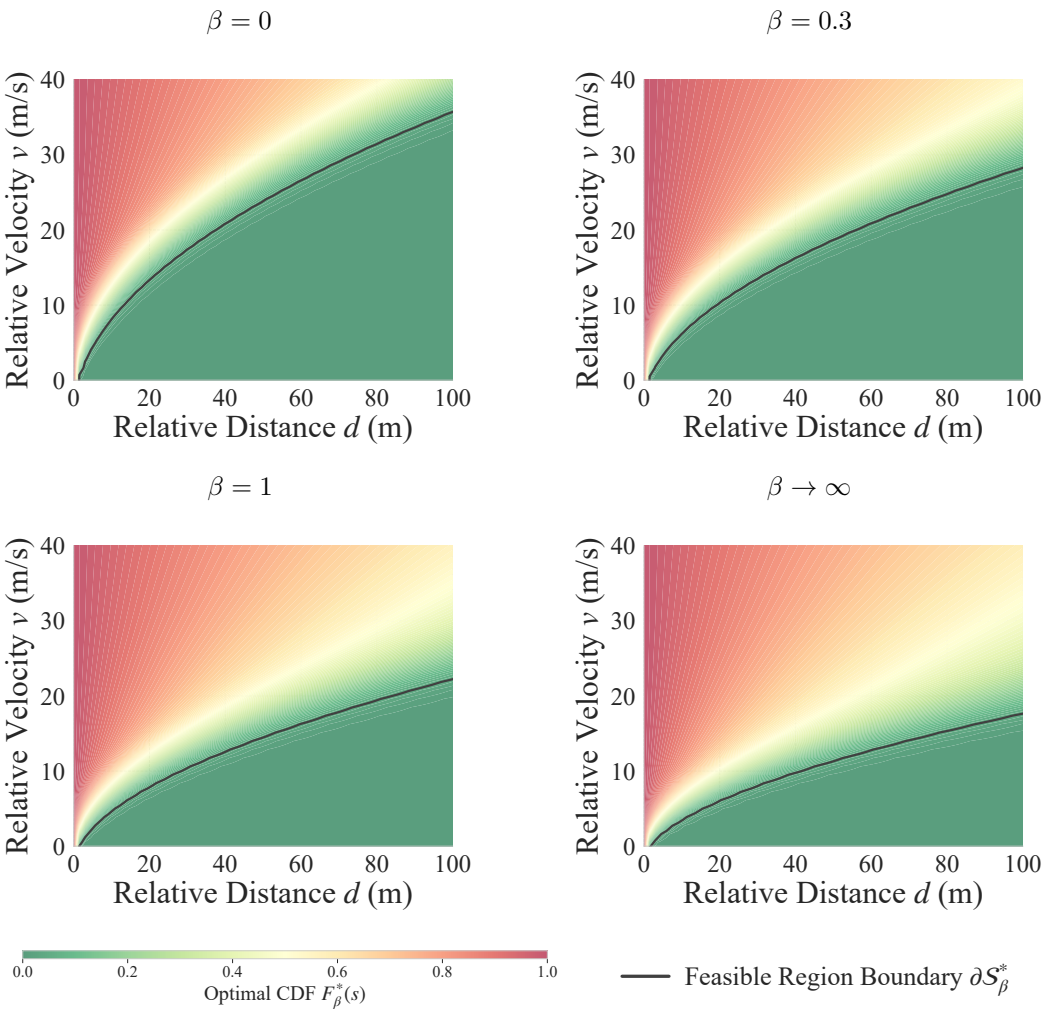


Figure 7: Optimal CDF fields and feasible region boundaries in the car-following interaction at $\beta = 0, 0.3, 1$, and $\beta \to \infty$. Relative distance and relative velocity define the two-dimensional state space, and the black line marks the feasible region boundary.

Protagonist–adversary cross-play. Figure 8 crosses protagonist training strength (rows) with adversary test strength (columns). Cell (i, j) reports the cross-play result for the protagonist policy π_{β_i} and adversary policy μ_{β_j}, trained at strengths β_i and β_j, respectively. The diagonal therefore reports matched-strength evaluations. For $\beta_i < \beta_j$, moving left from cell (j, j) to (j, i) tests the protagonist trained at the stronger adversarial strength against a weaker adversary. Cell (i, j) above the diagonal instead tests a protagonist trained at the weaker strength against a stronger adversary. The upper triangle provides the complementary stress test and shows that evaluating a protagonist against increasingly stronger adversaries raises its violation rate. For example, the strong-to-weak cell $(\infty, 0.3)$ has an SVR of 9.5%, between 7.0% for the matched weak cell and 18.9% for the matched strong cell. Figure 4(c) plots these pairings. Across the tested weaker strengths, the strong-to-weak SVR in the last row stays between the corresponding diagonal entry and the matched strong result. The empirical SVR ordering aligns with the stronger-to-weaker robust CDF ordering in Theorem 2.

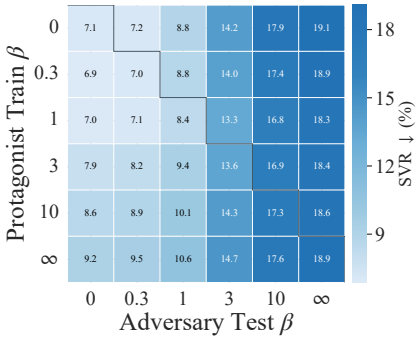


Figure 8: Protagonist–adversary cross-play across strengths (SVR, %). Rows index protagonist training strength and columns index adversary test strength. The diagonal gives matched-strength results. Below it, protagonists trained at stronger adversarial strengths face weaker adversaries, while above it, protagonists trained at weaker strengths face stronger adversaries.

Spatial CDF variation across traffic scenes. Across additional WOMD scenes, high CDF regions become more pronounced at larger β (Figure 9).

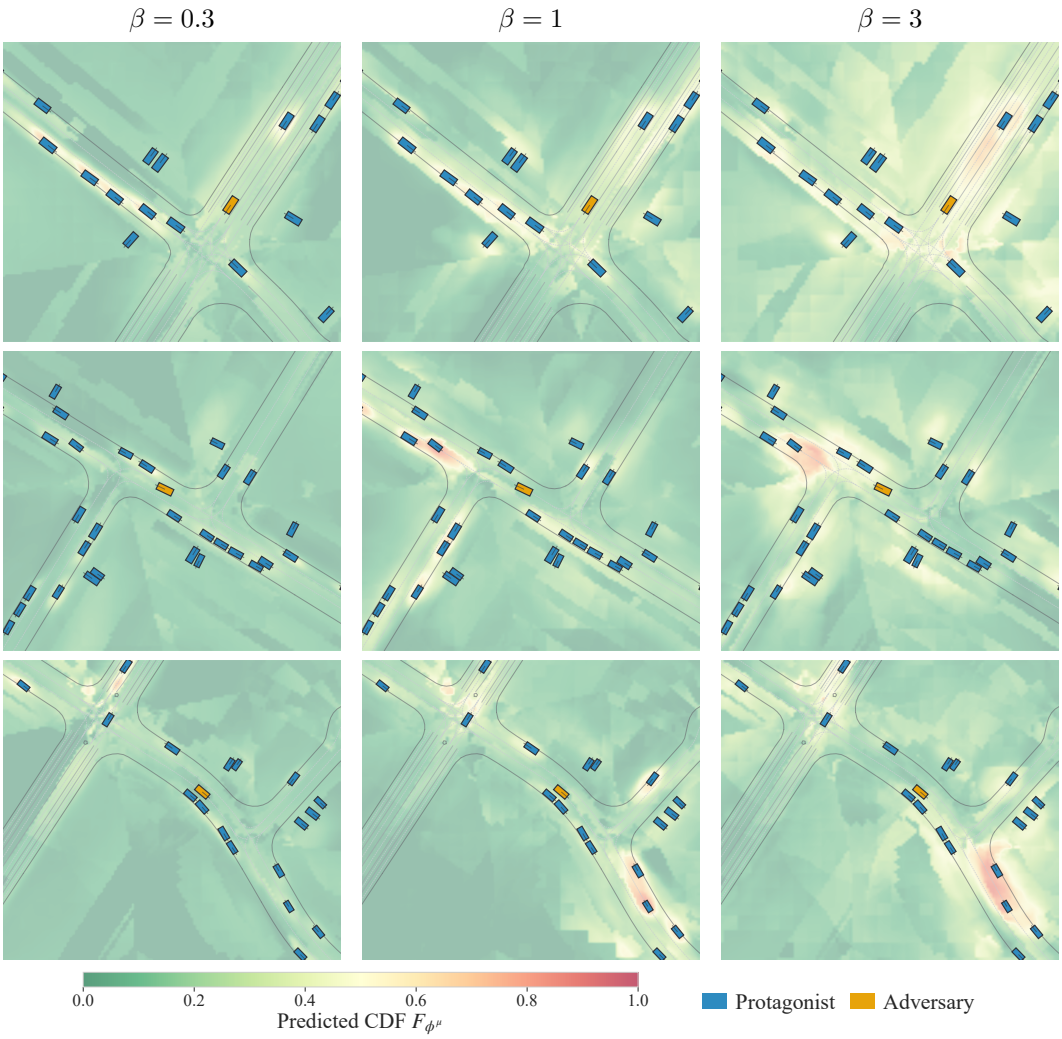


Figure 9: Spatial CDF predictions in three additional WOMD scenarios. Within each scenario, the adversarial vehicle is placed at candidate road positions with its heading aligned to the road.

Closed-loop trajectories across adversarial strengths. Each row focuses on a protagonist involved in a strong interaction with one adversary; the other vehicles appear as background traffic. The BC Trained protagonist and the Co-SPAR adversary trained at the indicated β both act in closed loop. Within each row, the columns show the same WOMD scenario at the same rollout step. Across these scenes, higher β brings the adversary closer to the protagonist, illustrating increasingly challenging closed-loop interactions (Figure 10).

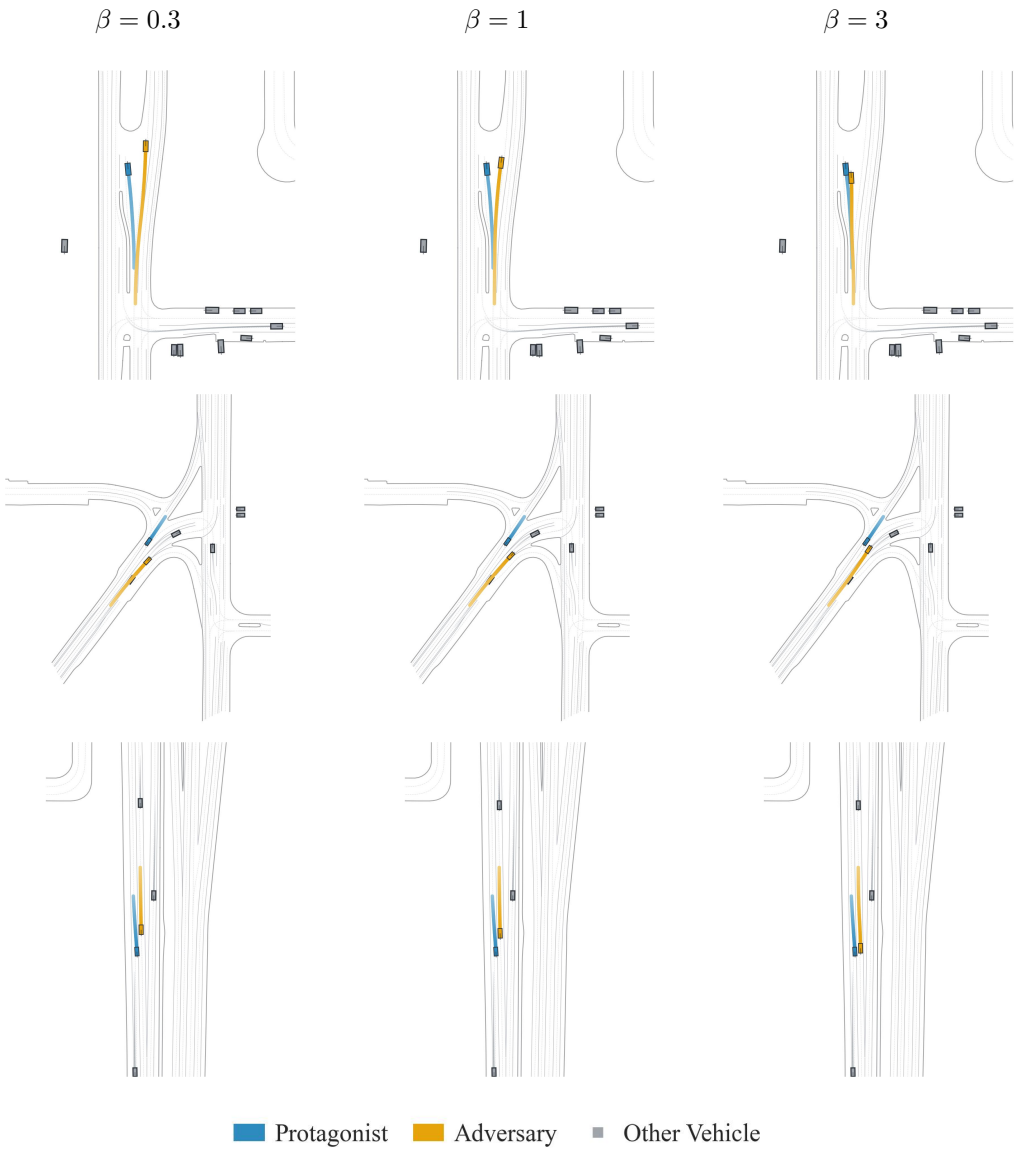


Figure 10: Trajectory visualizations in WOMD scenarios. Rows show separate scenarios; columns show $\beta = 0.3, 1, 3$ from left to right.

D.5 KL Budget Tracking

The adversary dual update uses β to control deviation from the behavioral prior. Figure 11 compares empirical adversary KL with its target at the four finite strengths during training. After the initial adjustment, the empirical KL remains near its corresponding budget at each strength.

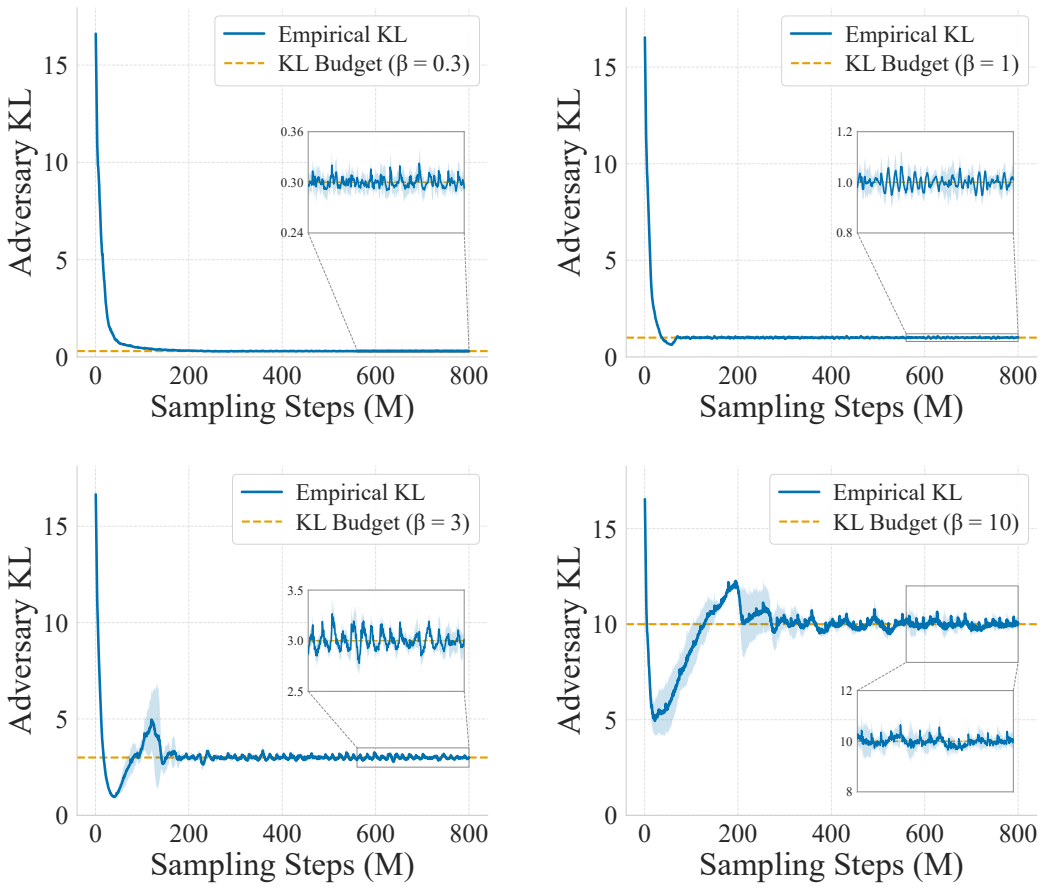


Figure 11: Empirical adversary KL during training at $\beta = 0.3, 1, 3, 10$. Solid curves show the measured KL, and dashed lines show the corresponding target budgets.

E Related Work

Adversarial scenario generation for autonomous driving. Adversarial scenario generation exposes safety weaknesses by searching over traffic behaviors, interaction states, or scenario parameters (Ding et al., 2023b). Search-based methods use importance sampling and adaptive stress testing to accelerate the discovery of rare failures (Zhao et al., 2016; Koren et al., 2018), while optimization-based methods either learn adversarial behaviors and scenario distributions (Wachi, 2019; Ding et al., 2020; Ransiek et al., 2024) or directly optimize trajectories and perturbations against a driving policy (Wang et al., 2021; Hanselmann et al., 2022). To balance criticality with plausible traffic behavior, existing methods draw on learned traffic distributions (Rempe et al., 2022), naturalistic driving data (Hao et al., 2023), causal structure (Ding et al., 2023a), feasibility constraints (Chen et al., 2024), controllable generative models (Chang et al., 2024; Ruan et al., 2026), or test-time interpolation of adversary policy weights (Nie et al., 2026a). Across these approaches, scenario generation generally treats the driving policy as fixed, so subsequent policy updates may shift its safety weaknesses beyond the scenarios generated earlier.

Closed-loop adversarial training. Closed-loop adversarial training couples challenge generation with protagonist policy learning so that the generated challenges adapt as the driving policy changes (Zhang et al., 2023). When the challenge generator is itself a policy, adversarial reinforcement learning jointly or iteratively optimizes it with the driving policy (Pinto et al., 2017; Ren et al., 2024; Ma et al., 2019; Kuutti et al., 2021). Such policy-level interaction has been structured through asymmetric self-play, where teacher and student objectives balance difficulty, solvability, and realism (Zhang et al., 2024), and through closed-loop min–max optimization, where the adversary tracks the long-tail weaknesses of an evolving policy (Nie et al., 2026b). These formulations do not characterize safety certificate transfer across adversarial strengths. Co-SPAR organizes closed-loop interaction across nested prior-constrained adversary classes defined by a behavioral prior and deviation budget, and shows that a protagonist optimal at a stronger strength retains its safety certificates against weaker classes.

Safe reinforcement learning under state constraints. Safe reinforcement learning under state constraints seeks to maximize task return while satisfying safety constraints at every step. Feasibility functions characterize the states from which infinite horizon constraints can be satisfied, and the resulting feasible regions guide policy improvement (Yu et al., 2022; Yang et al., 2026; 2023). Dual-robust safe reinforcement learning further considers adversarial disturbances in constrained zero-sum Markov games and learns policies against both safety and performance adversaries (Li et al., 2024). Co-SPAR constructs the robust CDF over the prior-constrained adversary class and uses its sublevel sets to certify safety against that class. The resulting feasible region guides region-wise protagonist improvement in closed-loop traffic self-play.