AN ANCHORED POLICY REGULARIZER FOR DEEP CONSTRAINED MULTI-AGENT REINFORCEMENT LEARNING UNDER TIGHT COST BUDGETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning teams must often keep costs like energy or collisions within budget. Standard training gives each budget a Lagrange multiplier that rises under violation and falls otherwise. At binding budgets (below what an unconstrained policy spends), policy and multiplier can cycle and deep training can collapse. It is unknown which cheap stabilizer prevents this, or whether the policy survives an adversary. We propose an anchored policy regularizer that pulls each policy toward a slow copy of itself. In the tabular robust constrained Markov game framework, the existence theorem needs a weaker equilibrium notion; we prove that the one behind our multiplier update exists under robust Slater. On cooperative navigation at tight budgets, unanchored training with our multiplier update collapses after a warm-up holding the multipliers at zero, and trains without it. The anchor removes the collapse and lifts the goal rate above every constrained variant without it, with or without warm-up. Under a bounded action-perturbing adversary retrained against each final policy, anchored variants keep most of their goal rate. Training against such an adversary adds no gain beyond seed noise at that perturbation size. The method meets every budget pair of a navigation grid without per-pair tuning, in seed-pooled mean cost and per agent in two thirds of runs. We enforce the budgets on the deployed deterministic policy.

1 INTRODUCTION

Consider a fleet of delivery robots sharing a warehouse floor. Each robot has its own battery budget and collision budget to keep, and its motors respond less precisely than its model assumes. Multi-agent reinforcement learning (MARL) answers each half separately. Constrained Markov games commonly enforce such budgets through Lagrange multipliers (Gu et al., 2021): penalty weights that rise while a budget is violated and fall while it is slack. Robust Markov games represent the model error by a learned adversary (Zhang et al., 2020). Robust constrained Markov game (RCMG) theory (Chang et al., 2026) combines both, but only for finite (tabular) state and action spaces. Its exact dynamic-programming algorithm does not scale to continuous control, and its equilibrium theory needs a weaker solution concept in general.

Whether primal–dual training reliably reaches a working deep policy at binding budgets is not known. We call a budget binding when it is below the cost measured in an unconstrained run of the same learner, whether or not its final multiplier ends positive. In that regime policy and multipliers can rotate about their joint target, as gradient play does in bilinear games (Mertikopoulos et al., 2018): the multiplier rises while the policy overspends and falls while it retreats. With deep policies, training can then end in a degenerate basin with the task unsolved: the multipliers pinned at their upper bound (the ceiling), or back near zero because the policy has retreated to cheap actions that stay within budget (Figure 1). Rotation is our working explanation of this collapse. Which cheap stabilizer, if any, prevents it is unsettled. One common candidate, PID Lagrangian control (Stooke et al., 2020), damps the multiplier alone, and we compare it with a pull on the policy.

This paper studies anchored primal–dual training for constrained Markov games at binding budgets. The method has two parts: per-agent vector multipliers that enforce the budgets, and an anchored policy regularizer that pulls each policy toward a slowly moving copy of itself. Its threat model is a

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.

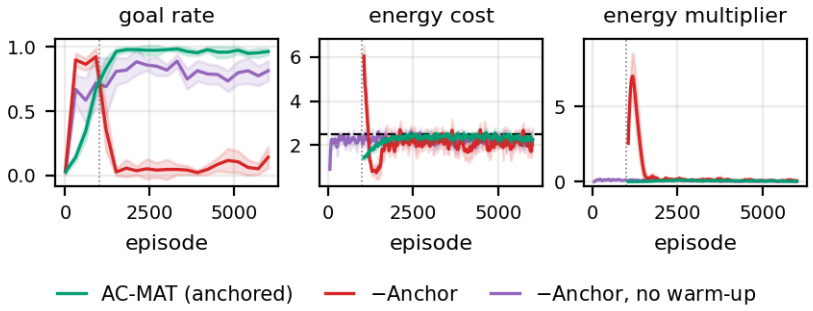

Figure 1: Released after a warm-up at zero (dotted), the unanchored run is already over the energy budget (dashed), so its multiplier spikes and its goal rate collapses. Trained without the warm-up it stays feasible at a lower goal rate. The anchored AC-MAT stays within budget with multipliers below 0.2. Navigation, budgets $(0.035, 2.5)$, five seeds, mean$\pm$s.d., 50-episode smoothing.

NatureActor per agent: a learned perturbation of that agent's own executed actions within a fixed size bound. At evaluation the attacker is a NatureActor retrained against the finished policy; the robust variant, which also trains against one, is one of the ablations. We instantiate the method on an on-policy transformer backbone, where the anchored constrained method is AC-MAT, and on an off-policy actor–critic backbone, and ask three questions of it. The first question asks whether the anchored method reaches a working policy at binding budgets, across a grid of budgets rather than at one tuned pair. A working policy keeps every cost below its budget in a mean over agents and seeds while the team still solves the task, measured by the goal rate, the fraction of agents on their goal at the end of an episode. The second question asks whether the anchor also preserves the goal rate against the retrained attacker, and whether training against a NatureActor is needed to keep it. The third question asks whether the anchor acts separably from the multipliers, the adversary, and generic stabilizers, and whether the method transfers across backbones and tasks.

Our contributions are fourfold. (1) We show on cooperative navigation that the method holds a grid of collision and energy budget pairs with no tuning per pair. This holds for costs averaged over agents and training seeds, and per agent in two thirds of runs. Tightening the budgets during training by about a fifth restores per-agent feasibility on those seeds and in most runs on five fresh ones. It is the third margin we tried, and we chose it on the seeds it is scored on. (2) We identify the anchored policy regularizer as the addition that removes the collapse of primal–dual training on cooperative navigation when the multipliers are released after a warm-up at zero. It lifts clean (adversary-free) and attacked goal rate above every constrained variant trained without it at binding budgets, including those without the warm-up. On pursuit–evasion, with no warm-up, it keeps the multipliers below their ceiling. (3) We separate the roles of the multipliers, the anchor, and training against the adversary by controlled ablation. Removing the multipliers loses feasibility: the seed-pooled mean cost leaves the budgets. Removing the anchor costs trainability (the clean goal rate) and most of retention, which is the goal rate kept under a retrained adversary. Training against the adversary at the evaluation perturbation size gives no improvement distinguishable from seed variance on navigation and discovery with the transformer backbone, and on tag with the off-policy one. On navigation no tested stabilizer without a slow copy matches the anchored method, clean or attacked. On the transfer tasks the anchored and unanchored variants stay within seed noise. (4) Separately, we refine the equilibrium-existence theorem of the tabular RCMG framework to a stationarity result. A finite instance meeting all of its assumptions has no equilibrium of the asserted kind. Under a robust Slater condition (against every opponent profile, each agent has a response whose costs stay strictly below budget under every kernel) we prove existence for a weaker equilibrium that requires feasibility and complementary slackness only at each agent's kernel minimizing reward minus weighted costs.

2 Related Work

Robust MARL supplies the threat model but lacks budgets. Constrained MARL often uses the primal–dual training the anchor stabilizes but models no uncertainty in the dynamics. Robust constrained MDPs (RCMDPs) have both, for a single agent. A fourth line supplies the backbone and stabilizer precedents.

Robust Markov games equip multi-agent games with transition or reward uncertainty. This line includes robust Q-learning and actor–critic under per-state uncertainty sets (Zhang et al., 2020) and minimax perturbation of the other agents' actions (M3DDPG; Li et al., 2019). A single-agent line uses adversarial disturbance forces (Pinto et al., 2017) and perturbation of the agent's own executed action (Tessler et al., 2019a) as physically realizable surrogates for model uncertainty. Maximum-entropy RL gives some of this robustness implicitly (Eysenbach & Levine, 2022). Further work covers equilibrium connections, sample-efficient robust equilibrium learning, and decentralized, single-agent offline, and regularized variants (Subramanian et al., 2023; Ma et al., 2023; Blanchet et al., 2023; McMahan et al., 2024; Shi et al., 2024; 2025; Gai & Shi, 2026; Farhat et al., 2026; Roch & Wang, 2026). None of them, however, enforces cost constraints. We measure what training against such an adversary adds once budgets and the anchor are in place (Section 5.3).

Constrained MARL theory covers constrained correlated equilibria (Chen et al., 2022), online safe MARL with sublinear regret (Ding et al., 2023b), constrained Markov potential games (Alatur et al., 2024; Jordan et al., 2024), anytime constraints (McMahan, 2025), Lagrangian primal–dual solvers (Das et al., 2025), and mean-field scaling (Mondal et al., 2024). Deep systems build on single-agent constrained-RL methods (Tessler et al., 2019b; Stooke et al., 2020; Achiam et al., 2017; Dalal et al., 2018; Bai et al., 2023) and scale the paradigm through value factorization (Liu et al., 2021), local trust regions (Zhang et al., 2024a), offline alignment (Guan et al., 2025), graph critics, coordination graphs, risk-aware communication, and epigraph reformulations (Du et al., 2025; Amaya-Corredor et al., 2026; Patel & Truong, 2026; Wang et al., 2026; Zhang et al., 2025). All of them, however, train under a single nominal dynamics model. Gu et al. (2021) report MACPO and MAPPO-Lagrangian meeting their budgets on Safe MAMuJoCo from unsafe initial policies while unconstrained learners violate them. Zhang et al. (2025) report Lagrangian methods training unstably at a zero violation threshold and MAPPO-Lagrangian depending on the initial value and learning rate of its multiplier. Neither trains in our regime: per-agent vector multipliers released after a warm-up onto a deep policy that violates its budgets.

RCMDPs treat both concerns for a single agent (Russel et al., 2021; Wang et al., 2022; Ghosh, 2024). Distributionally robust and epigraph extensions follow (Zhang et al., 2024b; Kitamura et al., 2025). Ma et al. (2025) exhibit an RCMDP with a strictly positive duality gap and optimize the primal directly. Bossens (2024) is the closest single-agent precursor in mechanism: it trains an explicit adversarial policy against the Lagrangian objective inside an RCMDP. Our method adds per-agent vector multipliers and an anchor; its robust variant adds interacting learner–NatureActor pairs. The temporally coupled adversary of Xu et al. (2026) is the nearest single-agent threat model. Its perturbations are coupled across time, while ours are bounded per step and per agent inside a game of strategically interacting agents. Wang et al. (2025) harden a cooperative driving team with a shared objective and multipliers but no per-agent budgets or equilibrium concept. Chang et al. (2026) extend the line to Markov games with a tabular algorithm. Section 4.2 refines its existence theorem to a stationarity result under a robust Slater condition.

Our primary instantiation, AC-MAT, builds on the Multi-Agent Transformer (MAT; Wen et al., 2022). AC-MATD3 builds on MATD3 (Ackermann et al., 2019), which pairs centralized MADDPG critics (Lowe et al., 2017) with TD3 stabilizers (Fujimoto et al., 2018). MAST (Wei et al., 2025) also equips a transformer backbone with Lagrangian constraints, but for cooperative teams under a nominal model. It lacks an adversary, per-agent vector budgets, and an equilibrium concept. We borrow regularization toward a reference policy from equilibrium learning (Perolat et al., 2021; Sokota et al., 2023; Perolat et al., 2022) as a primal stabilizer for the rotation, which min–max optimization damps with optimizer steps (Chavdarova et al., 2021; Yoon & Ryu, 2021). In single-agent constrained MDPs, regularized and optimistic primal–dual updates give last-iterate convergence (Ding et al., 2023a; Moskovitz et al., 2023). Our ablations separate the anchor from these stabilizers and from the other components (Section 5.3).

3 Preliminaries

We restate the RCMG framework of Chang et al. (2026). Full derivations are there. Its components descend from constrained (Altman & Shwartz, 2000) and robust Markov games (Zhang et al., 2020). An RCMG is a Markov game with N agents, a nominal transition kernel P over the joint action space $\mathcal{A} = \prod_i \mathcal{A}_i$, per-agent rewards R_i, J per-agent cost functions $C_{i,j}$ with budgets $c_{i,j}$ (stacked

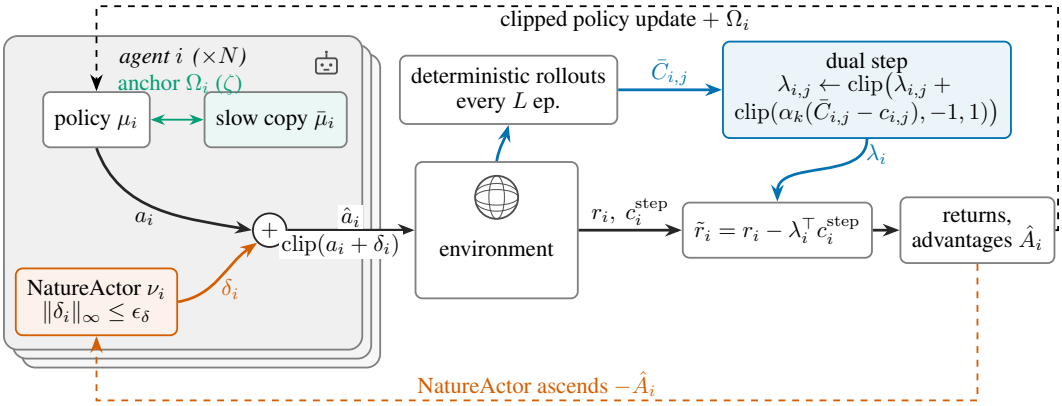

Figure 2: The on-policy method in one agent's slice: the NatureActor (orange) perturbs executed actions within ϵ_δ (3) at evaluation and, in the robust variant, in training; the dual step (blue) turns rollout costs (5) into the multipliers of the modified reward (6); the anchor term Ω_i (green) pulls toward the slow copy (4). Dashed: updates.

as $c_i \in \mathbb{R}^J$), discount $\gamma \in (0,1)$, and initial distribution ρ. Robustness enters through an $(s,\mathbf{a})$-rectangular uncertainty set $\mathcal{P} := \bigotimes_{(s,\mathbf{a})} \mathcal{P}(s,\mathbf{a})$ of transition kernels containing P. Rectangular means that the adversary may pick the next-state distribution independently at each state–action pair (Nilim & El Ghaoui, 2005). A stationary joint policy $\pi = \prod_i \pi_i$, $\pi_i : \mathcal{S} \to \Delta(\mathcal{A}_i)$, defines value V_i^π and cost-value $\Lambda_{i,j}^\pi$ functions in the standard way. Robustness is asymmetric: the agent maximizes worst-case reward while constraining worst-case cost:

$$\bar{V}_i^\pi(\rho) = \min_{P\in\mathcal{P}} \mathbf{E}_{\pi,P,\rho}\left[\sum_{t=0}^{\infty} \gamma^t R_i^t\right], \quad \bar{\Lambda}_{i,j}^\pi(\rho) = \max_{P\in\mathcal{P}} \mathbf{E}_{\pi,P,\rho}\left[\sum_{t=0}^{\infty} \gamma^t C_{i,j}^t\right]. \tag{1}$$

The ideal solution concept is the *Robust Feasible Nash Equilibrium* (RFNE): a robustly feasible joint policy π^* ($\bar{\Lambda}_{i,j}^{\pi^*}(\rho) \le c_{i,j}$ for all i,j) such that no agent i can improve its worst-case value $\bar{V}_i^{\pi_i,\pi_{-i}^*}(\rho)$ by a robustly feasible unilateral deviation.

Surrogate Lagrangian and Upper-RFNE. The exact Lagrangian takes its worst-case kernels separately for the reward and for each cost, $\bar{V}_i^L(\pi,\lambda) = \bar{V}_i^\pi(\rho) + \lambda_i^\top\big(c_i - \bar{\Lambda}_i^\pi(\rho)\big)$. In general no single kernel attains it, and it has no one-kernel Bellman form. Merging these into one unified kernel through Lagrange multipliers $\lambda_i \in \mathbb{R}_+^J$ yields the tractable *surrogate Lagrangian*:

$$\bar{V}_i^{\mathrm{SL}}(\pi,\lambda) = \lambda_i^\top c_i + \min_{P\in\mathcal{P}} \mathbf{E}_{\pi,P,\rho}\left[\sum_{t=0}^{\infty} \gamma^t \tilde{R}_i^t(\lambda_i)\right], \quad \tilde{R}_i^t(\lambda_i) := R_i^t - \lambda_i^\top C_i^t, \tag{2}$$

which upper-bounds $\bar{V}_i^L$ for all (π,λ). The *Upper-RFNE* is a profile in which (π_i,λ_i) is a saddle point of $\bar{V}_i^{\mathrm{SL}}$ for each agent i with the other agents' policies fixed. With the multiplier free over $\mathbb{R}_+^J$, such a saddle is robustly feasible and is an exact RFNE (Proposition 1 in Appendix A). The surrogate loses guaranteed existence of that saddle, because the minimum over kernels can leave a positive duality gap. Appendix A.2 gives an instance with no saddle.

Primal–dual learning dynamics. Simultaneous gradient steps on a policy and its multiplier form a min–max dynamics. For bilinear zero-sum games with an interior equilibrium, the continuous-time idealization of these steps conserves a Bregman energy and is Poincaré recurrent: it cycles around the saddle rather than converging to it (Mertikopoulos et al., 2018). We take this recurrence as the model of the rotation. Regularization toward a reference strategy converts this recurrence into convergence in entropy-regularized zero-sum games (Perolat et al., 2021; 2022).

4 Methodology: Anchored Primal–Dual Actor–Critic for Constrained Markov Games

Exact robust Q-iteration and analytic dual subgradients, on which the tabular RCMG algorithm relies, are unavailable in continuous domains. Our method replaces each with a deep counterpart, neural

critics under centralized training and per-agent multipliers updated from observed violations, and adds the anchor. We instantiate it twice. On the on-policy Multi-Agent Transformer the anchored constrained method is AC-MAT and its robust variant is RC-MAT, which also trains a NatureActor. On off-policy MATD3 they are AC-MATD3 and RC-MATD3 (Section 4.3). Every component is indexed by agent and cost dimension, so the method applies unchanged to N agents and J per-agent constraints (Figure 2, one agent's slice).

4.1 The threat model and the games played

Each agent i has a *NatureActor* ν_i. It replaces $\min_{P\in\mathcal{P}}$ in (2) by a realizable action-shift adversary, so $\mathcal{P}$ is never enumerated. It maps the state to an additive perturbation δ_i of the agent's selected action. Off-policy the perturbation is deterministic, $\delta_i = \nu_i(s)$, whereas on-policy it is sampled, $\delta_i \sim \nu_i(\cdot \mid s)$. Its ℓ_∞-norm is bounded by ϵ_δ. Here a_i is agent i's selected action: the mean of its stochastic policy at evaluation and a sample from it during training rollouts. The executed action is

$$\hat{a}_i = \mathrm{clip}\big(a_i + \delta_i,\ \mathcal{A}_i\big), \quad \|\delta_i\|_\infty \le \epsilon_\delta. \tag{3}$$

Retrained from scratch against a frozen policy, a NatureActor is the attacker of Section 5.3. Trained with the policies, it is the robust variant's adversary. On-policy it ascends a clipped surrogate of the negated advantage. Off-policy it descends the agent's critic.

4.2 The equilibrium behind the dual update

The dual update is modeled on an object weaker than the saddle of Section 3: it requires feasibility only under the kernel that minimizes the reward-minus-cost objective (Table 1). Write $J_{C,i,j}(\pi, P)$ for the cost under a fixed kernel P.

Definition 1 (Lagrangian-stationary Upper-RFNE)**.** *A profile (π^*, λ^*) with one kernel P_i^* per agent in the per-state-action convex hull of $\mathcal{P}$ is* Lagrangian-stationary *if for every agent i: (1) π_i^* maximizes $\bar{V}_i^{\mathrm{SL}}(\cdot, \pi^*_{-i}, \lambda_i^*)$; (2) P_i^* attains the minimum defining the surrogate Lagrangian (2) at (π^*, λ_i^*); (3) $J_{C,i,j}(\pi^*, P_i^*) \le c_{i,j}$ and $\lambda^*_{i,j}\,(c_{i,j} - J_{C,i,j}(\pi^*, P_i^*)) = 0$ for all j (feasibility and complementary slackness at P_i^*).*

Table 1: Only the Lagrangian-stationary notion is guaranteed to exist (tabular, robust Slater); columns: where feasibility is evaluated, the multiplier-side condition, existence.

notion	feasibility evaluated	multiplier side	existence
exact RFNE	worst case over $\mathcal{P}$	hard constraints	not established
Upper-RFNE (saddle)	worst case over $\mathcal{P}$ (Prop. 1)	saddle minimality	can fail (App. A.2)
Lagr.-stationary (Def. 1)	minimizing kernel P_i^*	KKT stationarity	Theorem 1 (tabular, robust Slater)

Under the conditions below, Appendix A proves that this object exists.

Theorem 1 (existence)**.** *Let $\mathcal{S}$ and every $\mathcal{A}_i$ be finite, rewards and costs bounded by $\bar{R}$, and each $\mathcal{P}(s, \mathbf{a})$ compact, and let robust Slater hold: for some $\xi > 0$, every agent has against every π_{-i} a response whose robust costs stay at least ξ below every budget. Then a Lagrangian-stationary Upper-RFNE exists, with kernels in the per-state-action convex hull of $\mathcal{P}$ and $\sum_j \lambda^*_{i,j} \le 2\bar{R}/((1-\gamma)\xi)$ for every agent.*

It upgrades to the saddle point and the exact RFNE when one kernel realizes both worst cases within budget (Theorem 2 in Appendix A).

The theory reaches the two deep methods only through tabular versions of the games they play. The deep methods' properties are measured, and Table 2 lists the system to which each main result applies. AC-MAT trains on the nominal dynamics, the special case $\mathcal{P} = \{P\}$, and its dual update targets the multiplier conditions of Definition 1 for that case. There one kernel realizes both worst cases within budget, so by Theorem 2 every Lagrangian-stationary point is a saddle point and an exact RFNE.

The robust variant trains against every NatureActor at once, a game of $2N$ players (N policies and N NatureActors). Because shifting an agent's action changes its reward, cost, and next state together, this adversary is not one of the kernel uncertainty sets of Section 3, and Theorem 1 does

Table 2: The system each main claim is established for, and the kind of evidence.

statement	status and system
a Lagrangian-stationary Upper-RFNE exists under robust Slater; an Upper-RFNE need not	proved, finite tabular RCMG (App. A)
the robust variant's game has an equilibrium, approximately Lagrangian-stationary if adversaries see the joint action	proved, its tabular version under uniform Slater (Theorems 3, 4)
the anchor preserves fixed points and makes the flow asymptotically stable	proved in a bilinear saddle model (App. B.6)
the duals track a grid of budgets; the anchor raises the clean goal rate	measured, AC-MAT on navigation (Sections 5.2, 5.3)
adversarial training at the evaluation perturbation size adds no separable retention margin	measured, fifteen paired seeds (Section 5.3)
the anchor damps a primal–dual rotation in the deep system	conjectured, consistent with the ablations

not cover it. Appendix A treats the tabular version of this game, in which each adversary picks its shift from a finite menu. Under a uniform Slater condition an equilibrium of that game exists with feasibility, complementary slackness, and the same multiplier bound (Theorem 3). If each adversary also observes the joint action, such an equilibrium is approximately Lagrangian-stationary in the robust game these menus generate, with an error set by how often the other agents' shifts change their executed actions (Theorem 4).

4.3 The anchored primal–dual update

Anchored policy regularizer. The anchor pulls each policy toward a slow copy $\bar{\mu}_i$ of itself with weight $\zeta \geq 0$, intended to damp the rotation. For each action coordinate we divide the difference between the current and slow-copy action means by its bound $a_{i,l}^{\max}$, average the squared differences over coordinates and states, and weight by ζ, so ζ is free of the action scale:

$$\Omega_i \;=\; \zeta\, \mathbf{E}\Big[\tfrac{1}{d_i}\textstyle\sum_l \big((\mu_{i,l}(s) - \bar{\mu}_{i,l}(s))/a_{i,l}^{\max}\big)^2\Big], \tag{4}$$

with $\mu_i(s)$ the action mean (sampled actions and the log-probability path are untouched), $\bar{\mu}_i$ its slow copy, and d_i the number of action coordinates. This copy is the target network off-policy and a parameter-space exponential moving average (EMA) of the policy on-policy. The anchor is a heuristic stabilizer, the deterministic-policy analogue of the reference regularization of Section 3. It preserves fixed points: the term and its gradient vanish wherever $\mu_i = \bar{\mu}_i$. It is also purely primal: it leaves the modified reward, the multipliers, and the dual update untouched. A bilinear saddle model with the same one-sided anchoring keeps both properties (Appendix B.6). There, for every $\zeta > 0$ and every rate $\tau > 0$ at which the copy tracks the policy, the flow is asymptotically stable with every eigenvalue negative in real part, whereas without the anchor it only rotates.

Per-agent Lagrangian update. Each agent i maintains a multiplier vector λ_i with one component per cost dimension, the multiplier structure of MAPPO-Lagrangian (Gu et al., 2021) and of the RCMG framework, and keeps it in the box $[\underline{\lambda}, \bar{\lambda}]^J$. Every L episodes it takes a projected subgradient-style step on a rollout estimate with a polynomially decaying step size (Appendix B.5). The inner clip bounds the increment per dimension and the outer clip projects the result onto the box:

$$\lambda_{i,j} \;\leftarrow\; \mathrm{clip}\Big(\lambda_{i,j} + \mathrm{clip}\big(\alpha_k\,(\bar{C}_{i,j} - c_{i,j}),\, -1, 1\big),\; \underline{\lambda},\; \bar{\lambda}\Big), \quad \alpha_k = \frac{\alpha_0}{\sqrt{1+k}}, \tag{5}$$

where $\bar{C}_{i,j}$ is agent i's mean episodic cost on dimension j over the M latest deterministic rollouts and k counts dual updates since the warm-up. This estimates the cost of the deployed deterministic policy rather than the worst-case value $\bar{\Lambda}_{i,j}$ of Section 3. The dual step therefore uses rollouts of the game actually played (Section 4.1): clean for AC-MAT and with the training-time adversary for RC-MAT. A warm-up holds the multipliers at zero so the policy learns the task first. Their release starts the updates inside the box. The floor $\underline{\lambda}$ is 0.01 on the off-policy navigation and tag runs, which leaves a residual penalty on slack dimensions, and 0 elsewhere.

Following (2), the modified reward fed to the critic is:

$$\tilde{r}_i = r_i - \lambda_i^\top c_i^{\text{step}} = r_i - \sum_{j=1}^{J} \lambda_{i,j}\, c_{i,j}^{\text{step}}, \tag{6}$$

Table 3: At the tight cell $(0.035, 2.5)$ the anchor, not the adversary, lifts goal rate and retention above every unanchored constrained variant. Mean$_{\pm \text{s.d.}}$, fifteen seeds (bottom three rows: five); all agents: seeds with every agent within both budgets on 500 held-out episodes; †: over budget; bold: anchored constrained variants. Variant labels are defined in Section 5.1.

	clean				retrained attack
Variant	collision	energy	goal rate	all agents	goal rate
AC-MAT (ours)	$0.026_{\pm 0.008}$	$2.32_{\pm 0.13}$	$\mathbf{0.947_{\pm 0.020}}$	14/15	$\mathbf{0.645_{\pm 0.051}}$
+Nature (= RC-MAT)	$0.025_{\pm 0.006}$	$2.28_{\pm 0.13}$	$\mathbf{0.950_{\pm 0.023}}$	12/15	$\mathbf{0.664_{\pm 0.057}}$
−Anchor	$0.025_{\pm 0.011}$	$2.14_{\pm 0.27}$	$0.123_{\pm 0.069}$	10/15	$0.025_{\pm 0.018}$
−Anchor +Nature	$0.019_{\pm 0.011}$	$2.06_{\pm 0.31}$	$0.082_{\pm 0.051}$	13/15	$0.023_{\pm 0.015}$
−Anchor +PID	$0.012_{\pm 0.007}$	$2.26_{\pm 0.18}$	$0.751_{\pm 0.109}$	13/15	$0.277_{\pm 0.099}$
−Anchor, no warm-up	$0.014_{\pm 0.008}$	$2.20_{\pm 0.13}$	$0.743_{\pm 0.097}$	15/15	$0.274_{\pm 0.087}$
−Anchor +PID, no warm-up, tuned	$0.009_{\pm 0.004}$	$2.41_{\pm 0.12}$	$0.777_{\pm 0.076}$	1/5	$0.381_{\pm 0.071}$
−Lagrangian +Nature	$0.042_{\pm 0.014}$†	$5.07_{\pm 0.20}$†	$1.000_{\pm 0.000}$	0/5	$0.999_{\pm 0.001}$
−Both (= MAT)	$0.055_{\pm 0.041}$†	$27.22_{\pm 3.26}$†	$0.955_{\pm 0.023}$	0/5	$0.918_{\pm 0.020}$

where $c_i^{\text{step}} \in \mathbb{R}^J$ is the per-step cost vector. The policy-independent constant $\lambda_i^\top c_i$ of (2) drops out of the gradient. Off-policy $\tilde{r}_i$ is assembled at update time under the current λ_i. On-policy it is written at collection under that iteration's λ_i.

Equation (5) is an integral controller: violation raises λ_i and the penalty in (6), and slack lets λ_i decay. Without the anchor this correction does not stop the joint (λ_i, μ_i) dynamics from settling into either degenerate basin: the multipliers saturated at their ceiling on pursuit–evasion, or near zero with the task unsolved on navigation at $(0.035, 2.5)$ after the warm-up (Figure 1). Both are consistent with an undamped primal–dual rotation, and on navigation a lagged multiplier–cost correlation supports it (Appendix C.7).

AC-MAT is built on MAT (Wen et al., 2022), an on-policy encoder–decoder sequence model updated with a PPO-style clipped objective. Because the modified reward (6) goes into the rollout buffer, λ_i enters via each agent's advantage terms alone, and heterogeneous multipliers coexist within one set of shared parameters. The multipliers follow (5), and the anchor (4) joins the policy objective. The robust variant, RC-MAT, adds a per-agent NatureActor network producing a $\tanh$-bounded stochastic perturbation $\delta_i \sim \nu_i(\cdot \mid s)$ of its agent's executed action. It trains on the same rollouts by the same clipped objective and ascends the negated λ-modified advantage $\hat{A}_i^\nu = -\hat{A}_i$. Algorithm 1 in Appendix B gives the full procedure.

The same method on Multi-Agent TD3 (Ackermann et al., 2019) yields AC-MATD3, and with training-time NatureActors RC-MATD3. Per-agent Bellman targets are separated so that each agent's backup sees its own perturbation alone. Because the off-policy instantiation undertrains at the navigation budgets of Section 5 within the same training length, it runs at looser pairs (Appendix C.5).

5 Experiments

Section 5.2 answers the first question of the introduction and Section 5.3 the second and third, on AC-MAT and its ablations with RC-MAT among them. Appendix B holds the protocol, hyperparameters, and drift check. Appendix C has the off-policy instantiation, which tracks its budgets but undertrains, together with further backbones, external baselines, multi-agent MuJoCo, and the stabilizer comparison. Appendix D has the remaining results, one subsection per question or environment.

5.1 Protocol

Environments. Our main domain is cooperative *navigation* from the VMAS benchmark (Bettini et al., 2022). Here $N{=}4$ holonomic agents must each reach a private goal while sharing the arena under a $J{=}2$ vector cost. The *collision* cost is the summed body-overlap depth and the *energy* cost the squared control effort per step. With the built-in collision reward penalty disabled, the constraint alone enforces avoidance. The NatureActor perturbs executed actions within the perturbation size ϵ_δ, set to 10% of the per-dimension action bound. As a competitive counterpart we use our own two-agent *pursuit–evasion* game with walls, zone-occupancy costs, and independent asymmetric budgets. Transfer tests use the VMAS *discovery* task, the MPE *tag* benchmark, and two multi-agent

Table 4: Seed-pooled clean mean cost stays inside both budgets in all nine cells, less often per seed and per agent (AC-MAT). Five seeds per cell, fifteen at the tight and mid cells; seeds: agent-mean feasible on 50 episodes. All agents and UCB (4×5 agent–seed pairs within marginal one-sided 95% bounds): first five seeds, 500 held-out episodes.

budgets (coll, en)	collision	energy	goal rate	seeds	all agents	UCB
(0.030, 2.5)	$0.023_{\pm0.008}$	$2.26_{\pm0.11}$	$0.954_{\pm0.020}$	4/5	3/5	8/20
(0.030, 3.0)	$0.027_{\pm0.005}$	$2.66_{\pm0.17}$	$0.980_{\pm0.009}$	3/5	1/5	5/20
(0.030, 3.5)	$0.027_{\pm0.009}$	$3.10_{\pm0.15}$	$0.994_{\pm0.005}$	4/5	2/5	3/20
(0.035, 2.5)	$0.026_{\pm0.008}$	$2.32_{\pm0.13}$	$0.947_{\pm0.020}$	12/15	5/5	16/20
(0.035, 3.0)	$0.027_{\pm0.009}$	$2.64_{\pm0.15}$	$0.978_{\pm0.012}$	11/15	3/5	6/20
(0.035, 3.5)	$0.028_{\pm0.008}$	$3.17_{\pm0.12}$	$0.990_{\pm0.008}$	4/5	2/5	8/20
(0.039, 2.5)	$0.028_{\pm0.005}$	$2.32_{\pm0.21}$	$0.955_{\pm0.016}$	4/5	5/5	16/20
(0.039, 3.0)	$0.025_{\pm0.005}$	$2.59_{\pm0.09}$	$0.984_{\pm0.009}$	5/5	5/5	18/20
(0.039, 3.5)	$0.032_{\pm0.010}$	$3.19_{\pm0.11}$	$0.986_{\pm0.012}$	4/5	4/5	16/20

MuJoCo tasks. Budgets are set once from the costs of unconstrained calibration runs, not per run. The warm-up lasts the first 1000 episodes. A *cell* is one budget pair, (collision, energy) on navigation, and a *variant* one method or ablation trained at it. The tight cell $(0.035, 2.5)$ and the mid cell $(0.035, 3.0)$ are the tightest and middle energy levels of the 3×3 grid of Table 4; off-policy runs use looser pairs.

Metrics and conditions. Evaluation runs deterministic policies on freshly sampled episodes from an environment seeded separately from training, 50 per run and 500 more held-out episodes per seed behind the per-agent counts. Like the dual, the budgets therefore bind the deterministic policy. Under sampled execution RC-MAT spends about three times its deterministic energy on the single-cost task, above every budget there, while the ported methods below regulate their sampled policy instead (Appendix D.13). *Feasible in mean* means that the episode-averaged cost (averaged over the agents) is within budget on every dimension, per seed or seed-pooled as stated; all-agent feasible means that every agent's cost is. We score every method *clean* and under the *retrained attack*: a fresh NatureActor trained against the frozen policy at the perturbation size ϵ_δ, minimizing $r - \lambda^\top c^{\text{step}}$ under the frozen final multipliers. Retention is the goal rate kept under this attack, and the *retention ratio* divides it by the clean goal rate. We report endpoints of fixed training lengths, with costs as discounted episodic sums.

Baselines and ablations. Our primary comparison is a controlled ablation: variants share backbone, network sizes, and training length and differ only in the components of interest. Relative to AC-MAT, +Nature adds the training-time NatureActor, −Anchor removes the anchor, −Lagrangian the multipliers, and −Both both multipliers and anchor (= the plain backbone); +PID replaces the dual step by the PID rule of Stooke et al. (2020); labels combine. We know no published deep method pairing an adversary with per-agent constraints in Markov games. On navigation with the stock reward and a single energy cost (Appendix C.4), MAPPO-Lagrangian and MACPO stay below a goal rate of 0.12 with $8.3\times$ our environment steps. Only MAPPO-Lagrangian with a PID multiplier trains, reaching 0.835 at the energy budget 8.7 and 0.909 at 12.4. RC-MAT reaches 0.994 and 0.992 there, with an energy of 4.6–4.8 under even the tightest budget (6.2). The ports regulate their sampled policy under undiscounted accounting, so we rest the comparison on the within-backbone ablations.

5.2 The duals track a grid of binding budgets

AC-MAT solves the tight cell inside both budgets, where the plain backbone violates both (Table 3). It completes $0.947_{\pm0.020}$ of goals over fifteen seeds, within one s.d. of that backbone. RC-MAT reaches a similar 0.950. At 12,000 episodes instead of 6000, AC-MAT's goal rate falls to 0.842 with its multipliers at 0.01, every seed feasible and retention kept, still ahead of the PID and no-warm-up variants (0.763 and 0.705). At a fixed pair with binding energy (0.064, 2.5) AC-MAT completes 0.92–0.95 of goals from $N{=}2$ to $N{=}8$ inside the energy budget at every size (Appendix D.2).

All nine fixed budget pairs, each trained as a separate policy with the same method hyperparameters, are feasible in seed-pooled mean cost (Table 4). Realized energy follows its budget across the grid, and collision stays at 0.023–0.032. Per seed and per agent, however, this weakens: 30 of 45 grid policies keep every agent within budget (all-agents column). That count uses 500 held-out episodes and the seeds column 50, so neither test implies the other: three of fifteen tight-cell seeds pass only the per-agent one (Table 3). The unanchored no-warm-up variant is all-agent feasible in 44 of 45, but at a goal rate 0.13–0.26 lower. A training margin, every cell trained at $0.80\times$ its collision and $0.83\times$

its energy budget, takes AC-MAT to 45 of 45 at a goal rate of 0.933. On five fresh seeds it gives 41 of 45, and at its own milder margin the no-warm-up variant reaches 45 of 45 at 0.781.

5.3 Retention and trainability follow the anchor

At the tight cell every unanchored constrained variant trails AC-MAT in clean goal rate (0.947), with or without the warm-up (Table 3). The −Anchor variant reaches 0.743 without the warm-up and collapses to 0.123 with it, once its multipliers are released onto an overspending policy (Figure 1).

Pursuit–evasion has no warm-up, yet unanchored training diverges there (off-policy backbone, floor 0). Without the anchor the pursuer and evader of the cell with equal budgets run at $17\times$ and $65\times$ their budgets in seed mean over frozen final policies, with one of five seeds feasible; at $\zeta{=}10$ the multipliers stay off their ceiling and four of five seeds are feasible on both sides (Appendix C.7).

No tested stabilizer without a slow copy closes the tight-cell gap: the PID multiplier update leaves the unanchored variant 0.17–0.20 below the anchored variants, and two extragradient variants hit the ceiling or diverge (Appendix C.7). Added to MAPPO-Lagrangian, the anchor itself gains nothing (Appendix D.13). A Gaussian-KL pull toward the same slow copy trains and retains at least as well at the tight cell (0.987 clean, 0.859 retained) but is feasible in two of five seeds, so the action-space form (4) is the more reliably feasible one.

Among the constrained variants retention follows the anchor (Table 3, last column): at the tight cell the anchored ones keep a ratio of 0.68–0.70 of their clean goal rate under the retrained attack, against 0.37–0.49 for the unanchored ones that train. Training against the adversary at the evaluation perturbation size ($\epsilon_\delta = 0.1$) adds nothing the paired fifteen-seed test can resolve: that test rules out a retention gain above 0.055 in goal rate but cannot rule out a smaller one. A larger training perturbation raises retention but uses up budget slack, and fewer seeds stay feasible.

At the tight cell the retrained attack leaves the costs inside both budgets, AC-MAT's collision and energy at $0.62\times$ and $0.64\times$ budget. Adversaries retrained to maximize a cost instead of the Lagrangian breach them (Appendix D.4): collision- and energy-directed attacks drive AC-MAT to $2.22_{\pm 0.41}\times$ and $1.46_{\pm 0.04}\times$ the budget each targets, and RC-MAT to $2.11_{\pm 0.17}\times$ and $1.42_{\pm 0.09}\times$.

On none of the transfer tasks does the unanchored variant collapse while the anchored one trains, nor does the anchor beat it beyond seed noise (Appendices D.5, D.6, and C.6). The multipliers hold at the slack discovery budgets and in four of five seeds for RC-MAT on Walker2d and for AC-MAT and −Anchor on HalfCheetah, while on tag the pooled mean energy exceeds its budget.

6 Conclusion

Anchored primal–dual training holds binding budgets on the deployed deterministic policy: in seed-pooled mean across a 3×3 grid, and per agent in two thirds of runs, or nearly every run with a training margin of about a fifth. Multipliers deliver seed-pooled feasibility. At the tight cell the anchor, not the adversary, carries the clean goal rate and retention, and adversarial training at the evaluation perturbation size adds no separable gain. Cost-directed attacks, which lie outside what the dual update targets, breach the budgets with and without adversarial training. No tested stabilizer without a slow copy matches the anchor in clean goal rate and retention. The anchor helps on navigation and pursuit–evasion and not beyond them. Its gain needs budgets that bind: it is largest when a warm-up releases the multipliers onto a policy that already overspends, and persists against unanchored training without one. Budget tracking transfers to the off-policy backbone except on tag. In the tabular framework the existence theorem needs a weaker equilibrium notion, and we prove that the Lagrangian-stationary one behind the dual update exists under robust Slater. For practice, check feasibility per agent and per seed, since pooled means hide the violations a training margin removes.

Reproducibility statement

All environments are public benchmark scenarios except pursuit–evasion, which Appendix B.4 specifies. On navigation the harness disables the built-in collision penalty, sets every goal radius to the agent radius, and adds the cost channels and energy tank of Appendix B.3. Every hyperparameter, threshold, and protocol choice is in Appendix B, including the drift check. Each study states its own

settings. Tables report per-seed statistics with sample standard deviations and per-seed feasibility counts. An anonymized code repository with training, attack, and evaluation scripts accompanies the submission.

AI USE STATEMENT

The research program, questions, algorithmic design, and evidentiary standards are the authors'; the problem setting comes from the published tabular framework restated in Section 3. Generative AI assisted with formulating mathematical claims and drafting their proofs (Appendix A), designing experiments, implementing methods, interpreting results, drafting sections of the text, retrieving and screening related work, analysis code, figures, and reference formatting; we generated no synthetic data. We checked every proof, number, and citation against its source and take responsibility for the content.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Johannes Ackermann, Volker Gabler, Takayuki Osa, and Masashi Sugiyama. Reducing overestimation bias in multi-agent domains using double centralized critics, 2019. Presented at the NeurIPS 2019 Deep Reinforcement Learning Workshop.

Pragnya Alatur, Giorgia Ramponi, Niao He, and Andreas Krause. Provably learning Nash policies in constrained Markov potential games. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, 2024.

Eitan Altman and Adam Shwartz. Constrained Markov games: Nash equilibria. In Jerzy A. Filar, Vladimir Gaitsgory, and Koichi Mizukami (eds.), *Advances in Dynamic Games and Applications*, pp. 213–221, Boston, MA, 2000. Birkhäuser Boston. ISBN 978-1-4612-1336-9.

Santiago Amaya-Corredor, Miguel Calvo-Fullana, and Anders Jonsson. Coordination graphs for constrained multi-agent reinforcement learning. In *Reinforcement Learning Conference (RLC)*, 2026.

Qinbo Bai, Amrit Singh Bedi, and Vaneet Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via conservative natural policy gradient primal-dual algorithm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Dimitri Bertsekas. *Convex optimization algorithms*. Athena Scientific, 2015.

Matteo Bettini, Ryan Kortvelesy, Jan Blumenkamp, and Amanda Prorok. VMAS: A vectorized multi-agent simulator for collective robot learning. In *Proceedings of the 16th International Symposium on Distributed Autonomous Robotic Systems (DARS)*, 2022.

Jose Blanchet, Miao Lu, Tong Zhang, and Han Zhong. Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 66845–66859. Curran Associates, Inc., 2023.

David M. Bossens. Robust Lagrangian and adversarial policy gradient for robust constrained Markov decision processes. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pp. 1227–1239, 2024.

Ningkang Chang, Chenyu Xu, Ziying Jia, Yue Wang, and Sihong He. Robust constrained Markov games: Multi-agent decision-making under model uncertainty and constraints. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2026.

Tatjana Chavdarova, Matteo Pagliardini, Sebastian U. Stich, François Fleuret, and Martin Jaggi. Taming GANs with lookahead-minmax. In *International Conference on Learning Representations*, 2021.

Ziyi Chen, Shaocong Ma, and Yi Zhou. Finding correlated equilibrium of constrained Markov game: A primal-dual approach. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 25560–25572. Curran Associates, Inc., 2022.

Gal Dalal, Krishnamurthy Dvijotham, Matej Vecerik, Todd Hester, Cosmin Paduraru, and Yuval Tassa. Safe exploration in continuous action spaces. *arXiv preprint arXiv:1801.08757*, 2018.

Soham Das, Santiago Paternain, Luiz F. O. Chamon, and Ceyhun Eksin. The Lagrangian method for solving constrained Markov games. *arXiv preprint arXiv:2503.10561*, 2025.

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Alejandro Ribeiro. Last-iterate convergent policy gradient primal-dual methods for constrained MDPs. In *Advances in Neural Information Processing Systems*, volume 36, 2023a.

Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo R. Jovanović. Provably efficient generalized Lagrangian policy optimization for safe multi-agent reinforcement learning. In *Proceedings of the 5th Annual Learning for Dynamics and Control Conference (L4DC)*, volume 211, pp. 315–332. PMLR, 2023b.

Haikuo Du, Fandi Gou, and Yunze Cai. Scalable safe multi-agent reinforcement learning for multi-agent system. *arXiv preprint arXiv:2501.13727*, 2025.

Benjamin Eysenbach and Sergey Levine. Maximum entropy RL (provably) solves some robust RL problems. In *International Conference on Learning Representations (ICLR)*, 2022.

Zain Ulabedeen Farhat, Debamita Ghosh, George K. Atia, and Yue Wang. Sample-efficient distributionally robust multi-agent reinforcement learning via online interaction. In *International Conference on Learning Representations (ICLR)*, 2026.

Arlington M. Fink. Equilibrium in a stochastic n-person game. *Journal of Science of the Hiroshima University, Series A-I*, 28(1):89–93, 1964.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Jingchu Gai and Laixi Shi. Taming the curses of multiagency in robust Markov games with large state space through linear function approximation. *arXiv preprint arXiv:2605.03125*, 2026.

John Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (eds.), *Bayesian Statistics 4*, pp. 169–193. Oxford University Press, 1992.

Arnob Ghosh. Sample complexity for obtaining sub-optimality and violation bound for distributionally robust constrained MDP. In *First Reinforcement Learning Safety Workshop*, 2024.

Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. Adversarial policies: Attacking deep reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2020.

Shangding Gu, Jakub Grudzien Kuba, Munning Wen, Ruiqing Chen, Ziyan Wang, Zheng Tian, Jun Wang, Alois Knoll, and Yaodong Yang. Multi-agent constrained policy optimisation. *arXiv preprint arXiv:2110.02793*, 2021.

Cong Guan, Tao Jiang, Yi-Chen Li, Zongzhang Zhang, Lei Yuan, and Yang Yu. Constraining an unconstrained multi-agent policy with offline data. *Neural Networks*, 186, 2025.

Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Juntao Dai, and Yaodong Yang. Safety-Gymnasium: A unified safe reinforcement learning benchmark. In *Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track*, 2023.

Philip Jordan, Anas Barakat, and Niao He. Independent learning in constrained Markov potential games. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4024–4032. PMLR, 02–04 May 2024.

Toshinori Kitamura, Tadashi Kozuno, Wataru Kumagai, Kenta Hoshino, Yohei Hosoe, Kazumi Kasaura, Masashi Hamaya, Paavo Parmas, and Yutaka Matsuo. Near-optimal policy identification in robust constrained Markov decision processes via epigraph form. In *International Conference on Learning Representations (ICLR)*, 2025.

Shihui Li, Yi Wu, Xinyue Cui, Honghua Dong, Fei Fang, and Stuart Russell. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4213–4220, 2019.

Chenyi Liu, Nan Geng, Vaneet Aggarwal, Tian Lan, Yuan Yang, and Mingwei Xu. CMIX: Deep multi-agent reinforcement learning with peak and average constraints. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2021.

Jiarong Liu, Yifan Zhong, Siyi Hu, Haobo Fu, Qiang Fu, Xiaojun Chang, and Yaodong Yang. Maximum entropy heterogeneous-agent reinforcement learning. In *International Conference on Learning Representations*, 2024.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Shaocong Ma, Ziyi Chen, Shaofeng Zou, and Yi Zhou. Decentralized robust V-learning for solving Markov games with model uncertainty. *Journal of Machine Learning Research*, 24(371):1–40, 2023.

Shaocong Ma, Ziyi Chen, Yi Zhou, and Heng Huang. Rectified robust policy optimization for model-uncertain constrained reinforcement learning without strong duality. *Transactions on Machine Learning Research*, 2025.

Jeremy McMahan. Anytime-constrained equilibria in polynomial time. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267, pp. 43399–43416. PMLR, 2025.

Jeremy McMahan, Giovanni Artiglio, and Qiaomin Xie. Roping in uncertainty: Robustness and regularization in Markov games. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235, pp. 35267–35295. PMLR, 2024.

Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2018.

Washim Uddin Mondal, Vaneet Aggarwal, and Satish V. Ukkusuri. Mean-field approximation of cooperative constrained multi-agent reinforcement learning (CMARL). *Journal of Machine Learning Research*, 25(260):1–33, 2024.

Ted Moskovitz, Brendan O'Donoghue, Vivek Veeriah, Sebastian Flennerhag, Satinder Singh, and Tom Zahavy. ReLOAD: Reinforcement learning with optimistic ascent-descent for last-iterate convergence in constrained MDPs. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25303–25336. PMLR, 2023.

Angelia Nedić and Asuman Ozdaglar. Subgradient methods for saddle-point problems. *Journal of Optimization Theory and Applications*, 142(1):205–228, 2009.

Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005. doi: 10.1287/opre.1050.0216.

Shrenik Patel and Christine Truong. Co2PO: Coordinated constrained policy optimization for multi-agent RL. *arXiv preprint arXiv:2602.02970*, 2026.

Julien Perolat, Rémi Munos, Jean-Baptiste Lespiau, Shayegan Omidshafiei, Mark Rowland, Pedro Ortega, Neil Burch, Thomas Anthony, David Balduzzi, Bart De Vylder, Georgios Piliouras, Marc Lanctot, and Karl Tuyls. From Poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization. In *International Conference on Machine Learning (ICML)*, 2021.

Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T Connor, Neil Burch, Thomas Anthony, et al. Mastering the game of Stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996, 2022.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pp. 2817–2826. PMLR, 2017.

Zachary Roch and Yue Wang. Distributionally robust Markov games with average reward. In *International Conference on Machine Learning (ICML)*, 2026.

Reazul Hasan Russel, Mouhacine Benosman, Jeroen Van Baar, and Radu Corcodel. Lyapunov robust constrained-MDPs: Soft-constrained robustly stable policy optimization under model uncertainty. *arXiv preprint arXiv:2108.02701*, 2021.

Laixi Shi, Eric Mazumdar, Yuejie Chi, and Adam Wierman. Sample-efficient robust multi-agent reinforcement learning in the face of environmental uncertainty. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235, pp. 44909–44959. PMLR, 2024.

Laixi Shi, Jingchu Gai, Eric Mazumdar, Yuejie Chi, and Adam Wierman. Breaking the curse of multiagency in robust multi-agent reinforcement learning. In *International Conference on Machine Learning*, 2025.

Samuel Sokota, Ryan D'Orazio, J Zico Kolter, Nicolas Loizou, Marc Lanctot, Ioannis Mitliagkas, Noam Brown, and Christian Kroer. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. In *International Conference on Learning Representations (ICLR)*, 2023.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by PID Lagrangian methods. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Jayakumar Subramanian, Amit Sinha, and Aditya Mahajan. Robustness and sample complexity of Model-Based MARL for General-Sum Markov games. *Dynamic Games and Applications*, 13(1): 56–88, March 2023.

Chen Tessler, Yonathan Efroni, and Shie Mannor. Action robust reinforcement learning and applications in continuous control. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019a.

Chen Tessler, Daniel J Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations (ICLR)*, 2019b.

Chuyao Wang, Ziwei Wang, and Nabil Aouf. Robust multi-agent reinforcement learning against adversarial attacks for cooperative self-driving vehicles. *IET Radar, Sonar & Navigation*, 19(1): e70033, 2025.

Xuefeng Wang, Lei Zhang, Henglin Pu, Husheng Li, and Ahmed H. Qureshi. Safe continuous-time multi-agent reinforcement learning via epigraph form. In *International Conference on Learning Representations*, 2026.

Yue Wang, Fei Miao, and Shaofeng Zou. Robust constrained reinforcement learning, 2022.

Suhang Wei, Xianwei Wang, Xiang Feng, and Huiqun Yu. MAST: Multiagent safe transformer for reinforcement learning. *IEEE Transactions on Cognitive and Developmental Systems*, 17(4): 976–986, 2025. doi: 10.1109/TCDS.2025.3533744.

Muning Wen, Jakub Grudzien Kuba, Runji Lin, Weinan Zhang, Ying Wen, Jun Wang, and Yaodong Yang. Multi-agent reinforcement learning is a sequence modeling problem. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Wentao Xu, Zhongming Yao, Weihao Li, Zhenghang Song, Yumeng Song, Tianyi Li, and Yushuai Li. TCRL: Temporal-coupled adversarial training for robust constrained reinforcement learning in worst-case scenarios. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2026. Extended abstract.

Taeho Yoon and Ernest K. Ryu. Accelerated algorithms for smooth convex-concave minimax problems with $\mathcal{O}(1/k^2)$ rate on squared gradient norm. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12098–12109. PMLR, 2021.

K. Zhang, Tao Sun, Yunzhe Tao, Sahika Genc, S. Mallya, and Tamer Başar. Robust multi-agent reinforcement learning with model uncertainty. In *Neural Information Processing Systems*, 2020.

Lijun Zhang, Lin Li, Wei Wei, Huizhong Song, Yaodong Yang, and Jiye Liang. Scalable constrained policy optimization for safe multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, 2024a.

Songyuan Zhang, Oswin So, Mitchell Black, Zachary Serlin, and Chuchu Fan. Solving multi-agent safe optimal control with distributed epigraph form MARL. In *Proceedings of Robotics: Science and Systems*, 2025.

Zhengfei Zhang, Kishan Panaganti, Laixi Shi, Yanan Sui, Adam Wierman, and Yisong Yue. Distributionally robust constrained reinforcement learning under strong duality. In *Reinforcement Learning Conference (RLC)*, 2024b.

A PROOFS: WHICH EQUILIBRIUM NOTION EXISTS

The equilibrium-existence theory of the tabular RCMG framework restated in Section 3 does not extend to every instance meeting its assumptions. We identify a step of its existence proof for the Upper-RFNE that needs an extra hypothesis, and exhibit an explicit RCMG that satisfies all of its assumptions and admits no Upper-RFNE while an exact RFNE exists. We then prove existence of the Lagrangian-stationary Upper-RFNE, which fixed points of the primal–dual update meet. These three objects differ in what they enforce. An exact RFNE enforces worst-case feasibility over the uncertainty set. An Upper-RFNE is a saddle of the surrogate Lagrangian with the multiplier free over $\mathbb{R}^J_+$. Proposition 1 shows that such a saddle is robustly feasible, so an Upper-RFNE is an exact RFNE and is Lagrangian-stationary. A Lagrangian-stationary Upper-RFNE keeps the policy side of the saddle but replaces multiplier-side minimality with KKT-type stationarity at the realized kernel. Converse implications fail: the counterexample below has an exact RFNE and no Upper-RFNE, and a Lagrangian-stationary point need not be either. Theorem 2 gives a sufficient condition (one aligned worst-case kernel feasible within budget) under which it is both. Only the Lagrangian-stationary notion is guaranteed to exist under our assumptions. Table 1 summarizes the three.

Setting and notation. Fix an RCMG as in Section 3. For a joint policy π and a stationary kernel P write $J_{R,i}(\pi, P) = \mathbf{E}_{\pi,P,\rho}[\sum_t \gamma^t R_i^t]$ and $J_{C,i,j}(\pi, P) = \mathbf{E}_{\pi,P,\rho}[\sum_t \gamma^t C_{i,j}^t]$, stacked as $J_{C,i} \in \mathbb{R}^J$. Abbreviate $G_i(\pi, P; \lambda_i) := J_{R,i}(\pi, P) - \lambda_i^\top J_{C,i}(\pi, P)$, so that the surrogate Lagrangian of Equation (2) is $\bar{V}_i^{\mathrm{SL}}(\pi, \lambda_i) = \min_{P \in \mathcal{P}} G_i(\pi, P; \lambda_i) + \lambda_i^\top c_i$.

Assumption 1. *(A1) $\mathcal{S}$ and all $\mathcal{A}_i$ are finite, rewards and costs are bounded in magnitude by a common constant $\bar{R}$ (so $\bar{R} \geq \max_i \sup |R_i|$ and $\bar{R} \geq \max_{i,j} \sup C_{i,j}$, with $C_{i,j} \geq 0$), and $\gamma \in (0, 1)$.*

(A2) each $\mathcal{P}(s, \mathbf{a}) \subseteq \Delta(\mathcal{S})$ is compact (implicitly used whenever the framework writes the inner minimum as attained).

(A3) robust Slater: *there is $\xi > 0$ such that for every i and every π_{-i} some π_i^S has $\max_{P \in \mathcal{P}} J_{C,i,j}(\pi_i^S, \pi_{-i}, P) \leq c_{i,j} - \xi$ for all j; by compactness of Π_{-i} and Berge's maximum theorem, the pointwise version of this condition already implies a uniform margin.*

Our experiments measure nominal feasibility at the deployed profile under the nominal kernel alone. Clean RC-MAT costs are $0.66\times$ (collision) and $0.90\times$ (energy) of budget at the tight cell (Table 16).

A.1 WHERE THE PUBLISHED ARGUMENT NEEDS A FURTHER HYPOTHESIS

The published proof derives convexity of the dual-optimal set map from the fact that a dual problem is always convex. That fact requires the function under the supremum to be affine in the multiplier. Here it is not. For fixed (π, P) the map $\lambda_i \mapsto G_i(\pi, P; \lambda_i) + \lambda_i^\top c_i$ is affine, but the inner minimum over P makes $\lambda_i \mapsto \bar{V}_i^{\mathrm{SL}}(\pi, \lambda_i)$ a pointwise minimum of affine functions, i.e. concave. The dual objective $D_i(\lambda_i) = \max_{\pi_i} \bar{V}_i^{\mathrm{SL}}(\pi, \lambda_i)$ is a pointwise maximum of concave functions, in general neither convex nor concave. Its minimizer set has no reason to be convex. Two more steps of that proof need further argument. Its convex-combination inclusion for the best-response map requires graph-convexity in λ. That is strictly stronger than convexity of each image, which is all the proof establishes. Its dual minimization runs over the unbounded set $\mathbb{R}^J_+$ without a coercivity argument. Under (A3) coercivity does hold, because evaluating at π_i^S gives

$$D_i(\lambda) \geq -\bar{R}/(1-\gamma) + \xi \sum_j \lambda_j.$$

Finally, even granting all of the above, a Kakutani fixed point of the best-response map composed with the dual-optimal set map yields the saddle-point inequalities of the Upper-RFNE under the additional requirement that the surrogate has zero duality gap at the fixed point. Robust constrained problems can lack strong duality: Ma et al. (2025) exhibit a two-state robust constrained MDP whose duality gap is strictly positive. Below, an instance transfers that single-agent phenomenon to the multi-agent existence theorem.

A.2 A COUNTEREXAMPLE TO THE PUBLISHED STATEMENT

Take two agents, discount $\gamma = \frac{1}{2}$, and a nine-state tree. Agent two is a dummy (one action, zero reward and cost, threshold $c_2 = 1$). From s_0 agent one picks a_k ($k \in \{1, 2, 3\}$) and moves deterministically to x_k with no reward or cost. From x_1 and x_2 nature picks between two absorbing successors, $\mathcal{P}(x_1, \cdot) = \{\delta_{o_1^A}, \delta_{o_1^B}\}$ and $\mathcal{P}(x_2, \cdot) = \{\delta_{o_2^A}, \delta_{o_2^B}\}$. Elsewhere the sets are singletons. Absorbing states have rewards and costs per step whose discounted values from s_0 are $(u, v) = (4, 1)$ at o_1^A, $(6, 3)$ at o_1^B, $(2, 1)$ at o_2^A, $(8, 3)$ at o_2^B, and $(0, 1)$ at o_3. Agent one's cost threshold is $c_1 = 2$. The uncertainty set is $(s, \mathbf{a})$-rectangular and finite, hence compact. Through a_3 the Slater condition holds with margin $\xi = 1$. Write $\theta \in \Delta^2$ for agent one's mixture over $\{a_1, a_2, a_3\}$. The branches decouple:

$$\bar{V}_1^{\mathrm{SL}}(\pi_\theta, \lambda) = \textstyle\sum_k \theta_k \, g_k(\lambda),$$
$$g_1 = \min(4 + \lambda,\, 6 - \lambda), \quad g_2 = \min(2 + \lambda,\, 8 - \lambda), \quad g_3 = \lambda.$$

Here the dual objective $D(\lambda) = \max_k g_k(\lambda)$ is the zigzag with

$$D(0) = D(2) = D(4) = 4, \qquad D(1) = D(3) = 5.$$

It is not convex, since

$$D(1) > \tfrac{1}{2} D(0) + \tfrac{1}{2} D(2),$$

and its minimizer set $\{0, 2, 4\}$ is not convex. Intermediate convexity steps of the published proof are therefore unavailable here. No Upper-RFNE exists. A case analysis over the five faces of $\arg\max_k g_k(\lambda^*)$ shows that for every candidate (θ^*, λ^*) the multiplier-side inequality of the saddle fails. On the first three faces the surrogate value decreases without bound in λ. On the last two, λ^* does not minimize. Equivalently, the surrogate primal value is 2 while the dual value is 4, a duality gap of 2. Yet the exact RFNE exists. Maximizing the robust value subject to the robust-cost constraint,

$$\max_{\theta} \; 4\theta_1 + 2\theta_2 \quad \text{subject to} \quad 3\theta_1 + 3\theta_2 + \theta_3 \le 2,$$

gives $\theta = (\frac{1}{2}, 0, \frac{1}{2})$ with robust value 2 and robust cost exactly 2. The Lagrangian-stationary points (Definition 2) are a_1 at $\lambda \in \{0, 1\}$, the mixture $(\frac{1}{2}, \frac{1}{2}, 0)$ at $\lambda = 2$, a_2 at $\lambda = 3$, and $(0, \frac{1}{2}, \frac{1}{2})$ at $\lambda = 4$. Tied kernels are mixed so that the realized cost equals 2 whenever $\lambda > 0$. That set contains the unconstrained robust optimum a_1 (worst-case cost 3) but excludes the exact RFNE $(\frac{1}{2}, 0, \frac{1}{2})$, because a_1 and a_3 never both attain $\max_k g_k(\lambda)$.

A.3 THE EXISTENCE THEOREM FOR THE STATIONARITY NOTION

The stationarity notion replaces the global multiplier-side minimality of the saddle by KKT-type conditions. That minimality encodes strong duality and can be unsatisfiable. The solutions of the KKT-type conditions are the fixed points of an idealized version of the dual update of Section 4, one with exact best responses on the policy side and an unbounded multiplier. They are also the fixed points of the algorithm of the tabular framework. The deep update approximates them, using a bounded multiplier and policy-gradient stationarity in place of a best response.

Definition 2 (Lagrangian-stationary Upper-RFNE; Definition 1 in the notation above)**.** *A triple (π^*, λ^*, P^*) with $\pi^* \in \Pi$, $\lambda_i^* \in \mathbb{R}_+^J$ and one kernel $P_i^* \in \mathcal{Q}$ per agent, where $\mathcal{Q}(s, \mathbf{a})$ is the closed convex hull of $\mathcal{P}(s, \mathbf{a})$ per state-action pair (a per-cell mixture of admissible kernels, a randomized nature), is a* Lagrangian-stationary Upper-RFNE *if for every agent i: (1) $\bar{V}_i^{\mathrm{SL}}(\pi_i, \pi_{-i}^*, \lambda_i^*) \le \bar{V}_i^{\mathrm{SL}}(\pi^*, \lambda_i^*)$ for all π_i; (2) $G_i(\pi^*, P_i^*; \lambda_i^*) = \min_{P \in \mathcal{P}} G_i(\pi^*, P; \lambda_i^*)$; (3) $J_{C,i,j}(\pi^*, P_i^*) \le c_{i,j}$ and $\lambda_{i,j}^* \, (c_{i,j} - J_{C,i,j}(\pi^*, P_i^*)) = 0$ for all j.*

Condition (1) is the policy side of the saddle, kept verbatim. Condition (3) is feasibility and complementary slackness at the realized kernel of (2). That is the stationarity condition of a projected dual update driven by costs measured under the kernel the robust best response returns. At such a point $\bar{V}_i^{\mathrm{SL}}(\pi^*, \lambda_i^*) = J_{R,i}(\pi^*, P_i^*)$: the surrogate value is the reward obtained under the realized kernel. Conditions (2)–(3) constrain the realized kernel alone, while the alignment condition of Theorem 2 extends feasibility to the whole uncertainty set, bounding $\sup_{P \in \mathcal{P}} J_{C,i,j}$. Cost-directed attacks of Section 5.3 probe that supremum empirically.

Proposition 1 (a surrogate saddle is robustly feasible)**.** *Let (π^*, λ^*) be an Upper-RFNE, so that for each i the multiplier λ_i^* minimizes $\bar{V}_i^{\mathrm{SL}}(\pi^*, \cdot)$ over all of $\mathbb{R}_+^J$ and π_i^* maximizes $\bar{V}_i^{\mathrm{SL}}(\cdot, \pi_{-i}^*, \lambda_i^*)$. Then π^* is robustly feasible, $\bar{V}_i^{\mathrm{SL}}(\pi^*, \lambda_i^*) = \bar{V}_i(\pi^*)$, π^* is an exact RFNE, and (π^*, λ^*, P^*) is Lagrangian-stationary for any P_i^* attaining the minimum in (2) at (π^*, λ_i^*).*

Proof. Write $h(\lambda) := \bar{V}_i^{\mathrm{SL}}(\pi^*, \lambda) = \min_{P \in \mathcal{P}} [J_{R,i}(\pi^*, P) + \lambda^\top (c_i - J_{C,i}(\pi^*, P))]$. Suppose some $P \in \mathcal{P}$ has $J_{C,i,j}(\pi^*, P) > c_{i,j}$. Then for every $t \geq 0$,

$$\begin{aligned} h(t\, e_j) &\leq J_{R,i}(\pi^*, P) + t\,(c_{i,j} - J_{C,i,j}(\pi^*, P)) \\ &\to -\infty \qquad (t \to \infty), \end{aligned}$$

so h has no minimizer on $\mathbb{R}_+^J$, against the multiplier-side saddle inequality. Hence $\max_P J_{C,i,j}(\pi^*, P) \leq c_{i,j}$ for every j: π^* is robustly feasible. Robust feasibility makes every affine piece of the concave map h componentwise nondecreasing, so

$$\begin{aligned} \min_{\lambda \geq 0} h &= h(0) \\ &= \min_P J_{R,i}(\pi^*, P) \\ &= \bar{V}_i(\pi^*), \end{aligned}$$

and $h(\lambda_i^*)$ equals this number by minimality. For a robustly feasible deviation π_i the same monotonicity applies to $h_{(\pi_i, \pi_{-i}^*)}$, so

$$\begin{aligned} \bar{V}_i(\pi_i, \pi_{-i}^*) &= h_{(\pi_i, \pi_{-i}^*)}(0) \\ &\leq h_{(\pi_i, \pi_{-i}^*)}(\lambda_i^*) \\ &\leq h(\lambda_i^*) \\ &= \bar{V}_i(\pi^*), \end{aligned}$$

where the second inequality is the policy-side saddle inequality. This is the exact RFNE condition. Finally let P_i^* attain the minimum in (2) at (π^*, λ_i^*). Condition (2) of Definition 2 holds by choice; feasibility at P_i^* follows from robust feasibility. If $\lambda_{i,j}^* > 0$ and $J_{C,i,j}(\pi^*, P_i^*) < c_{i,j}$ for some j, then $h(\lambda_i^*) = J_{R,i}(\pi^*, P_i^*) + \lambda_i^{*\top}(c_i - J_{C,i}(\pi^*, P_i^*)) > J_{R,i}(\pi^*, P_i^*) \geq h(0)$, against minimality. Complementary slackness holds, and the triple is Lagrangian-stationary. □

Restatement (Theorem 1, existence)**.** *Under (A1), (A2), and (A3), every RCMG admits a Lagrangian-stationary Upper-RFNE with worst-case kernels in the per-state-action convex hull of $\mathcal{P}$ (kernels in $\mathcal{P}$ itself when every $\mathcal{P}(s, \mathbf{a})$ is convex), with the multiplier bound $\sum_j \lambda_{i,j}^* \leq 2\bar{R}/((1-\gamma)\,\xi)$ for every agent.*

Proof. We first record three facts. Fact (i) is that convexification is harmless: replacing each $\mathcal{P}(s, \mathbf{a})$ by its closed convex hull $\mathcal{Q}(s, \mathbf{a})$ changes no robust value, cost, or surrogate. Worst-case evaluation of a stage payoff u uses the robust evaluation operator

$$(\mathcal{T}W)(s) = \sum_{\mathbf{a}} \pi(\mathbf{a} \mid s) \min_{p \in \mathcal{P}(s, \mathbf{a})} p^\top [u(s, \mathbf{a}, \cdot) + \gamma W],$$

a γ-contraction whose unique fixed point equals $\min_P V_u^{\pi, P}$ pointwise. This minimum is attained by per-$(s, \mathbf{a})$ minimizing selections, which exist by compactness. The inner minimand is affine in p, so its minima over $\mathcal{P}(s, \mathbf{a})$ and $\mathcal{Q}(s, \mathbf{a})$ coincide. The same argument with a maximum in place of the minimum shows that the robust cost $\max_P J_{C,i,j}$ is unchanged, so (A3) holds over $\mathcal{Q}$ exactly as stated over $\mathcal{P}$. From now on work over $\mathcal{Q} := \bigotimes \mathcal{Q}(s, \mathbf{a})$, which is compact, convex, and rectangular. An element is a per-cell randomization over kernels, and as above the minimum over $\mathcal{Q}$ is attained at an extreme kernel. Fact (ii) is contraction perturbation: if γ-contractions T_ϑ depend continuously on a parameter ϑ pointwise, their fixed points satisfy

$$\|x_\vartheta - x_{\vartheta'}\| \leq (1-\gamma)^{-1} \|T_\vartheta x_{\vartheta'} - T_{\vartheta'} x_{\vartheta'}\|$$

and hence vary continuously. Fact (iii) is value continuity: $(\pi, P) \mapsto J_{R,i}, J_{C,i,j}$ are continuous, because $J = \rho^\top (I - \gamma P_\pi)^{-1} r_\pi$ with $\|\gamma P_\pi\|_\infty \leq \gamma < 1$ uniformly.

Fix $\bar{\lambda} := 2\bar{R}/((1-\gamma)\xi) + 1$ and $\Lambda_i := [0, \bar{\lambda}]^J$, and work on the compact convex set $X := \Pi \times \prod_i \mathcal{Q} \times \prod_i \Lambda_i$. For $x = (\pi, (P_i), (\lambda_i)) \in X$ define three families of maps. Given $(\pi_{-i}, P_i, \lambda_i)$, agent i faces the ordinary MDP with kernel $\tilde{P}(s' \mid s, a_i) = \mathbf{E}_{\mathbf{a}_{-i} \sim \pi_{-i}}[P_i(s' \mid s, a_i, \mathbf{a}_{-i})]$ and reward $\mathbf{E}_{\mathbf{a}_{-i} \sim \pi_{-i}}[R_i - \lambda_i^\top C_i]$. Let $\mathcal{B}_i(x)$ be the policies supported per state on the argmax of its optimal Q-function. Each set $\mathcal{B}_i(x)$ is a nonempty compact product of simplex faces, hence convex. Its elements maximize $G_i(\cdot, \pi_{-i}, P_i; \lambda_i)$. It has closed graph: Q^* depends continuously on x by (ii), and support conditions pass to limits. Given (π, λ_i), let W_i be the fixed point of the robust evaluation operator for $\bar{u}_i = R_i - \lambda_i^\top C_i$, and let $\mathcal{N}_i(x)$ be the kernels in $\mathcal{Q}$ whose cell at every $(s, \mathbf{a})$ minimizes $p^\top[\bar{u}_i + \gamma W_i]$ over $\mathcal{Q}(s, \mathbf{a})$. Each such kernel attains $\min_P G_i(\pi, P; \lambda_i)$: a per-cell minimizing selection realizes the fixed point of the operator, and the minimum over $\mathcal{Q}$ equals the minimum over $\mathcal{P}$ by Fact (i). The set is a nonempty compact product of argmin sets of linear functions over compact convex cells, hence convex. It has closed graph because the coefficient vector $\bar{u}_i + \gamma W_i(x)$ is continuous in x by (ii), and the argmin of a linear map over a fixed compact convex cell has closed graph under coefficient convergence. No visitation restriction is imposed: cells the policy never visits have zero weight in G_i, so requiring the minimum there costs nothing and keeps the correspondence convex. The multiplier maps are $\mathcal{L}_i(x) := \arg\min_{\lambda \in \Lambda_i} \lambda^\top (c_i - J_{C,i}(\pi, P_i))$. Each is a face of the box, convex and compact, and upper semicontinuous by Berge's theorem and (iii).

The product correspondence maps X into nonempty closed convex subsets of itself with closed graph, so Kakutani's theorem yields a fixed point (π^*, P^*, λ^*). At the fixed point, for each i the pair (π_i^*, P_i^*) is a saddle of $G_i(\cdot, \cdot\,; \lambda_i^*)$ on $\Pi_i \times \mathcal{Q}$: the policy component maximizes against P_i^* and the nature component minimizes against π^*. Condition (2) of Definition 2 is the nature-side inequality, and condition (1) follows by chaining

$$\begin{aligned} \min_P G_i(\pi_i, \pi_{-i}^*, P; \lambda_i^*) &\le G_i(\pi_i, \pi_{-i}^*, P_i^*; \lambda_i^*) \\ &\le G_i(\pi^*, P_i^*; \lambda_i^*) \end{aligned}$$

and adding $\lambda_i^{*\top} c_i$. Here the first inequality holds because P_i^* is one admissible kernel inside the minimum. The second holds by the fixed-point optimality of π_i^* in the MDP induced by (π_{-i}^*, P_i^*). Its right-hand side equals $\min_P G_i(\pi^*, P; \lambda_i^*)$ by condition (2), which closes the chain. Any unilateral policy deviation therefore gains nothing at the worst-case kernel of the fixed point. This bounds the robust value of the deviation from above. For condition (3), let π_i^S be the Slater response to π_{-i}^*. Combining the policy-side inequality with $J_{C,i,j}(\pi_i^S, \pi_{-i}^*, P_i^*) \le c_{i,j} - \xi$ gives

$$G_i(\pi^*, P_i^*; \lambda_i^*) \ge -\bar{R}/(1-\gamma) - \sum_j \lambda_{i,j}^* c_{i,j} + \xi \sum_j \lambda_{i,j}^*,$$

while $\mathcal{L}_i$-optimality forces $\lambda_{i,j}^* J_{C,i,j}(\pi^*, P_i^*) \ge \lambda_{i,j}^* c_{i,j}$ componentwise and hence

$$G_i(\pi^*, P_i^*; \lambda_i^*) \le \bar{R}/(1-\gamma) - \sum_j \lambda_{i,j}^* c_{i,j}.$$

Combining these results,

$$\xi \sum_j \lambda_{i,j}^* \le 2\bar{R}/(1-\gamma),$$

so no component of λ_i^* lies at the box ceiling $\bar{\lambda}$, and optimality of an interior-or-zero point of the linear program forces $c_{i,j} - J_{C,i,j}(\pi^*, P_i^*) \ge 0$ with complementary slackness. The triple satisfies conditions (1)–(3) of Definition 2 with the stated multiplier bound. □

Remark 1 (mixing is essential)**.** *Both convexities carrying Kakutani's theorem come from mixing. Policy best-response sets are convex because policies randomize (faces of simplices), and nature argmin sets are convex because the rectangular cells are convex kernel sets. Restricting either object to deterministic elements removes these convexities, and the composed correspondence loses convex values.*

Remark 2 (information structure)**.** *Theorem 1 is stated for policies conditioning on the full state, matching the tabular framework it refines. Under a decentralized restriction to observation-conditioned policies, each agent's induced problem becomes a POMDP. Best responses then lose the per-state argmax characterization whose simplex faces supply the convex values above.*

Theorem 2 assumes one kernel realizes all reward and cost extrema.

Theorem 2 (tightness upgrade)**.** *Let (π^*, λ^*, P^*) be a Lagrangian-stationary Upper-RFNE, and suppose that for each i some aligned kernel $\hat{P}_i \in \mathcal{Q}$ simultaneously minimizes $J_{R,i}(\pi^*, \cdot)$ and maximizes every $J_{C,i,j}(\pi^*, \cdot)$, with $J_{C,i,j}(\pi^*, \hat{P}_i) \le c_{i,j}$ for all j. Then (π^*, λ^*) is an Upper-RFNE in the original saddle-point sense, and π^* is an exact RFNE.*

Proof. Alignment makes $\hat{P}_i$ minimize $G_i(\pi^*, \cdot\,; \lambda_i^*)$ term by term, so $G_i(\pi^*, \hat{P}_i; \lambda_i^*) = G_i(\pi^*, P_i^*; \lambda_i^*)$. Expanding and using $J_{R,i}(\pi^*, \hat{P}_i) \le J_{R,i}(\pi^*, P_i^*)$ and $J_{C,i}(\pi^*, \hat{P}_i) \succeq J_{C,i}(\pi^*, P_i^*)$ gives $J_{R,i}(\pi^*, \hat{P}_i) = J_{R,i}(\pi^*, P_i^*)$ and, with condition (3), $\lambda_{i,j}^*(c_{i,j} - J_{C,i,j}(\pi^*, \hat{P}_i)) = 0$. Robust feasibility of π^* makes every affine piece of the concave map $h(\lambda) := \bar{V}_i^{\mathrm{SL}}(\pi^*, \lambda)$ componentwise nondecreasing, so

$$\begin{aligned}\min_{\lambda \succeq 0} h &= h(0) \\ &= \min_{P} J_{R,i}(\pi^*, P) \\ &= J_{R,i}(\pi^*, \hat{P}_i),\end{aligned}$$

where the first equality uses monotonicity and the last uses alignment. The identities established above evaluate $h(\lambda_i^*)$ to the same number: λ_i^* attains the global minimum, which is the multiplier-side saddle inequality. For exactness, any robustly feasible deviation π_i has a nondecreasing $h_{(\pi_i, \pi_{-i}^*)}$, so

$$\begin{aligned}\bar{V}_i(\pi_i, \pi_{-i}^*) &= h_{(\pi_i, \pi_{-i}^*)}(0) \\ &\le h_{(\pi_i, \pi_{-i}^*)}(\lambda_i^*) \\ &\le h_{\pi^*}(\lambda_i^*) \\ &= \bar{V}_i(\pi^*),\end{aligned}$$

where the second inequality is condition (1). Hence (π^*, λ^*) satisfies both saddle-point inequalities of the Upper-RFNE and π^* is an exact RFNE. □

The game the deep updates target. The deep updates target the following stationarity conditions. The *action-perturbation game* has $2N$ players: policies $\pi_i : \mathcal{S} \to \Delta(\mathcal{A}_i)$ and natures $\nu_i : \mathcal{S} \to [-\epsilon_\delta, \epsilon_\delta]^{d_i}$, with d_i the number of action coordinates of agent i. Executed actions are $\hat{a}_i = \mathrm{clip}(a_i + \nu_i(s), \mathcal{A}_i)$ with all natures active. Given multipliers λ, policy i maximizes and nature i minimizes agent i's λ_i-modified value of the perturbed process. The multiplier λ_i follows projected ascent on the violations of the deterministic perturbed rollout. The targeted conditions mirror Definition 2 at the realized perturbed process: each π_i^* a best response given $(\pi_{-i}^*, \nu^*, \lambda_i^*)$, each ν_i^* a minimizing response given $(\pi^*, \nu_{-i}^*, \lambda_i^*)$, and feasibility with complementary slackness at the deterministic rollout under ν^*. Our updates are stochastic-gradient analogues of these three conditions. Theorem 1 covers the tabular rectangular counterpart in which independent per-agent kernel choices replace the coupled action perturbations. Theorem 3 below shows that the coupled game itself, in tabular form, has the stochastic-evaluation analogue of the targeted equilibrium.

The tabular action-perturbation game. The game lives on finite state and action spaces under the fixed nominal kernel: there is no kernel player. Each agent has a finite shift set $\mathcal{D}_i$ and a fixed perturbation map $m_i : \mathcal{A}_i \times \mathcal{D}_i \to \mathcal{A}_i$ with $\hat{a}_i = m_i(a_i, \delta_i)$. Policies $\pi_i : \mathcal{S} \to \Delta(\mathcal{A}_i)$ and natures $\nu_i : \mathcal{S} \to \Delta(\mathcal{D}_i)$ are randomized and stationary, and every nature acts in every transition. Given λ_i, policy i maximizes and nature i minimizes agent i's λ_i-modified value of the jointly perturbed process. This value is written $G_i(\pi, \nu; \lambda_i)$, with $J_{C,i,j}(\pi, \nu)$ its cost values. Because rewards and costs are evaluated at executed actions, values here are indexed by the profile rather than by (π, P) as in the setting.

Theorem 3 (existence for the tabular action-perturbation game)**.** *Assume rewards and costs bounded in magnitude by $\bar{R}$ and a uniform Slater condition: for some $\xi > 0$, every i and every profile (π_{-i}, ν) admit a response π_i^S with $J_{C,i,j}(\pi_i^S, \pi_{-i}, \nu) \le c_{i,j} - \xi$ for all j. Then there exists a profile $(\pi^*, \nu^*, \lambda^*)$ with $\sum_j \lambda_{i,j}^* \le 2\bar{R}/((1-\gamma)\xi)$ such that for every agent i: (1) π_i^* maximizes $G_i(\cdot, \pi_{-i}^*, \nu^*; \lambda_i^*)$; (2) ν_i^* minimizes $G_i(\pi^*, (\cdot, \nu_{-i}^*); \lambda_i^*)$; (3) $J_{C,i,j}(\pi^*, \nu^*) \le c_{i,j}$ with $\lambda_{i,j}^*\,(c_{i,j} - J_{C,i,j}(\pi^*, \nu^*)) = 0$ for all j.*

Proof. Facts (ii) and (iii) of the proof of Theorem 1 apply with the profile (π, ν, λ) as the parameter. Kernel and stage rewards of the perturbed process are polynomial in the profile with $\|\gamma P_{\pi,\nu}\|_\infty = \gamma$, so its values are continuous in the profile. Fixed points of parameter-continuous γ-contractions vary continuously. Work on the compact convex product $X := \prod_i \Delta(\mathcal{A}_i)^{\mathcal{S}} \times \prod_i \Delta(\mathcal{D}_i)^{\mathcal{S}} \times \prod_i [0, \bar{\lambda}]^J$ with $\bar{\lambda} := 2\bar{R}/((1-\gamma)\xi) + 1$. Given a profile, agent i faces the ordinary MDP whose kernel and λ_i-modified reward average the joint transition over $\mathbf{a}_{-i} \sim \pi_{-i}$ and $\delta \sim \nu$. Its own executed action is included: choosing a_i executes $m_i(a_i, \delta_i)$ with $\delta_i \sim \nu_i(s)$. Let $\mathcal{B}_i(x)$ be the policies supported per state on the argmax of its optimal Q-function. This is a nonempty compact product of simplex faces, convex, with closed graph as before. Nature i symmetrically faces the MDP over shifts δ_i with everything else averaged out and payoff $-G_i$. Its best-response set $\mathcal{V}_i(x)$ is again a product of per-state argmin simplex faces, convex and compact with closed graph. The kernel-selection subtleties of Theorem 1 (convexification) do not arise: the nature is an ordinary MDP player, defined at every state. The multiplier maps $\mathcal{L}_i(x) := \arg\min_{\lambda \in \Lambda_i} \lambda^\top (c_i - J_{C,i}(\pi, \nu))$ are faces of the box, upper semicontinuous by Berge's theorem and value continuity. Kakutani's theorem applied to the product correspondence yields a fixed point $(\pi^*, \nu^*, \lambda^*)$, and conditions (1)–(2) are the fixed-point optimality of its policy and nature components. For (3), the chain of Theorem 1 repeats verbatim with (π^*, ν^*) in place of (π^*, P_i^*). Condition (1) applied to the Slater response bounds G_i from below, and multiplier optimality bounds it above:

$$G_i(\pi^*, \nu^*; \lambda_i^*) \geq -\bar{R}/(1-\gamma) - \sum_j \lambda_{i,j}^* c_{i,j} + \xi \sum_j \lambda_{i,j}^*,$$

$$G_i(\pi^*, \nu^*; \lambda_i^*) \leq \bar{R}/(1-\gamma) - \sum_j \lambda_{i,j}^* c_{i,j}.$$

Combining these results,

$$\xi \sum_j \lambda_{i,j}^* \leq 2\bar{R}/(1-\gamma),$$

which keeps every component off the ceiling, and interior-or-zero optimality of the linear program forces feasibility with complementary slackness. This fixed point therefore satisfies conditions (1)–(3) with the stated multiplier bound, as claimed. □

Remark 3 (lineage of Theorem 3)**.** *The tabular action-perturbation game has a stationary equilibrium under the randomization and Slater assumptions above, in the lineage of stationary equilibria for finite stochastic games (Fink, 1964).*

Take a three-state, two-agent, three-action instance with one clipped $+1$ shift per nature. The uniform Slater condition holds by construction: the costly action lies two steps above the cheap one, one step beyond the shift range. Damped exact best-response iteration reaches a profile satisfying conditions (1)–(3) to residuals of at most 0.035 in value and 0.005 in constraint on a budget of 1.2, with interior multipliers (1.84).

The shift-generated RCMG. To relate the action-perturbation game to the kernel framework, equip the RCMG with the uncertainty the shifts generate. For each agent i and pair $(s, \mathbf{a})$ the *menu* is $\mathcal{U}_i(s, \mathbf{a}) := \{(R_i(s, \mathbf{a}^{i \leftarrow \delta}),\ C_i(s, \mathbf{a}^{i \leftarrow \delta}),\ P(\cdot \mid s, \mathbf{a}^{i \leftarrow \delta})) : \delta \in \mathcal{D}_i\}$ with $\mathbf{a}^{i \leftarrow \delta} := (m_i(a_i, \delta), \mathbf{a}_{-i})$. The adversary selects one entry per pair, $(s, \mathbf{a})$-rectangularly. Stage functions and transition move together, because an executed-action shift perturbs all three. Robust evaluation over such menus is a γ-contraction as in fact (i). The Lagrangian-stationary conditions of Definition 2 transfer verbatim with menu selections in place of kernels. Write $G_i^{\mathrm{sep}}(\pi, \nu_i; \lambda_i)$ for agent i's *separated* evaluation: its own executed action is shifted by ν_i and the other players are unperturbed. This is the object of the off-policy Remark of Appendix C.

Proposition 2 (separated evaluation is the menu-robust evaluation)**.** *Let natures condition on the state and the joint action before perturbation, $\nu_i : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{D}_i)$. Then $\min_{\nu_i} G_i^{\mathrm{sep}}(\pi, \nu_i; \lambda_i)$ equals the λ_i-modified rectangular robust value of π over the menus $\mathcal{U}_i$, attained by minimizing menu selections per pair (mixtures included, deterministic ones suffice).*

Proof. The evaluation operator of the separated process at $(s, \mathbf{a})$ minimizes, over the distribution $\nu_i(s, \mathbf{a})$, an expression affine in that distribution: the δ-indexed stage-plus-continuation value built from the corresponding menu entry. Its minimum over $\Delta(\mathcal{D}_i)$ is attained at a point mass. Choices at

distinct pairs are unconstrained by one another because ν_i conditions on the full pair. The resulting operator is the menu-rectangular robust operator, a γ-contraction whose fixed point defines both sides. The minimized separated evaluation therefore coincides with the menu-robust evaluation, attained by deterministic menu selections per pair. □

Theorem 4 (approximate correspondence)**.** *In the setting of Theorem 3 with natures conditioned as in Proposition 2, where the existence argument is unchanged by the enlarged, still finite, conditioning and the Slater hypothesis is taken over this enlarged class, let $(\pi^*, \nu^*, \lambda^*)$ satisfy conditions (1)–(3), let η bound every nature's probability per step of an* effective *shift at equilibrium ($\mathbf{P}_{\delta\sim\nu_k^*(s,\mathbf{a})}[m_k(a_k,\delta) \neq a_k] \leq \eta$ for all $k \neq i$ and $(s,\mathbf{a})$), and set $\varepsilon_i := 2\bar{R}\,(1+\|\lambda_i^*\|_1)\,(N-1)\,\eta/(1-\gamma)^2$ and $\varepsilon_J := 2\bar{R}(N-1)\eta/(1-\gamma)^2$. Then (π^*, λ^*), with the menu selection realizing ν_i^* (the pushforward per pair of $\nu_i^*(s,\mathbf{a})$ onto menu entries), is an approximate Lagrangian-stationary Upper-RFNE of the shift-generated RCMG: condition (1) holds up to $2\varepsilon_i$, condition (2) up to $2\varepsilon_i$, and condition (3) holds with $J_{C,i,j} \leq c_{i,j} + \varepsilon_J$ and $|\lambda_{i,j}^*(c_{i,j} - J_{C,i,j})| \leq \lambda_{i,j}^*\varepsilon_J$, all evaluated at the separated process.*

Proof. Couple the jointly perturbed and the separated process on common draws of actions and of agent i's shifts, writing $G_i^{\text{joint}}, J_{C,i,j}^{\text{joint}}$ and $G_i^{\text{sep}}, J_{C,i,j}^{\text{sep}}$ for the modified value and cost return of each process. The trajectories agree until the first step at which some other player's shift is effective, an event of probability per step at most $1-(1-\eta)^{N-1} \leq (N-1)\eta$. Because stage functions evaluate executed actions, the payoffs can differ already at that step. With the modified stage bounded by $\bar{R}(1+\|\lambda_i^*\|_1)$, the union bound $\mathbf{P}[\text{payoffs differ at } t] \leq (t+1)(N-1)\eta$ and $\sum_{t\geq 0}\gamma^t(t+1) = (1-\gamma)^{-2}$ give

$$\begin{aligned}|G_i^{\text{joint}} - G_i^{\text{sep}}| &\leq 2\bar{R}(1+\|\lambda_i^*\|_1)(N-1)\eta/(1-\gamma)^2 \\ &= \varepsilon_i,\end{aligned}$$

and likewise $|J_{C,i,j}^{\text{joint}} - J_{C,i,j}^{\text{sep}}| \leq \varepsilon_J$, uniformly over profiles that share ν_{-i}^*. Write $\bar{V}^{\text{menu}}(\pi) := \min_{\nu_i} G_i^{\text{sep}}(\pi, \nu_i; \lambda_i^*)$, the menu-robust evaluation of Proposition 2 at the fixed λ_i^*. For condition (1) and any π_i, Proposition 2 and the estimate give

$$\begin{aligned}\bar{V}^{\text{menu}}(\pi_i, \pi_{-i}^*) &\leq G_i^{\text{sep}}((\pi_i, \pi_{-i}^*), \nu_i^*) \\ &\leq G_i^{\text{joint}}((\pi_i, \pi_{-i}^*), \nu^*) + \varepsilon_i \\ &\leq G_i^{\text{joint}}(\pi^*, \nu^*) + \varepsilon_i,\end{aligned}$$

where the last inequality is Theorem 3(1). Similarly,

$$\begin{aligned}\bar{V}^{\text{menu}}(\pi^*) &\geq \min_{\nu_i} G_i^{\text{joint}}(\pi^*, (\nu_i, \nu_{-i}^*)) - \varepsilon_i \\ &= G_i^{\text{joint}}(\pi^*, \nu^*) - \varepsilon_i\end{aligned}$$

by Theorem 3(2). Combining these results, condition (1) holds up to $2\varepsilon_i$. The same sandwich bounds $G_i^{\text{sep}}(\pi^*, \nu_i^*)$ within $2\varepsilon_i$ of $\min_{\nu_i} G_i^{\text{sep}}(\pi^*, \nu_i)$, which is condition (2). Transporting Theorem 3(3) across the cost estimate gives the two condition-(3) statements. All three conditions therefore hold at the stated accuracies. □

Remark 4 (scope of the correspondence)**.** *Proposition 2 identifies the separated robust value when nature conditions on state and joint action and is optimized to completion. The implemented off-policy target (Appendix C) evaluates the current nature, so it bounds that robust value from above. Theorem 4 is a structural statement: the effective-shift rate η of the other players' natures separates the implemented game from the kernel framework, and the slack is linear in it.*

The menus perturb stage functions and transitions together (an executed-action shift moves all three), so the target is the shift-generated RCMG, not an arbitrary rectangular kernel set. This identity requires joint-action conditioning, whose behavioral effect is measured below.

Conditioning natures on the action before perturbation or the joint action would strictly enlarge the realizable class. We measure both enlargements. Retraining the full recipe with each nature reading state and its own agent's action before perturbation (five seeds, tight cell) falls within one s.d. of RC-MAT in Table 3: $0.937_{\pm 0.012}$ clean and $0.641_{\pm 0.038}$ under the retrained attack, against the RC-MAT values from five seeds 0.944/0.633. Even conditioning each nature on state and all agents'

actions before perturbation, the largest realizable class in this family, falls within the same range ($0.946_{\pm 0.022}$ clean, $0.641_{\pm 0.043}$ under the retrained attack). Both stronger classes therefore leave the result at this cell unchanged.

The alignment hypothesis of Theorem 2 holds when a single dominant hazard drives the worst cases of reward and costs, as in the grid world of the tabular framework (a kernel there simultaneously minimizes the constrained agent's reward and maximizes its cost). It does not apply in the non-tight regime its own second worked example occupies.

The dual-update justification of the tabular framework calls the map $\lambda \mapsto \bar{V}_i^{\mathrm{SL}}(\pi', \lambda)$ convex with a nonempty subdifferential. In fact the map is concave (a pointwise minimum of affine functions). Its *superdifferential* is the set written there, the convex hull of gradients at worst-case kernels. That set is nonempty, convex, and compact by a Danskin-type argument. No downstream formula therefore changes. The worked example of the tabular framework in its non-tight regime computes exactly a Lagrangian-stationary Upper-RFNE (complementary slackness at the realized kernel), the object Theorem 1 covers. Its grid-world experiment satisfies the Slater and alignment conditions, so Theorem 2 applies there.

B Implementation Details and Hyperparameters

Naming in the appendices follows the main text: AC-MAT and AC-MATD3 are the anchored constrained methods, RC-MAT and RC-MATD3 their robust variants with training-time NatureActors.

B.1 The NatureActor objective on RC-MAT

The perturbation budget of (3) is enforced by construction on both instantiations: the NatureActor output is $\tanh$-squashed and scaled by ϵ_δ, sampled perturbations are clipped to the same box, and the perturbed action is clipped to $\mathcal{A}_i$.

The on-policy NatureActor of Section 4.3 trains on the same rollouts by the same clipped machinery, ascending the negated λ-modified advantage:

$$\mathcal{L}_{\nu_i} = -\,\mathbf{E}\Big[\min\big(w_i\,\hat{A}_i^{\nu},\ \mathrm{clip}(w_i,\,1-\epsilon_c,\,1+\epsilon_c)\,\hat{A}_i^{\nu}\big)\Big], \qquad \hat{A}_i^{\nu} = -\hat{A}_i, \tag{7}$$

where ϵ_c is the shared clip parameter and w_i the likelihood ratio of the sampled perturbation δ_i; δ_i is clipped to the ϵ_δ box at execution. Minimizing $\mathcal{L}_{\nu_i}$ performs clipped ascent on $-\hat{A}_i$ through its leading minus.

B.2 Protocol details

For reference, the RCMG of Section 3 is the tuple $(\mathcal{N}, \mathcal{J}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, P, \{R_i\}_{i\in\mathcal{N}}, \{C_{i,j}\}_{(i,j)\in\mathcal{N}\times\mathcal{J}}, \{c_{i,j}\}_{(i,j)\in\mathcal{N}\times\mathcal{J}}, \gamma, \rho, \mathcal{P})$. Here $\mathcal{N} = \{1,\dots,N\}$ is the agent set, $\mathcal{J} = \{1,\dots,J\}$ the cost index set, $\mathcal{S}$ and $\{\mathcal{A}_i\}$ the shared state and per-agent action spaces, and the remaining components are as defined there. Each agent also has a loose hard energy tank that freezes it when depleted, sized so the dual-enforced threshold binds first. Evaluation reports episodic cost per agent and dimension against its threshold; the *worst-dimension satisfaction rate*; goal rate (agents on goal at episode end); and return. The worst-dimension satisfaction rate is the agent-averaged share of episodes in which an agent is within threshold on every dimension at once. It is harsher than the imposed expectation constraint. Robustness sweeps rescale the attack budget at test time. Our off-policy battery also adds Gaussian observation and actuation noise (Appendix C.3).

Evaluation pools and protocols. Every repeated number in this paper comes from one of four protocols on one of three seed pools, and the tables name which. The *clean* protocol (i) runs the deterministic policy of the final checkpoint with no adversary for 50 episodes per seed (the clean columns of every table). The *tail* protocol (ii) runs the same policy for 500 clean episodes per seed and serves every per-agent and upper-confidence-bound statistic and the all-agent feasibility counts. The *training-time attack* (iii) uses the training-time NatureActor of the run at $1\times\epsilon_\delta$ for 50 episodes (for rows of the robust variant). The *retrained attack* (iv) trains a fresh NatureActor for 800 episodes against the sampled policy of the frozen checkpoint, then runs 100 evaluation episodes per attacker budget. Its zero-budget row is its own clean reference (within 0.01 of protocol (i) for the anchored

variants and within 0.03 for the unanchored and PID variants). Each feasibility criterion admits seeds the other rejects. At the tight cell three of fifteen AC-MAT seeds fail the agent-mean test and pass the all-agent tail test, and one passes the first and fails the second. A count can therefore rise under the stricter criterion. The *training-time attack* columns of the appendix tables use a training-time adversary: the one a variant trains, if any, else the same-seed RC-MAT one. The grids, the scaling sweeps, the enforcement and frontier tables, and every appendix table use five seeds per cell unless a caption says otherwise. Fifteen seeds (the five plus ten more, trained and attacked under the same recipe) serve the paired test of training against the adversary at both the tight and mid cells, the rows with fifteen seeds of Table 4, and Table 3 except its last three rows. Every upper-confidence-bound column uses the tails of the first five seeds. Numbers under the retrained attack in Tables 3 and 16 use the attacker of Appendix D.4 on the final checkpoints. Retention figures for the same variant from five and from fifteen seeds therefore differ in the pool alone.

Multipliers rise under violation and decay under slack rather than pinning the constraint. They regulate the deterministic rollout of the game a variant plays, with the nature on where it trains one. The clean condition instead evaluates the same policy with the nature off, a different and less costly rollout distribution. Evaluation costs therefore lie below the threshold, so we test cost's monotone response to threshold. All reported costs and thresholds are discounted episodic sums, the convention of the training objective. We convert the undiscounted internal accounting of the ported baselines by measured ratios (Appendix D.13). We also stress-test reached equilibria with an empirical local deviation search, a *local exploitability* lower bound (Appendix B.7).

MAST (Wei et al., 2025) keeps one team-level multiplier. Its nearest counterpart here is the unanchored constraint-only variant with per-agent multipliers. The anchor added to the official MAPPO-Lagrangian implementation (Appendix D.13) tests anchor portability. Def-MARL (Zhang et al., 2025) solves cooperative safe multi-agent control through a distributed epigraph reformulation and targets zero violation under a nominal model. It uses no adversary, and its spread task fails our admission gate (Appendix D.8).

Algorithm 1 RC-MAT

Require: Transformer policy $\{\mu_i\}$ with EMA copy $\{\bar{\mu}_i\}$, NatureActors $\{\nu_i\}$, multipliers $\{\lambda_i \in [\underline{\lambda}, \bar{\lambda}]^J\}$, thresholds $\{c_i\}$, rollout buffer $\mathcal{B}$

1: **for** each iteration **do**
2: **for** each environment step **do**
3: Decode a_i by sampling agent i's stochastic policy (mean $\mu_i(s)$); sample $\delta_i \sim \nu_i(\cdot \mid s)$; execute $\hat{a}_i = \mathrm{clip}(a_i + \delta_i, \mathcal{A}_i)$; store the transition in $\mathcal{B}$ with the λ-modified reward (6)
4: **end for**
5: Compute returns and advantages $\hat{A}_i$ on $\mathcal{B}$
6: Update the critic by the value loss and $\{\mu_i\}$ by the clipped objective plus the anchor (4); update $\{\nu_i\}$ by (7) (one PPO update per iteration); update the EMA copy $\{\bar{\mu}_i\}$
7: Every L episodes, after the K_{warm}-episode warm-up: estimate $\bar{C}_{i,j}$ from M deterministic rollouts (NatureActor at its mean) and take the projected dual step (5)
8: **end for**

AC-MAT and RC-MAT (primary instantiation). AC-MAT is Algorithm 1 without the NatureActor steps. Both keep the upstream MAT settings: one encoder–decoder block, embedding size 64, learning rate 5×10^{-4}, clip parameter 0.2, 15 PPO epochs, and entropy coefficient 0.01. This pair reuses the dual and adversary settings of Table 5: dual cadence L=10 episodes on M=2 deterministic evaluation episodes, warm-up 1000 episodes, step scale α_0=0.8, a multiplier box of $[0, 20]$, anchor weight ζ=10 toward an exponential-moving-average copy at rate τ=0.005, and ϵ_δ at 10% of the action bound. The one change is the zero floor (Section 4). Relative to AC-MAT, RC-MAT adds the four NatureActors (340k parameters against 82k for the transformer), one nature forward pass per executed action, and one nature PPO update per iteration on the same rollouts. It adds no environment interaction. The dual evaluation rollouts (two deterministic episodes per ten training episodes, a 20% episode overhead) are shared by every constrained variant. All variants use the same policy network. Anchor penalties are normalized by the action bounds alone, independent of observation scale or parameterization. Here $a_{i,l}^{\max}$ is the per-dimension box half-width rather than the action range. This EMA copy moves after every optimizer step (every PPO minibatch) at rate τ. The buffer stores each agent's sampled action (the policy output before the NatureActor perturbs it), while perturbed actions

reach the environment alone. On a minibatch, the current policy's log-likelihood, the stored old log-likelihood, and both anchor decodes condition on the same stored prefix. The decoder for agent i is teacher-forced on the stored sampled actions of agents $1, \dots, i-1$, for the current transformer and for the EMA copy alike. The minibatch loss is the PPO clipped surrogate on the λ-modified advantage, plus the value loss, minus the entropy bonus, plus $\zeta \frac{1}{N} \sum_i \frac{1}{d_i} \sum_l \big((\mu_{i,l} - \bar{\mu}_{i,l})/a_{i,l}^{\max}\big)^2$ averaged over the minibatch states. Here μ and $\bar{\mu}$ are the teacher-forced decoder means of the current and EMA transformers. One backward pass follows, then the EMA step. Each NatureActor is a two-layer MLP. It produces a $\tanh$-bounded stochastic perturbation of its agent's executed action and trains with the same PPO machinery on the negated λ-modified advantage. Each nature uses its own agent's advantage rather than a shared team signal. Training uses the same 6000-episode interaction budget as the off-policy runs. One run takes under three hours on ten CPU threads.

All off-policy variants share the same actor–critic architecture: two-layer MLPs with 256 hidden units and twin critics with delayed policy updates (TD3); robust variants add one NatureActor per agent descending the same critics within the ℓ_∞ budget ϵ_δ. Multiplier clipping, warm-up, and dual-signal conditioning follow Section 4 and Appendix B.5. Table 5 lists every hyperparameter, shared across variants and domains except where two values are given. We fix the method hyperparameters during off-policy development and the pursuit–evasion sweeps, before any navigation battery runs: the anchor weight ζ and rate τ, dual cadence L, step scale α_0, warm-up length, and multiplier box. We adjust none per cell.

We test the step schedule, the dual-signal size, and the multiplier floor directly. Replacing the $1/\sqrt{k}$ schedule with the square-summable harmonic $\alpha_0/(1+k)$ on the off-policy instantiation at its operating tight cell (five seeds, doubled horizon) reaches a goal rate of $0.499_{\pm 0.068}$ against $0.697_{\pm 0.150}$ for the $1/\sqrt{k}$ run. A schedule meeting the step-size conditions of the tabular analysis thus trains materially slower here. Raising the dual signal from $M{=}2$ to $M{=}8$ evaluation rollouts on RC-MAT at its binding tight cell (five seeds) gives a clean goal rate of $0.968_{\pm 0.012}$, within one s.d. of RC-MAT in Table 3. It leaves the seed-level feasibility count unchanged. The $M{=}8$ variant gives $12/20$ agent and seed pairs by the one-sided 95% bound and $4/5$ seeds feasible at the worst agent against $15/20$ and $4/5$ for RC-MAT at $M{=}2$ (Appendix D.9, 500-episode tails). In isolation the per-update signal is noisy. Its standard error from the 500-episode spread is 0.047 collision and 0.67 energy at $M{=}2$ ($1.35\times$ and $0.27\times$ the tight budgets). The regulation comes from slow averaging by the dual across hundreds of updates rather than from per-update precision. Removing the off-policy multiplier floor ($\underline{\lambda} = 0$, five seeds, the same cell and horizon as the harmonic control) gives $0.809_{\pm 0.151}$ against $0.697_{\pm 0.150}$ with sign disagreement across seeds (paired $p = 0.39$). Residual bias of the floor is therefore within seed noise.

All experiments run on CPU workstations (small networks make CPU faster than GPU here). A navigation run takes roughly nine hours and a pursuit–evasion run up to two days of wall-clock time on two to three threads. The studies in this paper comprise roughly 500 reported training runs across the eleven scenarios tested. The on-policy family alone stores about 1400 runs with full cost and multiplier logs (about 1800 run directories) once development runs and reused checkpoints are counted. The drift check of Appendix B.7 scans that whole corpus (Section 5 and Appendix D–D.8), on the order of 10^4 CPU-core-hours. All robustness, noise, and exploitability evaluations reuse the trained checkpoints.

B.3 Cost channels and energy tank for the public scenarios

Navigation, discovery, dispersion, and tag are scenarios from the public VMAS and MPE suites. Our harness leaves their dynamics unchanged and alters the collision penalty and the goal radius of navigation. The built-in collision reward penalty is disabled (Section 5.1). Every goal radius is set to the agent radius 0.1 in place of the scenario value 0.05, so the on-goal test behind the goal rate and the terminal reward use that tolerance. Re-evaluating the checkpoints of the tight cell with the scenario radius 0.05 over 500 clean episodes gives goal rates of $0.689_{\pm 0.008}$ for AC-MAT, $0.692_{\pm 0.008}$ for RC-MAT, $0.039_{\pm 0.029}$ unanchored, $0.434_{\pm 0.078}$ unanchored with no warm-up, and $0.386_{\pm 0.071}$ for PID, fifteen seeds each. The tails of the same episodes at the radius used throughout give 0.942, 0.950, 0.123, 0.745, and 0.725, so the anchored lead of Table 3 holds. Collision cost per step of agent i sums the normalized penetration depths over all overlapping neighbours. With positions p_i and disc radii r_i taken from the unmodified scenario, $c_i^{\text{coll}} = \sum_{j \neq i} \max\{0,\ (r_i + r_j - \|p_i - p_j\|)/(r_i + r_j)\}$,

so each overlapping pair contributes in $[0, 1]$ and the cost is zero without body overlap. Energy cost per step is the squared control effort normalized per dimension by the action bound and averaged over action dimensions: $c_i^{\text{en}} = \frac{1}{d_i} \sum_{k=1}^{d_i} (a_{i,k}/u_k)^2$, where a_i is the commanded action after clipping per dimension to the bound u_k and d_i is the number of action coordinates. A frozen agent is commanded zero and pays zero. The navigation energy tank (Section 5.1) holds 80 units, refills at every episode start, and drains by the undiscounted energy cost per step. It freezes the agent once empty. Reported episodic costs are discounted sums, whereas the tank drains undiscounted. A constant effort per step that exactly meets the tightest energy budget (2.5 discounted at γ=0.99 over the 200-step episode) drains roughly 5.8 undiscounted units, an order of magnitude below capacity. At the reported cells the dual-enforced threshold therefore binds first.

B.4 The pursuit–evasion environment

Pursuit–evasion is our own environment. Its arena is the unit square $[0, 1]^2$. Its state (x_1, y_1, x_2, y_2) stacks both agents' positions and is observed by both. Each agent outputs a velocity command clipped per component to a speed bound of 0.05. Positions are clipped to the arena. An episode lasts 200 steps regardless of catches. The reward is zero-sum distance shaping: for inter-agent distance d the pursuer receives $-d$ and the evader $+d$. A ± 5 catch bonus is added on every step with $d < 0.06$. Each agent has a single cost dimension. It pays 1 per step inside any of its three private circular cost zones of radius 0.10. When the environment is created, their centers are drawn uniformly over the arena from a random-number stream separate from the training seed. Each layout is stored with its checkpoint, so every seed plays a different layout. Both agents also pay 1 per step within 0.02 of any border (a shared wall strip). Spawns are rejection-sampled outside the agent's own zones and the wall strip. The NatureActor perturbs each agent's executed action within an ℓ_∞ budget of 0.01, i.e. 20% of the speed bound. Budgets c_1, c_2 constrain the discounted episodic sums of these costs. Asymmetric cells of Appendix C.7 give the two agents different budgets.

B.5 Dual-signal conditioning

Dual warm-up. We hold the multipliers at zero and freeze the dual counter k for the first K_{warm} episodes. Each policy thus first learns to solve the task before the constraint becomes active. Once updates start, (5) confines $\lambda_{i,j}$ to $[\underline{\lambda}, \bar{\lambda}]$. On the off-policy navigation and tag runs the floor $\underline{\lambda}$ is positive, so the zero value is exclusive to the warm-up phase. It is 0 elsewhere, including on pursuit–evasion (Table 5). Our pursuit–evasion stack starts the dual at its first update (Table 5). With the same 1000-episode warm-up, all five unanchored pursuit–evasion seeds still diverge (evader cost 78–81$\times$ budget with its multiplier at the ceiling; three seeds fail both sides).

Exploration-free dual signal. The cost $\bar{C}_{i,j}$ that drives the dual step is measured on deterministic evaluation rollouts (exploration noise off, NatureActor on) rather than on the exploratory training rollouts. It is averaged over two episodes. Exploration noise enters the cost (most sharply as squared control effort) and inflates it by a roughly constant factor. This creates a spurious noise floor above which λ_i accumulates. Regulating the deterministic cost removes this train/evaluation gap. Each multiplier then tracks the cost the deployed policy incurs without exploration noise.

Step-size schedule. A $1/\sqrt{k}$ decay is a nonsummable diminishing step ($\alpha_k \to 0$, $\sum_k \alpha_k = \infty$) covered by projected subgradient analyses (Bertsekas, 2015), and we use it for its training speed. Nedić & Ozdaglar (2009) analyze averaged subgradient saddle-point iterates.

Table 5: Hyperparameters, shared across all variants and both domains except where two values are given as navigation | pursuit–evasion. The appendix domains reuse the navigation values except N, cost dimensions, thresholds, and the tag batch size (Appendices D.5 and D.6). Thresholds are absolute, sized from the calibration runs of Section 5.1.

	Parameter	Value
Backbone	hidden units per layer	256
	batch size	256 (tag: 512)
	actor / critic learning rate	10^{-4} / 3×10^{-4}
	discount γ	0.99
	gradient update period (steps)	4 | 8
	target Polyak coeff. (not the anchor rate τ)	0.005
	target smoothing noise / clip	0.1 / 0.3
	exploration (ϵ-uniform, linear decay)	$0.3 \to 0.05$ over 10^4 ep (0.15 at ep 6000)
	exploration noise σ	0.02
Dual / nature	multiplier range $[\underline{\lambda}, \bar{\lambda}]$	$[0.01, 20]$ off-policy navigation and tag | $[0, 20]$ on-policy, pursuit–evasion, RC-HASAC
	warm-up (episodes, λ frozen)	1000 | 0
	dual cadence L (episodes) / rollouts M	10 | 40 / 2
	dual step size, k dual updates after warm-up	$0.8/\sqrt{1+k}$
	anchor strength ζ	10 ($\{0, 2, 5, 40\}$ in the sweep of Appendix C.7)
	anchor tracking rate τ (EMA of the policy)	0.005
	NatureActor updates	one PPO update per iteration on-policy | 3 steps per actor update off-policy
	policy delay d	2
Environment	episode length (steps)	200
	training episodes	6000 | 10000 collection rounds of 4
	nature budget ϵ_δ (fraction of action bound)	10% (0.1) | 20% (0.01)

B.6 The anchor in a bilinear saddle model

The smallest system exhibiting the rotation that the anchor is designed to damp is the bilinear zero-sum game $f(x, y) = xy$ with one scalar decision variable per player. Simultaneous gradient play is the flow $\dot{x} = y$, $\dot{y} = -x$, whose orbits are closed circles. This is the game analogue of the Poincaré-recurrent dynamics of Section 3 (Mertikopoulos et al., 2018). Regularization by the anchor (4) applies to the policy players alone; the adversary has no such term. Accordingly, we anchor x alone and pull it toward its own EMA copy:

$$\dot{x} = y - \zeta\,(x - \bar{x}), \qquad \dot{y} = -x, \qquad \dot{\bar{x}} = \tau\,(x - \bar{x}), \tag{8}$$

where τ is the target-tracking rate. Fixed points of (8) require $x = \bar{x}$. There the anchor term vanishes and the remaining conditions are those of the unanchored game. For every ζ the origin is thus the unique equilibrium. Anchoring changes the attractor structure without moving the equilibrium. This is the exact analogue of the equilibrium-preservation property claimed in Section 4. The system is linear in $(x, y, \bar{x})$ with characteristic polynomial $\omega^3 + (\zeta + \tau)\omega^2 + \omega + \tau$ in the eigenvalue ω (we reserve λ for multipliers). With every coefficient positive, the Routh–Hurwitz condition for a cubic reduces to $(\zeta + \tau) \cdot 1 > \tau$, which holds exactly when $\zeta > 0$. At $\zeta = 0$ the polynomial factors as $(\omega^2 + 1)(\omega + \tau)$, which recovers the marginally stable rotation. Every $\zeta > 0$ with $\tau > 0$ places the spectrum strictly in the left half-plane. Anchoring one side is already enough to contract the coupled pair to the original equilibrium, because the free player y inherits stability through the game coupling. In the analogy x is the anchored policy and the free player y the unanchored opponent, which stands for the multiplier or the adversary. The empirical signature that separates damped rotation from ordinary optimization collapse is the multiplier tail oscillation of Appendix C.7. Damping halves its amplitude at an active constraint, whereas degeneracy flattens the multiplier by pinning it. Anchored and unanchored runs exhibit these two signatures respectively. The contraction rate $-\max \mathrm{Re}\,\omega$ is nonmonotone in ζ. It peaks at an intermediate value and decays to zero as $\zeta \to \infty$. For large ζ the real part of the oscillatory pair scales as $-1/(2\zeta)$ while the third root escapes along the fast anchored

direction. Figure 3 shows both regimes. The over-damped tail is the toy counterpart of the $\zeta = 40$ over-damping observed in the anchor sweep of Appendix C.7.

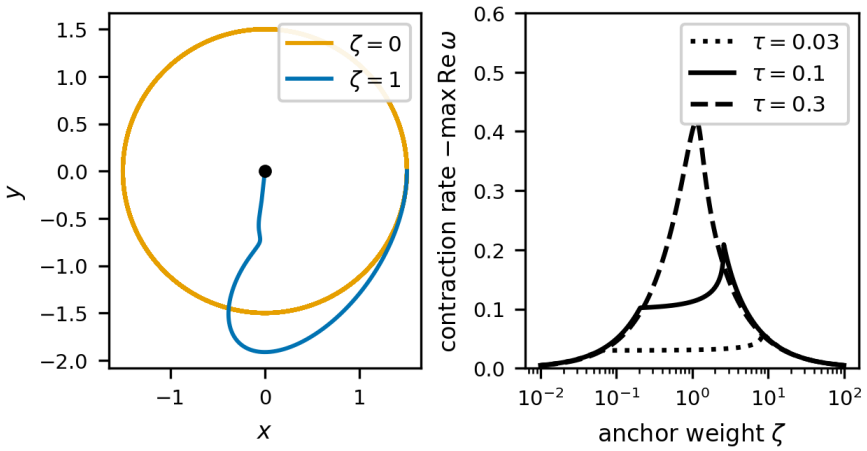

Figure 3: The one-sided anchored bilinear saddle model (8). Left: (x, y) trajectories from a common start ($\tau = 0.1$), unanchored ($\zeta = 0$, orange) and anchored ($\zeta = 1$, blue); black dot: equilibrium. Right: contraction rate $-\max \mathrm{Re}\,\omega$ against ζ for three tracking rates τ; the corner is where the leading eigenvalue switches from the oscillatory pair to the real root.

B.7 Convergence, reporting, and equilibrium stress-testing

Reporting windows. Training curves do not flatten. We investigate primal–dual rotation (Section 4) as one explanation. We report *cycle-averaged* statistics over the final window on pursuit–evasion (the last 500 collection rounds of four episodes, 2000 environment episodes). On the other tasks the reported endpoints are evaluations of the final policy, and multiplier tails use the final 20% of training episodes. We check all cost and multiplier series with a two-window shift diagnostic, a drift check rather than a significance test. In pools with a resume path we extend flagged runs. RC-MAT runs train for a fixed number of episodes; we take their endpoints from their tails and report the check for them after the fact. Deployment and every navigation table entry use the final deterministic policy. The pursuit–evasion tables report cycle averages over the last 500 collection rounds, and Table 10 scores the final checkpoints as deployed policies.

Drift check. For every run we store the per-agent, per-dimension cost and multiplier trajectories recorded during training (one deterministic dual estimate every ten collection rounds of four episodes on pursuit–evasion; ten-episode resolution elsewhere). This check splits the second half of training into two equal windows. For each series it reports the shift in window mean in units of the pooled standard deviation within the window. This is a two-window comparison in the spirit of Geweke (1992) but scaled by that standard deviation within the window, which stands in for the residual limit-cycle amplitude. A series is flagged when the shift exceeds one such unit. We scale by the standard deviation within the window rather than by a standard error because the series are positively autocorrelated. A standard-error scale detects that window means move at all: the naive standard error flags a multiplier series in 1353 and a cost series in 1374 of the 1376 stored RC-MAT runs, and ten-batch batch means flag 1345 and 1342. A standard-deviation scale instead measures the movement against the fluctuation of the series itself. Extension decisions use the latter. Reported endpoints are final iterates at a fixed budget (with the stopping-position check below). A resume path exists for the pursuit–evasion runs, the off-policy tag runs, and the RC-HASAC pools of Appendix C.5. We resume flagged runs there in fixed blocks from their full checkpoints (networks, optimizers, multipliers, and dual step counters restored) for up to two extensions. A run still flagged after two extensions is reported as non-stationary within budget rather than extended further. This applies to the tag cell of Appendix D.6 and to the doubled-horizon rows of Table 8, which are labeled as such. The RC-MAT tail test requires costs averaged over a cycle within thresholds with no drift toward violation, and multipliers at the floor or interior. The check on the 1376 stored RC-MAT runs flags at least one multiplier series in 1055 of them and a cost series in 42 of them. In the tight navigation cell every

seed is flagged, but flagged multipliers lie near the floor (second-half means between 0.005 and 0.11 against a ceiling of 20) and shift by up to 0.13, of the order of their own level. Every cost series in that cell passes (largest cost statistic 0.31 standard-deviation units).

We check stopping position directly. Evaluating the five RC-MAT seeds of that cell at training iterations 400, 500, and the final 599 (500 episodes each) gives goal rates $0.981_{\pm 0.006}$, $0.968_{\pm 0.011}$, and $0.941_{\pm 0.014}$. Agent-mean costs are inside both budgets at every point except one seed's energy at iteration 500 (2.61 against 2.5). Under the retrained attack at the training perturbation size the same three checkpoints keep $0.664_{\pm 0.079}$, $0.685_{\pm 0.081}$, and $0.633_{\pm 0.087}$ of the goal rate (five seeds, 100 episodes each). Training AC-MAT for $12{,}000$ episodes (five seeds, tight cell) shows the same decline: the clean goal rate settles at $0.842_{\pm 0.029}$ against its 0.947 over fifteen seeds at 6000. Feasibility improves: five of five seeds feasible, all four agents within budget on the 500-episode tails in every seed, energy 2.00 against 2.5. Retention under the retrained attack holds at $0.606_{\pm 0.034}$ against 0.645, and the multipliers end at 0.01. A longer horizon therefore lowers the goal rate without losing feasibility or retention; the added margin appears while the multipliers fall to 0.01, so the dual update does not demand it. The competing variants at the same $12{,}000$ episodes (five seeds each) remain behind: the PID variant gives $0.763_{\pm 0.062}$ clean, one of five seeds feasible for every agent, retention $0.37_{\pm 0.07}$; the no-warm-up unanchored variant $0.705_{\pm 0.077}$, four of five, retention $0.29_{\pm 0.10}$; against 0.842, five of five, 0.61 for AC-MAT. Doubling the horizon narrows the AC-MAT lead in goal rate from about 0.20 to about 0.10 and leaves its retention lead unchanged.

On the stored pursuit–evasion corpus the check flags 55 of 436 series in 47 of 109 runs. Those 47 runs include the variants without anchor ($\zeta{=}0$), whose non-stationarity is itself a reported finding. Binding-threshold cells systematically require the most training: the polynomially decaying dual outer loop is slowest exactly where the constraint is active.

Two orthogonal verdicts. Stationarity alone does not certify success: a run pinned in the degenerate basin of Section 4 is stationary. We therefore score each stationary run's attractor separately. It must show (i) costs averaged over a cycle within their thresholds, (ii) an interior multiplier on at least one agent, and (iii) active play. An interior multiplier means the constraint is actively shaping behavior rather than slack or saturated. Active play means the competitive interaction has not collapsed into one agent's forfeit. Cycle-averaged reporting is essential here: single end-of-training snapshots can catch a multiplier at the trough of its residual cycle and misclassify an actively regulated constraint as inactive.

Local exploitability (pursuit–evasion). This test asks whether a run passing both verdicts behaves as an approximate equilibrium. We freeze one agent's policy and give the other a K-episode fine-tuning budget for a best response. The fine-tuning restarts from the trained checkpoint with a fresh replay buffer, fine-tuning-scale exploration, and the Lagrange multipliers frozen at their trained values. Each deviation is thus evaluated inside the same penalized game in which the equilibrium was computed. A deviator's actor loss has no anchor term, because the anchor is a training-time stabilizer rather than part of the game. Our reported statistic is the *feasible-deviation* local Nash residual. Over the deviations observed along each agent's fine-tuning trajectory, it is the largest return improvement $\max_i \left[V_i^{\mathrm{BR}} - V_i^{\mathrm{base}}\right]_+$ among those whose own episodic cost still respects the deviator's budget. Deviations that violate the deviator's own constraint are filtered out rather than flagged: a gain bought by breaking one's own budget is not a profitable deviation of the constrained game. The residual is also an absolute return gain: a relative residual normalized by $|V^{\mathrm{base}}|$ degenerates exactly at balanced equilibria, where $V^{\mathrm{base}} \approx 0$ holds. This search is local, because global optimality of deep best responses is out of reach. The residual therefore lower-bounds the exploitability of the game in which both NatureActors keep adapting, not the gain of a fixed-nature unilateral deviation as in Definition 2. Only a large feasible gain is decisive, because it falsifies equilibrium outright.

We run this deviation search on three seeds per cell and anchor level of the original three-level sweep ($\zeta \in \{0, 10, 40\}$). This budget for the best response is 800 episodes with 25 updates per episode, the opponent frozen, and both NatureActors adaptive. This residual must be interpreted jointly with the attractor check above. A run pinned in the degenerate basin trivially admits no profitable feasible deviation. Its multipliers are saturated and its costs lie far above budget, so there is no equilibrium candidate to test. Pinned unanchored runs report near-zero residuals while failing feasibility outright. Read jointly, the levels separate. Without the anchor no tested run passes both checks: four of six lie in the pinned infeasible basin, one overshoots its budget in cycle mean, and the single cost-feasible run admits a 111.7-return budget-respecting deviation (a lopsided profile far from equilibrium). With

the anchor every tested run is unpinned. On this sweep the residuals are mixed: at $\zeta{=}10$ three of six runs admit no feasible gain, two admit gains of 2.6 and 10.6, and one symmetric-cost seed admits a pursuer gain of 54.6, against per-agent values of order 15–40. The two further seeds of that cell, one of them the pursuer-dominant run of Table 10, admit gains of 175 and 220. At $\zeta{=}40$ two of three symmetric-cost seeds admit gains of 73 and 295.

Because each pursuit–evasion seed instantiates a fresh arena layout, we repeat the search on the fixed-layout pool of Appendix C.7. That pool shares one zone map across five seeds per variant, tests both sides, and uses a budget of 1000 episodes for the best response. This basin separation reproduces. Every anchored run is unpinned with costs at or near budget (the one at-budget seed included). Every unanchored run either saturates a multiplier with its cost at $\sim 80\times$ budget or admits a feasible best-response gain. Anchored residuals themselves are larger there in relative terms, however. Their median is 8.8 and their maximum 15.3 against per-agent returns of 13–21. The largest single gain is 96% of the seed return on the evader side, a feasible deviation that an exact equilibrium would not admit. On the fixed layout the residuals are material in four of five seeds. Here the budget filter matters. On the unanchored asymmetric cell the evader's largest raw best-response gain comes solely by driving its cost to $8\times$ its budget (the infeasible basin of Section 4). Our feasible-deviation residual discards it.

Local exploitability (navigation). The same test ports to the cooperative navigation domain. There a deviating agent now plays against three frozen teammates rather than one frozen opponent. We restore the full checkpoint, freeze the other agents' policies, keep every Lagrange multiplier at its trained value, and fine-tune the deviator for 800 episodes. The fine-tuning uses 25 updates per episode, a fresh replay buffer, fine-tuning-scale exploration, and all NatureActors adaptive. Evaluation is deterministic every 200 episodes and filters deviations whose own cost exceeds the deviator's budget in either dimension. Testing the cell with collision budget 0.064 and energy budget 8.6 on three seeds and all four agents gives twelve deviation tests. Of the 60 evaluated deviations, 59 respect the deviator's budgets. The feasible-deviation residual has median 0.14 and maximum 0.28 against per-agent returns of 2.6–2.8. No agent finds a budget-respecting improvement above 11% of its own return, and the median gain across the twelve tests is 5%. This is the central cell of the off-policy grid.

This search requires an individually parameterized deviator and therefore runs on the off-policy backbones. Under the shared encoder–decoder of the transformer any update to the shared weights moves every agent's policy at once. A single-agent deviation is therefore unimplementable within that parameterization. Shared-decoder RC-MAT is stress-tested by the retrained attack of Appendix D.4 at equal budget. To test unilateral deviations on the transformer, we retrain the full recipe under the decentralized-actor option of the backbone: one actor network per agent with no weight sharing. At the tight cell it reaches the shared-decoder level ($0.956_{\pm 0.008}$ clean over five seeds, both mean costs interior). A unilateral deviation search then applies. We fine-tune each agent's actor head alone on its own λ-penalized return. The search runs for 800 episodes with clipped importance-weighted updates and deterministic evaluations on paired episode draws. The multipliers, encoder, critic, shared action-noise scale, other agents' actors, and the trained nature are all frozen at their checkpoint values, with the nature at its deterministic mean. This isolates the single-actor deviation the equilibrium conditions quantify over. Within this deviation class the search finds no feasible profitable deviation in any of the twenty deviation tests (five seeds, all four agents): every feasible-deviation gain is 0.000. These searches recover the base value ($20/20$ within 0.15 of base). Thirty-one of the 80 evaluated deviations are feasible.

As a positive control, we handicap one agent by scaling its decoder output layer to 0.1. This gives a slow, amply feasible deviator with own return 0.44–0.47 below its trained level. The identical search recovers strictly positive feasible gains on each of two handicapped agents ($+0.122$ and $+0.059$, every evaluation point feasible).

C The Off-Policy Instantiation, Further Backbones, and Anchor Stability

The off-policy instantiation follows in full: its backbone machinery and algorithm, and every run at the looser pairs of Section 4.3. Throughout, the protocol is that of Section 5.1. Here *operating-tight* and *operating-mid* refer to $(0.039, 5.2)$ and $(0.039, 8.6)$; in Appendices C.2 to C.4 the words tight

and mid mean these two cells. At the tight and mid cells of the main text the backbone trains slowly at every tested horizon. It improves with horizon but stays far below the transformer. Appendix C.5 and Table 8 report these runs.

C.1 METHOD DETAILS

Backbone: MATD3. We instantiate the construction on Multi-Agent TD3 (MATD3; Ackermann et al., 2019) and call the result **RC-MATD3**. MATD3 combines MADDPG-style centralized critics (Lowe et al., 2017) with the TD3 stabilizers (Fujimoto et al., 2018) under centralized training and decentralized execution (CTDE). Each agent i maintains a policy network $\mu_i : \mathcal{S} \to \mathcal{A}_i$ and two critic networks $Q_i^{(1)}, Q_i^{(2)} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. It also keeps their slow target counterparts $\bar{\mu}_i, \bar{Q}_i^{(1)}, \bar{Q}_i^{(2)}$ and a target NatureActor $\bar{\nu}_i$. We retain the TD3 stabilizing mechanisms: *twin critics*, whose target uses $\min_k \bar{Q}_i^{(k)}$; *delayed actor updates*, with actor and NatureActor updated every d critic steps; and *target policy smoothing*, with clipped Gaussian noise ϵ' added to target actions. The main change is that the Bellman targets are separated per agent. The target of agent i is evaluated at a joint next-action in which agent i alone is adversarially perturbed by its own (target) NatureActor,

$$y_i = \tilde{r}_i + \gamma(1 - \text{done}) \min_k \bar{Q}_i^{(k)}\big(s',\, \bar{\mathbf{a}}^{(i)}\big), \qquad \bar{a}_j^{(i)} = \begin{cases} \text{clip}\big(\bar{\mu}_i(s') + \bar{\nu}_i(s') + \epsilon_i',\, \mathcal{A}_i\big) & j = i, \\ \text{clip}\big(\bar{\mu}_j(s') + \epsilon_j',\, \mathcal{A}_j\big) & j \neq i, \end{cases} \tag{9}$$

so that each agent bootstraps against its own adversary. Here $\tilde{r}_i$ is the λ-modified reward of (6).

Remark (which robust value the separated target estimates; at an inner-optimal pair-conditioned nature it is the shift-generated RCMG robust evaluation, Proposition 2). Equation (9) is the sampled backup of the operator $(\mathcal{T}_i Q)(s, \mathbf{a}) = \tilde{r}_i + \gamma\, \mathbf{E}_{s' \sim P(\cdot|s,\mathbf{a})}\big[Q\big(s', (\hat{a}_i', \mu_{-i}(s'))\big)\big]$ with $\hat{a}_i' = \text{clip}(\mu_i(s') + \nu_i(s'), \mathcal{A}_i)$. Its fixed point is agent i's λ-modified value of the continuation process in which its own executed actions are shifted by its nature while the other players follow their unperturbed policies. With ν_i trained toward agent i's worst case via (10) below, this mirrors the per-agent minimization in (2). Each agent's robust value is defined by a kernel adversarial to that agent. No kernel simultaneously adversarial to all agents appears in the objective. Execution nonetheless activates every nature, so replay states are sampled from jointly perturbed play. An ablation over five seeds measures what the jointly perturbed backup costs. We rebuild the target of (9) with every other player's next action also shifted by that player's nature; all natures are active in the backup, and sampling and all other settings are unchanged. Training then collapses at the $(0.039, 5.2)$ cell. The energy multiplier saturates its ceiling and the policy freezes at goal rate $0.003_{\pm 0.002}$ against $0.294_{\pm 0.083}$ for the separated backup. Its constraints are trivially satisfied by immobility.

We train the NatureActor to minimize the estimated Q-value of agent i (adversarial) and the policy of the agent to maximize it against the current NatureActor. As in MADDPG, both gradients are taken with the other agents' actions fixed to the sampled replay actions $a_{-i} \sim \mathcal{D}$; the action of agent i alone is re-evaluated under the current networks:

$$\mathcal{L}_{\nu_i} = \quad \mathbf{E}_{(s,a_{-i})\sim\mathcal{D}}\Big[Q_i^{(1)}\big(s,\, \hat{a}_i,\, a_{-i}\big)\Big], \tag{10}$$

$$\mathcal{L}_{\mu_i} = -\mathbf{E}_{(s,a_{-i})\sim\mathcal{D}}\Big[Q_i^{(1)}\big(s,\, \hat{a}_i,\, a_{-i}\big)\Big] + \zeta\, \mathbf{E}\Big[\tfrac{1}{d_i} \textstyle\sum_l \big((\mu_{i,l}(s) - \bar{\mu}_{i,l}(s))/a_{i,l}^{\max}\big)^2\Big], \tag{11}$$

where the NatureActor step holds μ_i fixed and the policy step holds ν_i fixed. The ζ-weighted term is the *anchored policy regularizer*, a squared deviation from the slow target policy $\bar{\mu}_i$. Normalizing per coordinate by the action bound keeps ζ free of the action scale. It does not affect the inner minimax and is motivated in Section 4 ($\zeta = 0$ recovers the base actor loss). Both updates use the first twin critic $Q_i^{(1)}$ rather than the minimum. The min-critic target enters only the Bellman backup of Algorithm 2 to bound overestimation. It stays out of the policy-gradient step, where pessimistic bias would harm convergence. Algorithm 2 gives the full procedure. All N agents share a single replay buffer storing $(s, \mathbf{a}, \boldsymbol{\delta}, \mathbf{r}, \mathbf{c}^{\text{step}}, s', \text{done})$, where $\boldsymbol{\delta} = (\nu_1(s), \ldots, \nu_N(s))$ are the realized NatureActor perturbations. Nesting an inner update per step inside an outer update at the dual cadence mirrors the tabular algorithm.

Algorithm 2 RC-MATD3

Require: Policies $\{\mu_i\}$, NatureActors $\{\nu_i\}$, twin critics $\{Q_i^{(1)}, Q_i^{(2)}\}$ with targets, multipliers $\{\lambda_i \in [\underline{\lambda}, \bar{\lambda}]^J\}$, thresholds $\{c_i\}$, replay buffer $\mathcal{D}$

1: **for** each episode $e = 1, 2, \ldots$ **do**
2: **for** each environment step **do**
3: Select $a_i = \mu_i(s)$ + exploration noise; execute $\hat{a}_i = \mathrm{clip}(a_i + \nu_i(s), \mathcal{A}_i)$; store the transition in $\mathcal{D}$
4: Update each $Q_i^{(1)}, Q_i^{(2)}$ on a minibatch with the λ-modified reward (6) and the per-agent separated target (9)
5: Every d steps: update ν_i by (10), μ_i by the anchored loss (11); soft-update targets
6: **end for**
7: Every L episodes, after the K_{warm}-episode warm-up: estimate $\bar{C}_{i,j}$ from M deterministic rollouts (NatureActor on) and take the projected dual step (5)
8: **end for**

C.2 Budget-schedule grid

We train the full method on a 3×3 grid of absolute thresholds to test whether the dual updates per agent and dimension enforce a schedule of budgets rather than a single hand-tuned pair. The grid is collision $\{0.039, 0.064, 0.097\}$ $\times$ energy $\{5.2, 8.6, 13\}$ with five seeds per cell (45 runs). Table 20 reports cost per cell, min-dimension satisfaction, and return, all measured under each run's trained adversary (its training condition). The min-dimension satisfaction rate (min-dim sat. in Table 20) takes the smaller of each agent's two satisfaction rates per dimension and averages it over agents. The on-policy tables' worst-dimension satisfaction rate (Appendix B.2) instead counts episodes within threshold on every dimension at once. It is at most the min-dimension rate. Across the entire grid the energy cost responds monotonically to its threshold. Across budgets 5.2, 8.6, and 13 the mean energy cost rises from $3.4-3.8$ to $5.4-5.9$ to $8.1-8.5$ in that order. Collision cost stays low ($0.02-0.06$) at every cell. Collision is below its budget in seven of nine cells. It overshoots marginally (0.041–0.045 against 0.039) where the tightest collision budget meets the looser energy budgets. More movement there means more contact. As in Section 5.1, the signal is this monotone tracking rather than saturation. Min-dimension satisfaction is correspondingly high ($0.71-0.79$ agent-averaged, $0.62-0.74$ at the worst agent). Task completion is also budget-controlled. The goal rate rises from $0.29-0.32$ at budget 5.2 to $0.46-0.51$ at 13. Per-cell numbers, training dynamics, heterogeneous budgets, and agent-count scaling are in Appendix D.

C.3 Unseen perturbations

We compare the robust variant RC-MATD3 against AC-MATD3 at the cell $(0.064, 8.6)$ with five seeds each. On clean episodes the constraint-only policy reaches a goal rate of 0.80 versus 0.46 for the adversarially trained one, which Table 20 scores at 0.29 under its training-time adversary. Unseen unstructured perturbations leave both policies near their clean goal rates: Gaussian observation noise and Gaussian actuation noise with standard deviation 0.05, 0.1, and 0.2 against the adversary's action bound of 0.1. Min-dimension satisfaction stays at 0.73–0.77 across the perturbation sweeps.

C.4 Component ablation and external baselines

Table 6: Native single-cost task (stock reward, $J{=}1$; five seeds, mean$_{\pm\text{s.d.}}$) at three energy budgets. Ported methods ($8.3\times$) run their native interaction budget, cost limits converted by the measured discounting ratio 1.287. †: cost over budget. Bold: best goal rate among budget-feasible policies at our interaction budget; $8.3\times$ ports are co-reported, not ranked.

	budget 6.2		budget 8.7		budget 12.4	
Variant	energy cost	goal rate	energy cost	goal rate	energy cost	goal rate
RC-MAT	$4.62_{\pm 0.21}$	$\mathbf{0.991_{\pm 0.012}}$	$4.78_{\pm 0.28}$	$\mathbf{0.994_{\pm 0.009}}$	$4.66_{\pm 0.13}$	$\mathbf{0.992_{\pm 0.011}}$
−Anchor	$4.79_{\pm 0.43}$	$0.762_{\pm 0.319}$	$6.72_{\pm 0.48}$	$0.933_{\pm 0.031}$	$9.00_{\pm 0.55}$	$0.972_{\pm 0.021}$
RC-MATD3	$3.66_{\pm 0.19}$	$0.078_{\pm 0.022}$	$5.96_{\pm 0.33}$	$0.195_{\pm 0.074}$	$6.79_{\pm 0.67}$	$0.303_{\pm 0.044}$
MAPPO-Lag ($8.3\times$)	$0.00_{\pm 0.00}$	$0.000_{\pm 0.000}$	$0.00_{\pm 0.00}$	$0.000_{\pm 0.000}$	$0.00_{\pm 0.00}$	$0.000_{\pm 0.000}$
MACPO ($8.3\times$)	$1.09_{\pm 0.08}$	$0.029_{\pm 0.023}$	$2.01_{\pm 0.47}$	$0.065_{\pm 0.050}$	$3.42_{\pm 0.47}$	$0.111_{\pm 0.082}$
PID-Lag ($8.3\times$)	$3.97_{\pm 0.20}$	$0.291_{\pm 0.124}$	$5.72_{\pm 0.41}$	$0.835_{\pm 0.067}$	$8.54_{\pm 0.61}$	$0.909_{\pm 0.028}$

Off-policy ablation. Our primary comparison is a controlled ablation. All variants share the same MATD3 backbone, network sizes, and training budget. They differ solely in the components of interest: AC-MATD3, the constraint-only off-policy method; −Lagrangian+Nature, robustness without budgets; and −Both (= plain MATD3). −Lagrangian+Nature and −Both never read the cost budgets. We therefore share their trained policies across budget cells and judge them against each cell's budgets at evaluation time. An M3DDPG-style port (Li et al., 2019) is unconstrained: the other agents' actions take one signed-gradient descent step on the agent's value in an ℓ_∞ ball of the adversary's perturbation size. It never met a budget on the native variant (energy 30 against budgets of 6.2–12.4, goal rate 0.006) and is not tabulated. On the constraint-only axis we port MACPO and MAPPO-Lagrangian (Gu et al., 2021), each in its native paradigm. That paradigm has no adversary, a single cost channel on the binding energy dimension, and a cost limit converted to undiscounted accounting by a measured ratio. We run them at both our interaction budget and their native $8.3\times$ more environment steps. We add a variant of the MAPPO-Lagrangian port with its own PID-controlled multiplier (Stooke et al., 2020) at the native budget (a change to the controller alone). All run through the identical deterministic harness. We evaluate these ports on the native task with a single active energy constraint. We retrain RC-MATD3 and AC-MATD3 with the energy budget as the sole active constraint (collision unconstrained but still measured), five seeds per cell.

Table 7 evaluates every ablation variant clean at the tight cell of the problem with two constraints (collision budget 0.039, energy budget 5.2; five seeds). Beside them are the same checkpoints' goal rates under the training-time or transferred attack and under a retrained attack. For RC-MATD3 and −Lagrangian +Nature the training-time or transferred attack is the adversary trained alongside the policy. For the other variants it is the same-seed RC-MATD3 adversary. Pre-attack evaluation within the retrained-attack protocol differs from the clean column by at most 0.015. Removing the Lagrangian dual removes feasibility. −Lagrangian+Nature and −Both spend $2.0\times$ and $5.4\times$ the energy budget in both conditions. Plain MATD3 posts the best clean goal rate (1.00) because it faces neither the budgets nor an adversary. The same backbone with the anchor alone added ($\zeta{=}10$, last row) reaches 0.94 at $1.9\times$ the energy budget. Clean columns show what robustness costs. Like the full method, AC-MATD3 is within every budget and completes more goals (0.34 against 0.22 clean). This gap is within seed noise here and unambiguous at the cell $(0.064, 8.6)$ (0.80 against 0.46, Appendix C.3). Under the training-time or transferred attack the full method holds ($0.22 \to 0.31$) while AC-MATD3 falls to $0.19_{\pm 0.19}$ as individual seeds collapse (values per seed in Appendix D.11). There, however, the attacked rate exceeds the clean one. For the nature-trained policy the training-time attack is its training condition, so the clean rollout is the larger distribution shift. This is consistent with the mild rise and the tighter spread under attack. Rows that stay high under attack are exactly the budget-violating ones. Under the retrained attack, fitted to each frozen policy at the same perturbation size, both variants collapse (0.03 and 0.14). The advantage of the robust variant is thus specific to the training-time and transferred attack class (protocol in Appendix D.4). The on-policy instantiation differs (Section 5.3).

The external comparison runs on the native task with a single constraint (Table 6; stochastic execution in Table 21). Appendix D.13 analyzes the MAPPO-Lagrangian collapse and the control run. At an interaction budget equal to ours neither MACPO nor MAPPO-Lagrangian reaches a working point.

With $8.3\times$ more steps the PID-controlled MAPPO-Lagrangian variant alone does, at the two looser budgets. MAPPO-Lagrangian and MACPO stay below a goal rate of 0.12 at every budget even with $8.3\times$ more steps. Our M3DDPG-style port stays over every budget (energy 30 against budgets of 6.2–12.4).

Table 7: Variant ablation at the tight cell of the two-constraint problem (collision budget 0.039, energy budget 5.2; $N{=}4$, five seeds, mean$_{\pm\text{s.d.}}$), clean condition of Section 5.1, plus the goal rate under the training-time or transferred attack and under the retrained attack (n/r: not run). †: mean cost exceeds its budget; bold: best goal rate among budget-feasible, task-completing variants.

			goal rate		
Variant	collision	energy	clean	training-time or transferred attack	retrained attack
RC-MATD3	$0.021_{\pm 0.006}$	$3.90_{\pm 0.25}$	$0.22_{\pm 0.10}$	$0.31_{\pm 0.03}$	$0.03_{\pm 0.04}$
AC-MATD3	$0.034_{\pm 0.005}$	$3.91_{\pm 0.26}$	$\mathbf{0.34_{\pm 0.16}}$	$0.19_{\pm 0.19}$	$0.14_{\pm 0.13}$
−Lagrangian +Nature	$0.055^{\dagger}_{\pm 0.015}$	$10.53^{\dagger}_{\pm 0.54}$	$0.79_{\pm 0.04}$	$0.50_{\pm 0.06}$	n/r
−Both (= MATD3)	$0.117^{\dagger}_{\pm 0.028}$	$28.16^{\dagger}_{\pm 1.12}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	n/r
−Both, anchored	$0.053^{\dagger}_{\pm 0.008}$	$9.76^{\dagger}_{\pm 0.32}$	$0.94_{\pm 0.01}$	$0.86_{\pm 0.02}$	n/r

C.5 Backbone generality and training horizon

Backbone gap on ports. On the native single-cost navigation task (stock reward, one energy constraint) RC-MAT reaches 0.99 with energy interior. There the off-policy instantiation is at 0.08–0.30. At the tight budget no port attains a feasible, performing point even at $8.3\times$ interaction. At the two looser budgets PID-Lag does (0.835 and 0.909 within budget, Table 6). No early stopping, cost-limit curriculum, or entropy floor rescues MAPPO-Lag (Table 6; Appendix D.13).

Backbones. Sample efficiency depends on the backbone at binding budgets. At its own operating cell RC-MATD3 goes from 0.29 to 0.70 at a doubled horizon, no higher at a quadrupled one (each with its training-time adversary active, Table 8). It stays far below the transformer. A third, maximum-entropy backbone trains at both cells with the constraint-only variant ahead; the adversary price depends on the backbone (Table 8).

The next tests ask whether the off-policy gap to the transformer at the operating cells reflects the machinery or the training length, whether the construction depends on the specific off-policy backbone, and whether either off-policy backbone trains at the tight and mid cells of the main text at all. We retrain RC-MATD3 at its operating tight cell with the same configuration and seeds for $12{,}000$ episodes, twice the equal horizon. We also mount the same nature, dual, and anchor machinery on a HASAC-style backbone (Liu et al., 2024) and call the result RC-HASAC. HASAC is a maximum-entropy actor–critic with stochastic squashed-Gaussian policies. Our implementation updates agents under the centralized-critic convention. This backbone learns slowly: its RC-HASAC reaches a mean goal rate of 0.41 over two pilot seeds at the equal budget and still climbs. We therefore report all its variants at $12{,}000$ episodes as well, in pools that combine runs resumed from 6000 episodes with fresh ones. At the main-text tight and mid cells we retrain both backbones: RC-MATD3 across its variant family and at a doubled horizon, and RC-HASAC with its full and constraint-only variants. Table 8 reports all three batteries. Below we report the RC-MATD3 binding-cell variants beyond RC-MATD3 itself.

Table 8: Off-policy backbones on the two-constraint problem (N=4, five seeds, mean$_{\pm\text{s.d.}}$). Upper block: off-policy operating cells; lower block: main-text tight and mid cells. Goal rate is each run's final evaluation under its training condition: adversary active for RC-MATD3 and RC-HASAC, clean otherwise. The 6000-episode RC-MATD3 row is Table 7's pool (0.22 clean). † as in Table 3.

Variant	episodes	budgets	collision	energy	goal rate	feasible seeds
RC-MATD3	12,000	$(0.039, 5.2)$	$0.029_{\pm 0.010}$	$4.08_{\pm 0.40}$	$0.697_{\pm 0.150}$	4/5
RC-MATD3	24,000	$(0.039, 5.2)$	$0.035_{\pm 0.006}$	$4.56_{\pm 0.35}$	$0.476_{\pm 0.310}$	2/5
RC-MATD3	6000	$(0.039, 5.2)$	$0.024_{\pm 0.004}$	$3.80_{\pm 0.17}$	$0.294_{\pm 0.083}$	5/5
RC-HASAC	12,000	$(0.039, 5.2)$	$0.043^{\dagger}_{\pm 0.013}$	$4.11_{\pm 0.53}$	$0.818_{\pm 0.130}$	3/5
RC-HASAC	12,000	$(0.039, 8.6)$	$0.043^{\dagger}_{\pm 0.007}$	$6.66_{\pm 0.52}$	$0.944_{\pm 0.072}$	2/5
AC-HASAC	12,000	$(0.039, 5.2)$	$0.048^{\dagger}_{\pm 0.018}$	$4.41_{\pm 0.46}$	$0.984_{\pm 0.012}$	1/5
−Both (plain)	12,000	$(0.039, 5.2)$	$0.095^{\dagger}_{\pm 0.023}$	$22.39^{\dagger}_{\pm 4.05}$	$0.959_{\pm 0.086}$	0/5
RC-MATD3	6000	$(0.035, 2.5)$	$0.021_{\pm 0.003}$	$2.42_{\pm 0.21}$	$0.137_{\pm 0.064}$	4/5
RC-MATD3	12,000	$(0.035, 2.5)$	$0.027_{\pm 0.006}$	$2.48_{\pm 0.40}$	$0.402_{\pm 0.165}$	3/5
RC-MATD3	6000	$(0.035, 3.0)$	$0.021_{\pm 0.003}$	$2.64_{\pm 0.22}$	$0.245_{\pm 0.131}$	5/5
RC-MATD3	12,000	$(0.035, 3.0)$	$0.022_{\pm 0.005}$	$2.66_{\pm 0.25}$	$0.548_{\pm 0.210}$	4/5
RC-HASAC	12,000	$(0.035, 2.5)$	$0.021_{\pm 0.004}$	$2.31_{\pm 0.26}$	$0.682_{\pm 0.143}$	4/5
AC-HASAC	12,000	$(0.035, 2.5)$	$0.024_{\pm 0.007}$	$2.55^{\dagger}_{\pm 0.34}$	$0.964_{\pm 0.026}$	3/5
RC-HASAC	12,000	$(0.035, 3.0)$	$0.024_{\pm 0.003}$	$3.20^{\dagger}_{\pm 0.53}$	$0.865_{\pm 0.113}$	2/5
AC-HASAC	12,000	$(0.035, 3.0)$	$0.028_{\pm 0.006}$	$2.83_{\pm 0.35}$	$0.975_{\pm 0.018}$	4/5

A horizon control answers the first question. RC-MATD3 reaches $0.697_{\pm 0.150}$ at 12,000 episodes at its operating tight cell, with multiplier tail components at or below 0.22, both mean costs interior, and four of five seeds feasible. Doubling the training horizon thus raises its goal rate there from 0.29 to 0.697. A further doubling to 24,000 episodes adds nothing to the climb. Its mean falls to $0.476_{\pm 0.310}$ with two of five seeds feasible (Table 8). Four of five seeds end below their 12,000-episode values. One seed collapses late in training: its goal rate of 0.55 at episode 21,500 decays to 0.00 by 24,000 as its energy multiplier rises. Recovery is thus nonmonotone in horizon. Additional budget alone leaves the gap to the transformer open. On the third backbone the construction trains without modification. An unconstrained control exceeds both budgets (energy at $4.3\times$) yet completes goals. The constraint-only variant completes more goals clean but overshoots the collision budget (two further seeds within 0.001 of it). RC-HASAC ends at 0.818 with interior energy and interior multipliers, with three of five seeds feasible. At its $(0.039, 8.6)$ cell the same RC-HASAC reaches $0.944_{\pm 0.072}$ with energy interior ($6.66_{\pm 0.52}$) and a mean collision cost of $0.043_{\pm 0.007}$ at the 0.039 budget: the same signature at a higher goal rate.

The tight and mid cells of the main text separate the backbones chiefly by sample efficiency, with a late-horizon stability limit reported below. RC-MATD3 trains there slowly. At the main text's tight and mid cells it reaches $0.137_{\pm 0.064}$ and $0.245_{\pm 0.131}$ at the equal horizon, rising to $0.402_{\pm 0.165}$ and $0.548_{\pm 0.210}$ when the horizon doubles. Multipliers stay interior throughout, with the largest tail component 1.2 against the ceiling of 20, and mean costs ride their budgets (Table 8). At the equal horizon the transformer is at 0.94–0.98. Other family variants agree at the tight cell. Constraint-only, ported-Lagrangian, and PGD-nature variants are at 0.13–0.23. A random-perturbation control learns fastest ($0.677_{\pm 0.135}$, four of five seeds feasible). The cells therefore measure the sample efficiency of this backbone, not a failure mode of any single component. RC-HASAC trains at both main-text cells with a consistent reading. AC-HASAC outperforms RC-HASAC clean (0.964 against 0.682 tight; $0.975_{\pm 0.018}$ against 0.865 mid). On this backbone training against the adversary thus costs clean goal rate, whereas on the transformer it costs nothing. This ordering holds under the adapted best-response attacker of each family. Construction and validation follow Appendix D.4, refitted per family; a validation checkpoint of the HASAC-style backbone falls from a clean goal of 0.48 to 0.24 under attack. The evaluation protocol of the HASAC-style backbone has unattacked baselines lower than the deterministic ones (0.263 full, 0.952 constraint-only at the main-text tight cell). Within it the attack leaves 0.015 and 0.577 respectively. At the operating tight cell after 12,000 episodes the full variants of both off-policy backbones collapse under the attack: RC-MATD3 full ends at 0.04, and RC-HASAC full at 0.14 against the 0.36 of its constraint-only variant. Best-response robustness evidence thus remains that of the transformer (Section 5.3).

All three backbones thus order consistently at binding budgets. The on-policy transformer trains at the equal horizon; the maximum-entropy backbone pays a cost in clean goal rate for training against the adversary; the deterministic TD3 backbone trains with interior multipliers and costs at budget but is still climbing when the horizon doubles. These backbones differ in update conservatism, policy stochasticity, and the value estimator. Removing the entropy bonus tests the second: both transformer constrained variants, retrained at the tight cell, turn near-deterministic (mean final policy entropy ≈ -2.5). Both variants still train (0.87/0.86 clean, every seed feasible). The robust variant gains no best-response margin over the anchored constraint-only variant ($0.517_{\pm 0.045}$ against $0.500_{\pm 0.037}$); both retain less than the $\approx$0.62 of their stochastic counterparts. Policy stochasticity alone therefore neither blocks transformer training nor unlocks the contribution of the adversary.

C.6 Multi-agent MuJoCo: the duals hold in four of five seeds for one variant on Walker2d and two on HalfCheetah

We run RC-MAT unmodified on Walker2d 2x3, a MuJoCo task from the multi-agent velocity suite of Safety-Gymnasium (Ji et al., 2023). Two agents control one three-joint leg each, with 11-dimensional observations per agent, actions in $[-1, 1]^3$, and 1000-step episodes. The suite's native cost is a velocity-threshold indicator per step computed globally and broadcast to both agents (J=1). The budget of 4.86 translates a cost limit of 25 unsafe steps per episode into our discounted convention by a ratio measured on unconstrained evaluation rollouts. The natural cost of the unconstrained backbone is $49.9_{\pm 7.9}$ (five seeds), ten times the budget, so the constraint binds. Hyperparameters are unchanged (Table 5; ϵ_δ at 10% of the action bound), with five seeds per variant at 6000 episodes. This task lacks a goal channel, so we report reward and cost. The full method trains at the budget: mean velocity cost $4.79_{\pm 3.90}$ with four of five seeds feasible and multiplier tails interior (0.03–2.3 against the ceiling of 20). Its reward is $1973_{\pm 244}$; the unconstrained backbone, whose spread includes failed gaits, reaches $1626_{\pm 896}$. Here the dual prices the velocity budget without collapsing locomotion. Per-seed feasibility is more volatile than on the particle tasks. Walker gaits differ qualitatively across seeds. One full seed ends at 2.3$\times$ budget, and the constraint-only variants split (1/5 feasible anchored, 3/5 unanchored). The trainability role of the anchor thus does not replicate on this cooperative task with a single constraint in either direction, consistent with the tag ablation (Appendix D.6). A fresh best-response adversary per run (protocol of Appendix D.4) moves the reward channel by under 4% either way on every variant. It moves one feasibility count per seed, that of the anchored constraint-only variant, from one to two of five. RC-MAT holds its four feasible seeds under attack. On HalfCheetah 2$\times$3 from the same Safety-Gymnasium velocity suite (budget 1.94) the anchored and unanchored constraint-only variants hold in four of five seeds and RC-MAT in one (Table 17; Appendix D.13).

C.7 Anchor stability on pursuit–evasion and the stabilizer comparison

The extragradient variant of the main text's stabilizer comparison takes each policy update at a lookahead point with hyperparameters otherwise unchanged. Both lookahead variants here are extragradient forms. The optimistic primal–dual updates with last-iterate guarantees for single-agent constrained MDPs (Ding et al., 2023a; Moskovitz et al., 2023) are the nearest prior stabilizers of this kind. At both the tight and mid cells every seed drives its energy multiplier to the ceiling with clean goal rates at most 0.01, and seven of ten runs diverge to numerical failure. A joint version takes the lookahead on the multiplier as well: the dual step is taken at the lookahead policy and the policy step at the lookahead multiplier, with our dual formula, box, and clip unchanged. It fares no better at the tight cell. At the default dual step size three of five seeds finish with the energy multiplier pinned at the ceiling and a clean goal rate of 0.000, and two diverge. At half and double the step size all six seeds diverge before the end of training. AC-MAT's multipliers, by contrast, end near zero. Whether the adaptive feedback is needed at all then has a direct test: a fixed penalty at the scale the dual passes through, and the default rule with the multiplier frozen after the shaping phase. A fixed $(\lambda_{\text{coll}}, \lambda_{\text{en}})$ of $(0.035, 0.15)$, half the peak multipliers of the default variant, trains to $0.515_{\pm 0.063}$ clean with energy at 0.76 against a budget of 2.5. The peak scale $(0.07, 0.30)$ gives $0.088_{\pm 0.023}$ and three times it $0.020_{\pm 0.014}$ (five seeds each, anchored, no warm-up). A penalty the dual merely passes through is far too strong as a constant. Freezing the multiplier after 3000 episodes gives $0.900_{\pm 0.085}$ clean but the energy cost drifts over budget (2.61; two of five seeds feasible). Freezing after 4500 gives $0.878_{\pm 0.076}$ with four of five feasible and retention $0.52_{\pm 0.14}$ under the retrained attack, against the default 0.943, four of five, and 0.61. Neither a fixed penalty nor a frozen multiplier matches the adaptive rule within

0.03 clean at four of five feasible: the small late-training multipliers still do work. Whether the unanchored variant collapses depends on the warm-up protocol. At the tight cell, the unanchored constraint-only variant with the warm-up tripled to 3000 episodes collapses further ($0.010_{\pm 0.006}$ clean goal rate, five seeds). The anchored constraint-only variant trained with no warm-up reaches $0.954_{\pm 0.014}$ (energy 2.46 against 2.5, two of five seeds feasible in mean; Welch t=1.0 against the default 0.943 of the variant with warm-up). Over fifteen seeds the unanchored variant trained with no warm-up reaches $0.743_{\pm 0.097}$ (0.60–0.94) with every seed feasible in mean and energy multipliers resting near 0.03–0.06. Its retained goal rate under the retrained attack is $0.274_{\pm 0.087}$, the level of the PID-stabilized variant and less than half of what the anchored variants keep (0.61–0.63 retained goal rate). The main-text collapse (0.082/0.123) therefore belongs to the protocol that releases the dual onto a budget-violating policy after warm-up. Shaped gradually by the dual from the first episode, the unanchored variant ends at 0.74, below the 0.94–0.95 of the anchored variants but far from degenerate. Part of what the anchor damps in the default protocol is thus the release transient itself. Retained robustness, by contrast, still comes from the anchor. Evaluation-time weight averaging is a generic smoothing that touches no training step (uniform average of the checkpoints at iterations 400, 500, and 599; 50 clean episodes). It leaves the collapsed unanchored variant with warm-up collapsed ($0.147_{\pm 0.087}$, five seeds). It lifts the anchored constraint-only variant from 0.943 to $0.971_{\pm 0.016}$ and the no-warm-up unanchored variant from its final-iterate 0.80 to $0.878_{\pm 0.036}$ on the same five seeds. Averaging at evaluation thus helps every variant that trained and closes none of the gap to the anchor. Table 9 fills the warm-up grid. Every variant is trained without warm-up at the tight cell at three settings of its own parameter, five seeds each: dual step $\alpha_0 \in \{0.4, 0.8, 1.6\}$ for the unanchored and anchored variants, and a common gain scale $\{0.5, 1, 2\}$ on the PID controller. The best unanchored setting reaches a clean goal rate of $0.803_{\pm 0.087}$, the best PID setting $0.777_{\pm 0.076}$, and every anchored setting 0.947–0.954. Feasibility in agent mean is where the anchored variant is weaker: two to four of five seeds (energy 2.38–2.46 against 2.5), against five of five unanchored. Ramping the dual step scale linearly from zero over the 1000 or 3000 episodes after the warm-up (unanchored, default rule otherwise, five seeds each) gives a clean goal rate of $0.700_{\pm 0.123}$ and $0.834_{\pm 0.026}$. Retention is $0.29_{\pm 0.17}$ and $0.28_{\pm 0.08}$ under the retrained attack. Under the retrained attack the best settings retain $0.645_{\pm 0.060}$ (anchored), $0.381_{\pm 0.071}$ (PID), and $0.263_{\pm 0.044}$ (unanchored), so the retention gap persists under equal tuning without a warm-up. The unanchored variant degrades monotonically with warm-up length (0.743, 0.212, 0.123, 0.010 at 0, 500, 1000, and 3000 episodes). The anchored variant is insensitive up to 1000 (0.943–0.954) but degrades ($0.554_{\pm 0.302}$) at 3000 episodes. There three of five seeds end below 0.7 with energy well under budget and multipliers at zero, since the dual is released with half the training episodes left. The anchor therefore widens the window of warm-up lengths over which training succeeds without removing the sensitivity.

warm-up (episodes)	0	500	1000	3000
unanchored constraint-only	$0.743_{\pm 0.097}$ (15)	$0.212_{\pm 0.051}$	$0.123_{\pm 0.069}$ (15)	$0.010_{\pm 0.006}$
anchored constraint-only	$0.954_{\pm 0.014}$	$0.949_{\pm 0.016}$	$0.943_{\pm 0.022}$	$0.554_{\pm 0.302}$

Table 9: Clean goal rate at the tight cell over warm-up length: the anchor widens the window in which constrained training succeeds, though not to a 3000-episode warm-up. Five seeds unless marked; feasible seeds in Appendix C.7's text.

We test whether the anchor damps the primal–dual rotation without moving the equilibrium (Section 4) in the pursuit–evasion domain. We sweep $\zeta \in \{0, 2, 5, 10, 40\}$ with five seeds on the *symmetric* binding cell and $\zeta \in \{0, 10, 40\}$ on the *asymmetric* one. The symmetric cell has both budgets tight; the asymmetric cell has a tight pursuer and a moderate evader budget. Figure 4 reports statistics averaged over a cycle. Without the anchor (ζ=0) the dynamics diverge. The multipliers saturate at the ceiling, costs averaged over a cycle run $16-65\times$ over budget on the symmetric cell ($5-17\times$ asymmetric), and a single seed of five is feasible on either cell. Raising the ceiling tenfold ($\bar{\lambda} = 200$, symmetric cell, five seeds) leaves no seed feasible. Four of five saturate the new ceiling at costs near $80\times$ budget, and the fifth keeps interior multipliers while still running $3.8\times$ over. This anchored–unanchored separation also appears with one fixed layout. Runs above draw a fresh zone map per seed. Rerunning both variants with one map held fixed across five seeds (a dedicated zone seed, everything else identical) reproduces the split. Four of five unanchored seeds are infeasible at costs up to $80\times$ budget with pinned multipliers, against four of five anchored seeds feasible with interior multipliers; the remaining anchored seed rides at about twice budget. The anchor removes the runaway: at ζ=10

no seed reaches the ceiling and the multipliers settle in the interior. All five symmetric-cell seeds are feasible at $0.60\times/0.80\times$ of budget (pursuer/evader), and the asymmetric cell gives $1.14\times/0.72\times$ with the pursuer riding just over budget. Stronger anchoring ($\zeta{=}40$) stays stable but slightly over-damps the dual on the symmetric cell, which costs feasibility. We adopt $\zeta{=}10$. Smaller anchor weights also avoid multiplier saturation. On the symmetric cell, $\zeta{=}2$ and $\zeta{=}5$ already keep every multiplier in the interior (all tail values at most 4.2 against the ceiling of 20) with mean costs at $0.64\times/0.67\times$ and $0.96\times/0.54\times$ of budget. Their spread across seeds is wider than at $\zeta{=}10$, though. The cycle averages above are taken over a still-moving policy, so we also score the final checkpoint of every run as a deployed policy: 500 deterministic episodes per seed with both trained NatureActors active and no learning (Table 10). Frozen evaluations reproduce the split. Without the anchor, the frozen pursuer and evader run $17\times$ and $65\times$ over budget on the symmetric cell in seed mean ($17\times$ and $4.9\times$ on the asymmetric cell), with one and two of five seeds feasible on both sides. At $\zeta{=}10$ the frozen policies are within budget on both sides for four of five seeds on each cell, at $0.53\times/0.66\times$ (symmetric) and $0.77\times/0.29\times$ (asymmetric) of budget. Once frozen, the asymmetric pursuer that rides just over budget in cycle average is inside it. Weaker anchors $\zeta \in \{2, 5\}$ give four and one of five, and $\zeta{=}40$ two and four of five. On the symmetric cell the adopted weight thus ties for the most feasible seeds. Catches per episode under the frozen policies are also listed there.

cell	ζ	pursuer cost/budget	evader cost/budget	catches per episode	both feasible
symmetric	0	$16.83_{\pm 36.47}$	$64.92_{\pm 36.27}$	$88.8_{\pm 93.3}$	1/5
symmetric	2	$0.54_{\pm 0.40}$	$0.50_{\pm 0.29}$	$4.0_{\pm 2.3}$	4/5
symmetric	5	$0.70_{\pm 0.51}$	$3.13_{\pm 5.71}$	$2.6_{\pm 1.8}$	1/5
symmetric	10	$0.53_{\pm 0.45}$	$0.66_{\pm 0.22}$	$27.5_{\pm 50.9}$	4/5
symmetric	40	$1.04_{\pm 0.67}$	$0.57_{\pm 0.30}$	$40.6_{\pm 77.8}$	2/5
asymmetric	0	$16.51_{\pm 36.14}$	$4.86_{\pm 4.30}$	$37.8_{\pm 81.9}$	2/5
asymmetric	10	$0.77_{\pm 0.22}$	$0.29_{\pm 0.14}$	$2.4_{\pm 1.5}$	4/5
asymmetric	40	$0.82_{\pm 0.62}$	$0.27_{\pm 0.09}$	$4.7_{\pm 8.9}$	4/5

Table 10: Without the anchor ($\zeta{=}0$) pursuit–evasion blows both budgets; with it ($\zeta{=}10$) four of five seeds are feasible on both sides. Final checkpoints, 500 deterministic episodes per seed with both NatureActors active, five seeds per row, mean$\pm$s.d. Cost/budget below 1 is feasible.

The local deviation search of Appendix B.7 agrees: no unanchored run is simultaneously feasible and unexploitable, and the anchored runs' residuals are reported there. Repeating the $\zeta{=}0$ ablation on the tag benchmark (five seeds, Appendix D.6) produces no divergence, with interior multipliers and realized energy within 6% of budget.

Five further variants are run at the tight cell $(0.035, 2.5)$ of the transformer on navigation. *Reference form*: swapping the action-space anchor for a Gaussian-KL pull toward the same EMA copy (weight 50, the smallest calibrated value) trains to $0.987_{\pm 0.014}$ clean and retains $0.859_{\pm 0.054}$ under the retrained attack. Its price is riding both budgets (two of five seeds feasible, mean energy 2.50 at the 2.5 threshold). An L2 pull in parameter space (weight 10^4) trains partly ($0.807_{\pm 0.156}$ clean, $0.247_{\pm 0.052}$ retained, four of five feasible). The effect thus comes from the slow-reference family rather than the specific action-space form. Of the three, the adopted form is the most budget-conservative (Table 11). The comparison at matched feasibility trains both forms at the tightened budgets $(0.032, 2.35)$ and judges them at the paper's (five seeds each, 500-episode tails, retrained attack at the training perturbation size). It removes the KL-form lead. It gives $0.863_{\pm 0.203}$ clean (one seed at 0.50) and $0.60_{\pm 0.35}$ retained against $0.910_{\pm 0.027}$ and $0.58_{\pm 0.05}$ for the action-space anchor. Both have all four agents within budget in five of five seeds on the 500-episode tails. The feasible-seed column of Table 11 uses the agent-mean criterion of the 50-episode evaluation and gives four of five for each. At equal per-agent feasibility the two forms are within seed noise on retention and the action-space form is the more reliable in clean goal rate. We thus choose the adopted form on that ground rather than on budget conservatism alone.

Table 11: Reference-penalty forms at the tight cell, five seeds each. At the default budgets the Gaussian-KL pull trains and retains best but rides both budgets. Its lead is gone once trained at tightened budgets and judged at the paper's (last two rows). Retention = goal rate under the retrained attack at size 0.1; feasible seeds judged at the paper's budgets.

reference penalty	clean goal	retention	feasible seeds	mean energy
action-space anchor (default)	$0.943_{\pm 0.022}$	$0.613_{\pm 0.047}$	4/5	2.32
Gaussian-KL pull (weight 50)	$0.987_{\pm 0.014}$	$0.859_{\pm 0.054}$	2/5	2.50
parameter-space L2 pull (weight 10^4)	$0.807_{\pm 0.156}$	$0.247_{\pm 0.052}$	4/5	2.38
action-space anchor, trained at $(0.032, 2.35)$	$0.910_{\pm 0.027}$	$0.583_{\pm 0.048}$	4/5	2.18
Gaussian-KL pull, trained at $(0.032, 2.35)$	$0.863_{\pm 0.203}$	$0.604_{\pm 0.346}$	4/5	2.00

Reference rate: the one-dimensional sweep of τ is contained in the joint window below (its ζ=10 row). Because the EMA moves once per optimizer step, τ sets a timescale of $1/\tau$ minibatch updates. A reference that tracks the policy too closely is no anchor at all.

Joint window: a 3×3 grid over $\zeta \in \{5, 10, 20\}$ and $\tau \in \{0.001, 0.005, 0.02\}$ on the anchored constraint-only variant at the tight cell (three seeds per new cell, the default five at ζ=10; Table 12) shows that the two rates must match. A strong anchor with a slow reference over-damps (ζ=20, τ=0.001: $0.495_{\pm 0.077}$ clean). A fast reference collapses under a weak or the default anchor (τ=0.02: 0.042 at ζ=5, 0.200 at ζ=10). Under a strong anchor it does not collapse ($0.927_{\pm 0.016}$ at ζ=20). At ζ=5 with the default rate the cell trains marginally ($0.878_{\pm 0.025}$ clean). This stabilization window is therefore two-dimensional. One-dimensional sweeps above trace its two axes through the default point. Direct measurement on two instrumented runs at the tight cell (parameter-gradient norms of the anchor and PPO terms every twenty updates) puts the anchor term at the same order as the PPO term. The median ratio is 0.93 and 0.89 per seed with interquartile range about 0.5–1.2, falling to about 0.4 in the last fifth of training. In this regime the anchor damps the update without dominating it. Warm-up length has its own grid (Table 9). The multiplier ceiling has its own control on pursuit–evasion, where raising it tenfold leaves the unanchored failure unchanged.

Table 12: Anchored constraint-only variant at the tight cell $(0.035, 2.5)$ over anchor weight ζ (rows) and EMA rate τ (columns): clean goal rate, mean$_{\pm \text{s.d.}}$, with clean-feasible seeds; three seeds per cell, five at ζ=10.

$\zeta \setminus \tau$	0.001	0.005	0.02
5	$0.977_{\pm 0.020}$ (2/3)	$0.878_{\pm 0.025}$ (3/3)	$0.042_{\pm 0.029}$ (2/3)
10	$0.942_{\pm 0.015}$ (5/5)	$0.943_{\pm 0.022}$ (4/5)	$0.200_{\pm 0.138}$ (5/5)
20	$0.495_{\pm 0.077}$ (3/3)	$0.990_{\pm 0.010}$ (3/3)	$0.927_{\pm 0.016}$ (3/3)

Reference dynamics: we replace the EMA with a reference hard-reset to the current policy every 200 updates, matching the EMA timescale $1/\tau$. Its pull is then purely proximal toward a recent frozen copy without temporal smoothing. With this reference the anchored constraint-only variant still trains ($0.914_{\pm 0.028}$ clean, five of five seeds feasible with interior multipliers) and retains $0.538_{\pm 0.094}$ under the retrained attack. An EMA reference gives $0.945_{\pm 0.019}$ and $0.613_{\pm 0.047}$. The paired differences are $-0.031_{\pm 0.026}$ clean, negative in all five seeds, and $-0.075_{\pm 0.082}$ attacked. Proximity to a slow reference provides the stabilization, and the smoothing adds a small, consistently signed increment. *Anchoring side*: extending the anchor symmetrically to the NatureActors, each pulled toward an EMA copy of itself with the same form and weight, stays within noise of 0.944/0.633 ($0.959_{\pm 0.011}$ clean and $0.659_{\pm 0.050}$ under the retrained attack). Our adopted form therefore anchors the policies alone. *Constraints removed*: the plain backbone with only the anchor added trains to goal $1.000_{\pm 0.000}$ and retains $1.000_{\pm 0.001}$ under the same retrained attack (against 0.918 for the plain backbone under the same attacker; both ignore budgets). Best-response robustness of the anchor thus arrives by a route other than the constraints. Of the generic dampers, a five-fold learning-rate reduction on the unanchored variant recovers partially and unstably ($0.788_{\pm 0.262}$, seeds split between success and collapse). A tighter PPO clip ($0.218_{\pm 0.225}$) does not. To test directly whether the anchor acts as a per-iteration trust region, we tune an adaptive KL penalty to the previous iterate online to hold the unanchored per-iteration policy KL at the anchored measured median (2.5×10^{-4} nats). It still collapses to the unanchored band ($0.086_{\pm 0.060}$ clean, $0.015_{\pm 0.010}$ under the retrained attack), with

its coefficient climbing to 5–13 once the multipliers activate. Slowing the dual instead of the policy fails the same way. A dual step of $\alpha_0/5$ stays in the failed band ($0.215_{\pm 0.052}$ clean over five seeds). A step of $\alpha_0/20$ recovers partially ($0.667_{\pm 0.081}$); both are far from the anchored level. The multiplier trajectories add direct evidence for damping: the residual tail oscillation of the energy multiplier halves under the anchor (0.037–0.042 s.d. against 0.083 unanchored at the tight cell).

Two hypotheses for why the anchor also confers best-response robustness (Section 5.3) fit the data and need not exclude each other. A trainability effect: without the anchor the constrained variants drift into degenerate basins during training, so their frozen policies present low-competence behavior that any retrained adversary exploits cheaply. Robustness is then a property of the competent, anchored policies. The PID variant shows competence alone does not suffice: trained to 0.757/0.812 clean without the anchor, it keeps just 0.293/0.407 against 0.613/0.774 for the anchored variant under the retrained attack. An implicit trust region: the anchor regularizes the policy toward a slowly moving copy of itself throughout training. This biases it away from high-frequency, adversary-specific behavior that a retrained attacker could invert. Contraction in the bilinear model (Appendix B.6) is the linear case of this effect. The KL-equalized control above sharpens this reading: conservatism toward the previous iterate does not even train at these budgets. Any operative trust region is therefore the one anchored to the slow reference. This split of roles agrees in direction with the regularization line the anchor imports (Perolat et al., 2021; Sokota et al., 2023). Its moving-magnet variant likewise lets the reference trail the current iterate. Of the two components, the proximal term has a convergence argument toward regularized equilibria. For a practitioner, the anchored constraint-only variant is the measured recommendation for clean-deployment feasibility at binding budgets on this backbone. The demonstrated value of the training-time adversary lies in training-time and transferred-attack retention on the off-policy backbone and in enlarging the class the policy is stress-tested against. On the primary backbone at the default budget the comparison favors omitting it. Its cost is worth paying solely when the deployment perturbation matches the off-policy setting where the training-time-attack margin appears. To separate the two, we train the anchored constrained variant with a periodically refit best-response adversary in place of one trained alongside the policy throughout. We retrain the nature on frozen-policy rollouts in the last ten of every fifty iterations and keep it frozen but active elsewhere (five seeds per cell). The policy and the dual both pause during those windows, so the policy sees about a fifth fewer updates at equal experience. Under the retrained attack the goal rate is $0.709_{\pm 0.041}$ tight and $0.809_{\pm 0.047}$ mid. The widening threshold, chosen before training, is the mean of the adversary-free anchored variant plus two standard deviations (0.862 against $0.774_{\pm 0.044}$ mid; 0.708 against $0.613_{\pm 0.047}$ tight). At the mid cell it is below its threshold, and at the tight cell exactly at it. Extending that refit variant to fifteen paired seeds, the same execution standard as the comparison of Section 5.3, sharpens the tight cell to the edge of separability. The margin is $+0.046$ $[+0.001, 0.091]$ ($p = 0.05$, uncorrected for the multiple paired comparisons reported). This is a small refit-adversary gain where the energy budget has no slack, consistent with scarcity lowering retention. The mid cell remains non-separable ($+0.035$ $[-0.026, 0.097]$, five seeds), and the margins of Section 5.3 from training against the adversary remain non-separable at both cells. A companion variant that adds the same refit windows to RC-MAT moves its retention by a similarly small amount ($0.633_{\pm 0.087}$ to $0.682_{\pm 0.039}$ tight, $0.826_{\pm 0.020}$ to $0.837_{\pm 0.039}$ mid).

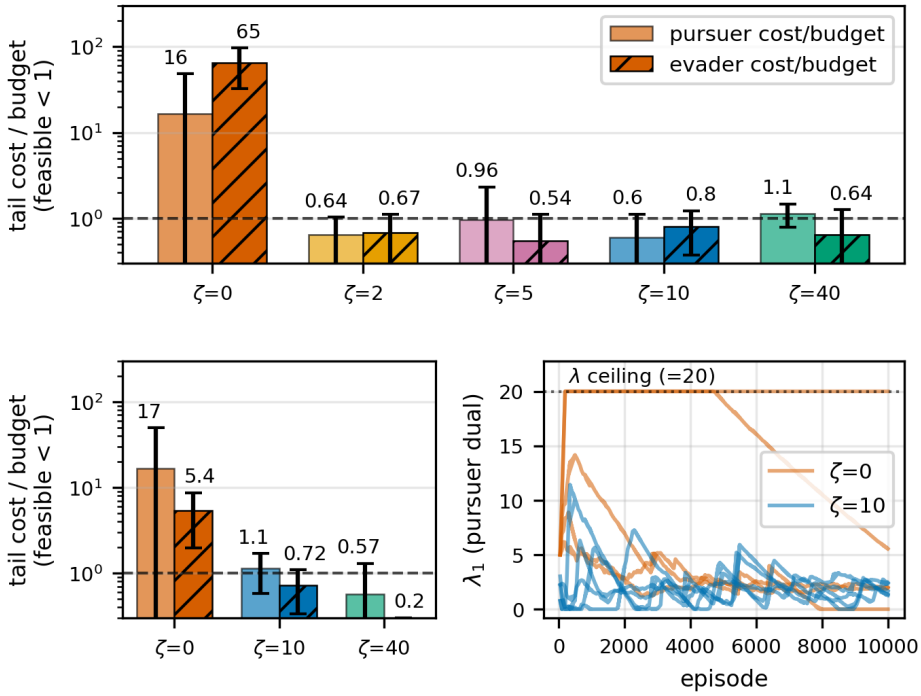

Figure 4: Pursuit–evasion anchor ablation, five seeds per variant. Top and bottom left: cycle-averaged cost relative to budget (feasible < 1, log scale), symmetric cell ($\zeta \in \{0, 2, 5, 10, 40\}$) and asymmetric cell ($\zeta \in \{0, 10, 40\}$). Bottom right: pursuer multiplier trajectories on the symmetric cell, anchored (ζ=10, blue) and unanchored (ζ=0, orange).

A phase-lag read of the navigation collapse. A lag signature supports the rotation reading of Section 4.3 in the deep system. On the pools of fifteen seeds at the tight cell we correlate the multiplier level at one dual update with the budget violation a fixed number of updates later. Correlation is taken per agent and on the energy dimension over the first 300 updates after warm-up. In the unanchored variant the most negative correlation occurs at a median lag of 16 updates (160 episodes; interquartile range 14–19) with value $-0.32_{\pm 0.08}$. A high multiplier is followed after that delay by a lower cost, the delayed corrective response of an integral controller acting on a slowly responding policy. In AC-MAT the same correlation is $-0.00_{\pm 0.05}$, the cost sitting at its threshold with no lagged dependence on the multiplier.

D Additional Experimental Results

Complete evidence behind Section 5 follows, one subsection per question or environment. Appendices D.1 to D.4 serve the retention question: the robust variant under the retrained attack (Figure 5) and under dynamics shifts, the agent-count sweep in Table 14, the binding frontier of the energy budget, and the attack at equal budget with its training-size frontier and cost-directed breaches in Tables 15 and 16. Appendices D.5 to D.8 serve the transfer question: discovery, tag, dispersion (Table 17), and the tasks that fail the admission gate. Appendices D.9 to D.12 serve the feasibility question: the full budget grids (Tables 18, 19, and 20), the training trajectories, the dispersion across seeds of the contested cells, and the heterogeneous per-agent budgets. Appendix D.13 records the external baselines' training dynamics.

D.1 RC-MAT UNDER BEST-RESPONSE PERTURBATION AND DYNAMICS SHIFT

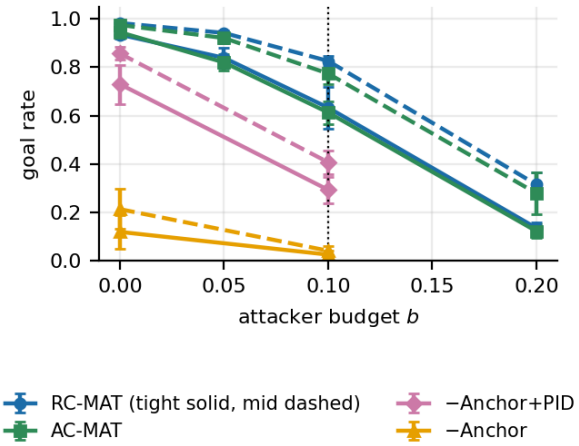

Figure 5: Retention under a retrained attacker follows the anchor: goal rate against the attacker perturbation size b, five seeds, s.d. bars; tight cell solid, mid cell dashed, dotted line the training size; RC-MAT and AC-MAT are measured at four sizes, the other variants at 0 and 0.1 only.

Tail behavior and stopping position. Per-seed feasibility for Table 3's constrained variants uses the clean agent-mean criterion with 50-episode final evals and the caption's seed pools. It gives $15/15$ RC-MAT, $12/15$ AC-MAT, $11/15$ −Anchor+Nature, $10/15$ −Anchor, and $14/15$ −Anchor+PID. Collapsed-variant counts are trivial feasibility. Table 13 asks whether the variants differ in the upper tail as well as in the mean. For each variant at the tight cell it reports the distribution over 500 clean episodes of the maximum over agents per episode. Anchored variants have the tightest upper tails on energy ($\text{CVaR}_{0.9}$ of 4.5–4.6 against 6.3 for the PID dual and 8.2 unanchored). On collision the PID and adversarially trained unanchored variants have the tightest tails, with p95 of 0.113 and 0.134 against 0.230–0.249. Stopping position within the primal–dual cycle is checked in Appendix B.7. Checkpoints 100–200 iterations earlier show higher goal rates with slightly higher energy at the worst agent.

Table 13: Tail statistics at the tight cell (500 clean episodes per seed, five seeds; per-episode maximum over agents, then mean over episodes and seeds, not the maximum of per-agent expected costs). upper_{95}: mean plus 1.96 standard errors, the two-sided upper end (the grid tables' UCB is one-sided, 1.645); $\text{CVaR}_{0.9}$: mean of the worst decile. Budgets 0.035/2.5.

	energy				collision	
Variant	mean	upper_{95}	p95	$\text{CVaR}_{0.9}$	mean	p95
AC-MAT (ours)	3.12	3.18	4.34	4.45	0.051	0.232
+Nature (= RC-MAT)	3.13	3.20	4.49	4.61	0.055	0.249
−Anchor +PID	3.61	3.73	6.04	6.26	0.020	0.113
−Anchor +Nature	3.72	3.89	7.55	8.19	0.027	0.134
−Anchor	3.88	4.06	7.75	8.22	0.052	0.230

Dynamics shift is outside the action-realizable threat model. We therefore also evaluate the frozen policies of the tight cell of the full and anchored constraint-only variants under four transition perturbations that no additive action shift expresses. Evaluation runs with nature off and 100 deterministic episodes per seed. The perturbations are world drag halved and doubled, and agent mass scaled by 1.5 and by 0.67. Both variants move together. Goal rates are 1.00/1.00 at halved drag, 0.47/0.49 at doubled drag, 0.70/0.71 at 1.5× mass, and 1.00/1.00 at 0.67× mass (full/constraint-only, five seeds each, Welch $|t| \leq 1.1$ on every condition). Training against the adversary therefore transfers nothing to dynamics shift. Feasibility is a property of the trained-on dynamics rather than a test-time guarantee. Under the two slowing shifts both variants leave the energy budget (agent-mean energy 3.6–3.7 at doubled drag and 3.1–3.2 at 1.5× mass against 2.5, no seed feasible). The two

speeding shifts keep the budget but push collisions toward their threshold (0.033 against 0.035 at halved drag, four and three of five seeds feasible). Figure 5 sweeps the retrained attack from size 0 to 0.2. RC-MAT degrades monotonically, from 0.935 at b=0 through 0.633 at the training size to 0.135 at twice it at the tight cell (0.983, 0.826, and 0.315 at the mid cell). The anchored constraint-only variant follows it at every size, 0.01–0.02 below at the tight cell (0.820, 0.613, and 0.122 at $b = 0.05$, 0.1, and 0.2) and 0.02–0.05 below at the mid cell (0.921, 0.774, and 0.279). The unanchored constraint-only variant is degenerate throughout (0.119/0.213 before attack, 0.025/0.041 under it). Between the two lies the PID-dual constraint-only variant. It is close to the anchored variant before attack (0.730/0.857) but loses more than half of its goals at the training size (0.293/0.407). Section 5.3 attributes best-response robustness to the anchor rather than to stabilization in general.

Under the retrained attack at the training perturbation size, the pool of fifteen seeds behind the paired test of Section 5.3 gives $0.664_{\pm 0.057}$ tight and $0.815_{\pm 0.046}$ mid for RC-MAT against $0.645_{\pm 0.051}$ and $0.790_{\pm 0.038}$ for the anchored constraint-only variant (paired intervals in Appendix D.11). In the widened pool the original five seeds' $+0.048$ margin at the mid cell disappears. The same pairing on the costs under that attack finds no advantage from training against the adversary on feasibility either. RC-MAT minus the anchored constraint-only variant is $+0.004$ $[-0.001, 0.008]$ on collision and $+0.02$ $[-0.18, 0.22]$ on energy at the tight cell. At the mid cell it is $+0.002$ $[-0.002, 0.005]$ and $+0.15$ $[-0.04, 0.33]$ (agent-mean costs, fifteen shared seeds, paired 95% intervals). Under the attack 14 and 14 of fifteen seeds are feasible in mean at the tight cell and 12 and 14 at the mid cell (RC-MAT and AC-MAT respectively). Competence normalization separates what the anchor retains from its effect on clean learning. On the fifteen seeds per cell of Table 3, the ratio per seed of attacked to clean goal rate at the tight cell is $0.70_{\pm 0.07}$ for RC-MAT and $0.68_{\pm 0.05}$ for the anchored constraint-only variant (mean of the ratios per seed; the ratio of the pooled means is the same to two digits). It is $0.36_{\pm 0.09}$ for the PID-stabilized variant and $0.37_{\pm 0.09}$ for the unanchored variant trained with no warm-up (fifteen seeds each); both reach clean goal rates of 0.74–0.75. For the collapsed unanchored variant it is $0.20_{\pm 0.14}$. Three measurable properties of the frozen policies leave that difference unexplained. A finite-difference norm of the deterministic action with respect to the observation, averaged over 200 visited states, is $4.0_{\pm 1.3}$ for RC-MAT, $3.8_{\pm 1.6}$ for the anchored constraint-only variant, $5.1_{\pm 1.6}$ for PID, and $4.4_{\pm 1.7}$ with no warm-up. These differences are within seed noise (Welch $|t| \leq 1.3$). No variant saturates its action box. A uniform random action perturbation of the attack size (0.1) costs no variant any goal rate (0.99–1.01 of clean). The retrained adversary therefore exploits structure that random perturbation and local sensitivity do not reach. Pairing is by training seed. The two variants share policy initialization and the environment seed stream and differ solely in the adversary, so their trajectories diverge once the adversary acts. Pairing removes initialization and layout variance.

D.2 Agent-count scaling, on-policy

AC-MAT completes 0.92–0.95 of goals from N=2 to N=8 with a per-agent binding energy budget held fixed (collision 0.064, energy 2.5) while the team grows. RC-MAT, the robust variant, completes 0.93–0.96 (Table 14). Both variants are within 0.03 of each other at every tested team size. The largest gap, 0.026 at N=8, has Welch $|t| = 2.0$ on five seeds. The realized energy tracks its budget from below at every size (2.10–2.36 for AC-MAT, 2.28–2.40 for RC-MAT, against 2.5). The three seeds counted infeasible overshoot a budget by at most 0.04 (two on energy, one on collision by 0.003). Collision cost rises with density. An unconstrained, unperturbed MAT reference at N=8, five seeds trained without dual or adversary, completes $0.966_{\pm 0.018}$, AC-MAT $0.922_{\pm 0.023}$, and RC-MAT $0.948_{\pm 0.016}$ (the last two from Table 14). On that backbone the small loss relative to the unconstrained reference travels with the constraint machinery under density, not with the adversary. Collision stays inside its budget at every size, from 0.012–0.018 at N=2 to 0.054–0.058 at N=8 against 0.064. We hold the collision budget slack by design. The scenario's contact floor rises with density (mean collision 0.038–0.043 at N=6 and 0.054–0.058 at N=8 under a binding energy budget). No policy we trained at N=6 or N=8 reaches the collision levels of the N=4 floor. Energy is the binding axis this sweep holds fixed per agent. The scaling sweep of the off-policy instantiation at its own pair $(0.064, 8.6)$ degrades from 0.96 to 0.06 over the same sizes (five seeds).

Table 14: Agent-count scaling of AC-MAT and of the robust variant RC-MAT at a fixed per-agent budget pair with the energy axis binding (collision 0.064, energy 2.5) across team sizes (five seeds, mean$_{\pm\text{s.d.}}$, clean condition). The last column counts seeds whose mean cost is feasible on both dimensions.

variant	N	goal rate	collision	energy	feasible seeds
AC-MAT	2	$0.948_{\pm 0.025}$	$0.012_{\pm 0.009}$	$2.10_{\pm 0.06}$	5/5
	4	$0.942_{\pm 0.028}$	$0.032_{\pm 0.013}$	$2.27_{\pm 0.15}$	4/5
	6	$0.947_{\pm 0.023}$	$0.038_{\pm 0.012}$	$2.36_{\pm 0.09}$	5/5
	8	$0.922_{\pm 0.023}$	$0.054_{\pm 0.008}$	$2.33_{\pm 0.12}$	4/5
RC-MAT	2	$0.928_{\pm 0.058}$	$0.018_{\pm 0.011}$	$2.30_{\pm 0.12}$	5/5
	4	$0.937_{\pm 0.038}$	$0.022_{\pm 0.007}$	$2.28_{\pm 0.12}$	5/5
	6	$0.963_{\pm 0.016}$	$0.043_{\pm 0.014}$	$2.39_{\pm 0.08}$	5/5
	8	$0.948_{\pm 0.016}$	$0.058_{\pm 0.006}$	$2.40_{\pm 0.07}$	4/5

D.3 The binding frontier of the energy budget

Goal rates with a single constraint at 3.0 and 3.5 agree with the grid columns of Table 19 to within 0.01.

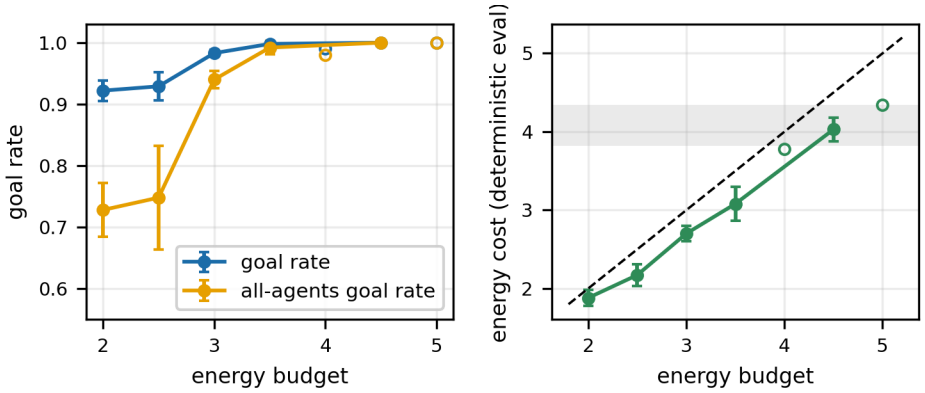

Figure 6: Binding frontier of the energy budget for RC-MAT (single active constraint, $N{=}4$; filled points five seeds, mean±s.d., open points single-seed edge runs). Left: goal rate and all-agents goal rate against the budget. Right: realized deterministic-evaluation energy (agent mean); dashed line: the budget; shaded band: per-seed spend at the two loosest budgets.

Both axes are priced jointly by the grid of Table 19. This sweep isolates what the energy dual does alone and extends the axis below the grid's lowest level. The sweep removes the collision dimension from the surrogate reward and fixes its multiplier at zero (the logged multiplier is exactly 0.0 throughout every run). Energy budgets run from 5.0 down to 2.0. The top of that range is above the policy's natural reach-and-stop spend of ≈4, and the bottom is half that spend. Budgets $\{2.0, 2.5, 3.0, 3.5, 4.5\}$ have five seeds each. The budgets 4.0 and 5.0 are single-seed edge-finding points, drawn open in Figure 6 and entering no statistic.

Figure 6 reports the sweep. The realized energy follows the budget over the whole range, from $4.03_{\pm 0.15}$ at budget 4.5 down to $1.88_{\pm 0.10}$ at 2.0. At every point the seed-mean of the agent-mean cost is inside the budget. The multiplier rises while the policy adapts, then settles small and stationary (tail means 0.03–0.06, drifting downward where it drifts at all). The goal rate falls monotonically, from $1.000_{\pm 0.000}$ at 4.5 through $0.983_{\pm 0.006}$ at 3.0 to $0.922_{\pm 0.017}$ at 2.0. The all-agents rate is the stricter measure and falls from 1.000 to $0.728_{\pm 0.044}$ over the same range. The agent-mean cost is feasible in five of five seeds everywhere except the deepest budget (four of five; the crossing seed is 0.1% over). Away from the edge the worst single agent exceeds its budget marginally (+0.5% in one seed at 3.5, +4.8% in one seed at 4.5). At 2.0 it does so in three of five seeds, by up to +7.5%. At extreme scarcity the dual holds the team average on budget while individual agents ride slightly over. Satisfaction rates add the per-episode view. Per-episode satisfaction stays near one half (0.50–0.57) across the entire sweep, at the slack budgets as well as the tight ones. Episodic energy varies by

more than the margin between the realized mean and the threshold at every point. Episodes therefore straddle a threshold that the mean respects. We judge enforcement of the expectation constraint from the counts per seed above rather than from the diagnostic per episode.

The frontier isolates the cooperative regime, in which all pressure on the budget comes from the task itself and the multiplier can retire once it has reshaped the policy. Competitive tag (Appendix D.6) measures the complementary regime, where an opponent supplies the pressure. There the zero-sum incentive keeps pushing spend upward. A budget sweep from $0.4\times$ to $0.8\times$ the unconstrained scale shows the realized energy tracking the budget with multipliers that stay engaged instead of retiring. That evidence is from the off-policy instantiation. A transformer backbone requires a common per-agent observation width. Tag's heterogeneous pursuer and evader observations violate it. Zero-padding every observation to the common maximum width removes this obstacle. Five padded seeds train stably end to end. At a binding budget (1.0 against a measured unconstrained energy scale of $\sim$1.4) the multipliers engage without divergence. The largest tail component is 5.4 against the ceiling of 20. At the equal horizon the constraint has not yet settled (clean energy $1.26_{\pm 0.46}$ against budget, satisfaction rate $0.61_{\pm 0.16}$). This matches the slow-learning binding-cell profile of Appendix C.5.

D.4 The retrained attack

Our stress test retrains a fresh best-response NatureActor at the training perturbation size for every frozen policy (800 episodes, 25 updates per episode, the same protocol as the local deviation search of Appendix B.7). Best response throughout means an attacker trained for a fixed length, whose objective value need not be optimal. During its 800 training episodes the attacker faces the sampled victim policy and is scored against the deterministic victim policy, the deployment condition of Appendix B.2. An attacker trained against the deterministic victim instead (same recipe, seed, and budget) leaves more of the goal rate at the tight cell: $0.739_{\pm 0.051}$ for AC-MAT and $0.719_{\pm 0.053}$ for RC-MAT, against $0.613_{\pm 0.047}$ and $0.633_{\pm 0.087}$ for the attacker fitted to the sampled policy. Tables report the lower, sampled-policy value. This attacker objective is the training-time object itself. It descends the frozen policy's multiplier-penalized return $r - \lambda^\top c$ with the run's own final multipliers held frozen (a budget-blind run has $\lambda = 0$, a pure reward attack). Attack and training adversary therefore optimize the same functional. We call this attack *Lagrangian-directed*, and the pure-reward control sets $\lambda = 0$. Because the frozen multipliers differ across variants, the objective differs across variants too. We therefore run a λ=0 attack shared by every variant (multipliers set to zero, same trainer and budget) as a control at the tight cell. It changes no variant by more than 0.02 in retained goal rate over five seeds (AC-MAT $0.594_{\pm 0.043}$ against 0.613, RC-MAT $0.616_{\pm 0.051}$ against 0.633, the PID and no-warm-up variants 0.285/0.247 against 0.293/0.263, the collapsed variants 0.02 either way). This ordering of the variants is therefore independent of the attack objective. The collision- and energy-directed attacks of Section 5.3 are additional single-axis stress tests on top of it.

The transformer attacker needs its own control. The off-policy recipe trains the best-response nature on the gradients of critics fitted to the frozen policy. Such critics are absent on the transformer backbone. A Monte-Carlo attacker for it therefore estimates best responses from return-to-go values. That attacker fails a positive control. It leaves a fragile mid-training checkpoint of the plain backbone (clean goal 0.767) at 0.73 under both the standard and an eight-fold attacker training length. That checkpoint lies within the clean-goal range where the fitted-critic attacker below takes the off-policy constraint-only variant from 0.72 to 0.10. We therefore fit TD3-style twin critics to the frozen transformer under attack, with a warm-up of the critics alone replacing the victim-critic warm start. We then let the nature descend their gradients as above. This attacker passes the positive control ($0.767 \to 0.065$) and produces the retrained-attack column of Table 3. At the tight and mid cells RC-MAT degrades monotonically in the attacker perturbation size (Figure 5), and the anchored constraint-only variant follows it at every size, at most 0.05 below. The PID-dual variant loses more than half of its partial pre-attack level ($0.730 \to 0.277$ tight, fifteen seeds). Degenerate before any attack, the unanchored variant falls further ($0.119 \to 0.025$). This attack is also stable in its own initialization. We retrain the best-response nature from two further random restarts against every frozen policy behind these variants and score each policy by its worst restart. Worst-of-three restart scoring lowers the reported goal rates by at most 0.025. Fourfold attacker training (3200 episodes) moves the retention of the default RC-MAT by -0.028 over five seeds. These restarts also measure attacker optimization quality across variants. The per-victim spread over the three independently initialized attackers is as large against RC-MAT as against AC-MAT (mean 0.025/0.013 against 0.015/0.016, tight/mid). Every logged attacker retraining terminates using 0.85–0.93 of its

perturbation budget (mean 0.88 across 250 retrainings). Appendix C.7 tests whether adversarial training against a periodically refit best response widens the retention of the anchored variants. It does not at the mid cell. At the tight cell the paired margin over fifteen seeds is $+0.046$ with an interval $[+0.001, 0.091]$, a small gain at the edge of separability. Checkpoints with similar clean goal rates can have different attacked goal rates. RC-MAT checkpoints chosen at the PID-variant clean level (seven from the pool of fifteen seeds, protocol clean $0.79_{\pm 0.09}$, six mid-training) retain just 0.04–0.07 under the same retrained attack. That lies far below the PID-variant end-of-training 0.293 at equal competence. Best-response retention therefore tracks the final checkpoint's solution rather than its clean goal rate, consistent with the fragile mid-training positive control above.

Training-size frontier. Scarcity lowers retention too: the attack that leaves 0.83 at the mid cell leaves 0.63 at the tight one, where the energy budget has no slack left. We test whether the default training perturbation size ϵ_δ=0.1 is a special point. We retrain the full recipe at the tight cell with the adversary training perturbation size set to 0.05, 0.15, and 0.2 (five seeds at 0.05, ten at 0.15 and at 0.2; every other setting as published). We then attack every run with the retrained attack at the paper's test size 0.1 (Table 15). Retention is flat between 0.05 and 0.1 (0.637 against 0.633). It rises at larger training perturbation sizes, to $0.731_{\pm 0.044}$ at 0.15 (ten seeds) and $0.732_{\pm 0.084}$ at 0.2. Welch t against the default variant is 2.4 and 2.1. Clean goal is unchanged or higher (0.96 and 0.97 against 0.944 at the default size). On this axis the default size therefore sits below a plateau rather than at an optimum. This gain depends on threshold slack. The larger-size variants spend closer to the collision threshold (nine of ten seeds clean-feasible at 0.15 and seven of ten at 0.2, against five of five at 0.05). Retraining either variant against tightened thresholds (0.032, 2.35) pulls its spend and its retention back to the default level. They retain $0.611_{\pm 0.070}$ at 0.15 and $0.685_{\pm 0.100}$ at 0.2. Neither separates from the default variant. Slack alone does not produce the margin either. The anchored constraint-only variant and RC-MAT at the default training perturbation size are trained at the $(0.039, 2.5)$ cell, where realized collision cost matches that of the 0.15 variant (0.026–0.028). They retain $0.620_{\pm 0.068}$ and $0.626_{\pm 0.065}$ under the same attack. The null of Section 5.3 on training against the adversary is a statement at the default perturbation size and spend. A stronger training adversary gains Lagrangian-directed retention jointly with the threshold slack it consumes.

Table 15: RC-MAT at the tight cell $(0.035, 2.5)$ retrained with the adversary training perturbation size varied: clean goal, clean-feasible seeds (agent-mean cost within both thresholds), and goal rate under the retrained attack at the paper's test size 0.1 and at its own training size (mean$_{\pm \text{s.d.}}$). Tightened rows train at thresholds $(0.032, 2.35)$ and are judged at the paper's.

training ϵ_δ	seeds	clean goal	feasible	retrained attack at 0.1	at own ϵ_δ
0.05	5	$0.956_{\pm 0.014}$	5/5	$0.637_{\pm 0.042}$	$0.851_{\pm 0.030}$
0.1 (default)	5	$0.944_{\pm 0.023}$	5/5	$0.633_{\pm 0.087}$	$0.633_{\pm 0.087}$
0.15	10	$0.959_{\pm 0.020}$	9/10	$0.731_{\pm 0.044}$	$0.503_{\pm 0.078}$
0.15, tightened	5	$0.935_{\pm 0.048}$	4/5	$0.611_{\pm 0.070}$	$0.377_{\pm 0.101}$
0.2	10	$0.970_{\pm 0.023}$	7/10	$0.732_{\pm 0.084}$	$0.182_{\pm 0.117}$
0.2, tightened	5	$0.943_{\pm 0.031}$	4/5	$0.685_{\pm 0.100}$	$0.154_{\pm 0.119}$

Off-policy read. Table 7 places the clean, transferred-attack, and retrained-attack goal rates of the off-policy variants side by side. Cells with a single constraint repeat the pattern (RC-MATD3 and AC-MATD3 clean 0.17/0.18 tight and 0.36/0.73 mid, 0.01–0.10 under the retrained attack). The retrained attack collapses both variants at every cell, consistent with fixed deep policies admitting adversarial policies (Gleave et al., 2020). Feasibility is unbroken where task completion is not. Under the attack the agent-mean costs stay within budget, at 0.63–0.80$\times$ of the collision budget and 0.63–0.74$\times$ of the energy budget across both variants and all three cells (feasible in mean in 29 of 30 runs). Worst-dimension satisfaction stays at 0.58–0.76.

Table 16: Budget accounting at the tight cell $(0.035, 2.5)$ for AC-MAT and RC-MAT (five seeds each, paired attacker protocol). Cost columns are multiples of the corresponding budget; the cost-directed rows retrain the attacker against one cost dimension at the full perturbation size; every attacker is fitted to the sampled policy (Appendix D.4). † as in Table 3.

condition	collision (×budget)	energy (×budget)	goal rate
AC-MAT			
clean	$0.61\times$	$0.93\times$	$0.943_{\pm 0.022}$
retrained attack (Lagrangian-directed)	$0.62\times$	$0.64\times$	$0.613_{\pm 0.047}$
collision-directed attack	$2.22\times^{\dagger}$	$1.15\times^{\dagger}$	$0.893_{\pm 0.023}$
energy-directed attack	$1.22\times^{\dagger}$	$1.46\times^{\dagger}$	$0.995_{\pm 0.005}$
RC-MAT (robust variant)			
clean	$0.66\times$	$0.90\times$	$0.944_{\pm 0.023}$
training-time adversary ($1\times\epsilon_\delta$)	$0.71\times$	$0.89\times$	$0.938_{\pm 0.038}$
retrained attack (Lagrangian-directed)	$0.70\times$	$0.68\times$	$0.633_{\pm 0.087}$
collision-directed attack	$2.11\times^{\dagger}$	$1.15\times^{\dagger}$	$0.899_{\pm 0.033}$
energy-directed attack	$1.31\times^{\dagger}$	$1.42\times^{\dagger}$	$0.994_{\pm 0.005}$

Cost-directed attacks. The attacker also admits a cost-directed objective: we retrain the nature to maximize one cost dimension of the frozen policy. At the tight and mid cells and at the full perturbation size, the collision-directed attack drives the mean collision cost of AC-MAT to $2.07\times$ its budget at the tight cell and $2.11\times$ at the mid cell, at goal rates of 0.90 and 0.95 (single-attacker runs, five seeds per cell). Our paired-attacker protocol of Table 16 gives the values quoted below for the tight cell. At the tight cell the energy-directed attack extracts $1.46\times$ energy from AC-MAT and $1.42\times$ from RC-MAT at goal rates above 0.99 (Table 16). Policies concede constraint margin while continuing to play. Co-training adds no margin on this axis either: under the same single-attacker runs the robust variant is at $2.20\times/1.99\times$ under the collision-directed and $1.41\times/1.32\times$ under the energy-directed best response at the same two cells. Under the paired protocol of Table 16 at the tight cell, AC-MAT is at $2.22_{\pm 0.41}\times$ and $1.46_{\pm 0.04}\times$ against $2.11_{\pm 0.17}\times$ and $1.42_{\pm 0.09}\times$ for the robust variant (paired seed differences $+0.11_{\pm 0.28}$ and $+0.04_{\pm 0.07}$). At the resolution of five seeds, RC-MAT thus shows no margin over AC-MAT on the cost-directed axis.

Energy-directed extraction stays within a narrow 1.39–$1.46\times$ band for every dual-regulated variant whatever its stabilizer or adversary (PID $1.39\times$, unanchored $1.40\times$, adversarially trained unanchored $1.41\times$). This attack drives the budget-blind backbone to $11.2\times$. Collision-directed extraction instead varies by variant, from $1.06\times$ on the PID variant through 2.1–$2.2\times$ on the working anchored variants to 2.1–$3.0\times$ with seed spreads above 1.1 on the degenerate unanchored variants. Under this attack class the anchor and the adversary neither create nor remove the vulnerability. A jointly adaptive attacker ascends per agent whichever cost dimension currently shows the largest normalized violation (re-picked throughout its training). At the tight cell it converges to the energy axis for every agent and seed. It extracts $1.41\times/1.44\times$ energy against the full and anchored constraint-only variants, inside the energy-directed band above. Against both variants a forced collision-directed attack remains the strongest of those tested. Paired over shared seeds, the costs of the two variants under this adaptive attacker differ by $-0.000_{\pm 0.006}$ collision and $-0.06_{\pm 0.19}$ energy (95% intervals straddling zero). We measure how collision overshoot grows with the perturbation size on the earlier and looser operating pairs $(0.039, 5.2)$ and $(0.039, 8.6)$. The perturbation budget sweeps over $\{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\} \times \epsilon_\delta$ (five seeds per point, protocol otherwise unchanged). Those cells give the response shape uncontaminated by the scarcity of the tight and mid cells. Its response is a smooth monotone ramp free of any cliff. Collision runs from $0.90\times$ to $1.73\times$ of budget at $(0.039, 5.2)$ and from $0.94\times$ to $1.71\times$ at $(0.039, 8.6)$. It is at $1.02\times$ and $1.09\times$ at one quarter of ϵ_δ, with goal rate at least 0.998 at every point. The constraint-only variant at $(0.039, 8.6)$, deep inside the budget unattacked ($0.45\times$), stays feasible across the sweep. The plain backbone, in violation without any attack, moves within 1.2–$1.8\times$ across the sweep (little margin left to extract). The anchored variants overshoot in the same band as RC-MAT (1.55–$1.69\times$), and energy-directed attacks leave collision at 0.92–$1.23\times$. RC-MATD3 under the same attack at $1 \times \epsilon_\delta$ at those cells moves to $1.11\times/1.24\times$ of budget with goal rates $0.22/0.36$, unchanged from its clean levels ($0.22/0.34$). On this axis neither instantiation preserves strict feasibility at the full perturbation size. On-policy, the variant cedes bounded, budget-graded margin while continuing to complete goals. A final test asks whether this breach is an artifact of the Lagrangian-directed training adversary. We retrain the $(0.039, 5.2)$ RC-MAT with the nature

ascending the discounted collision cost itself (five seeds, policy and dual updates unchanged). The result is unchanged (clean $0.999_{\pm 0.002}$, collision-directed overshoot $1.67\times$, 0.994 under the retrained attack). A cost-directed training adversary therefore leaves the gap open. Neither does an adaptive one always ascending the currently most-violated dimension (five seeds: clean 0.967, retention 0.652, overshoot $1.94\times$/$1.42\times$). The graded shape admits a deployment-side mitigation: a practitioner holding a hard threshold can train against a tightened budget and absorb the attack inside the reclaimed margin. On the binding frontier (Appendix D.3), an energy-directed attack retrained against every frontier checkpoint at the full perturbation size extracts a roughly constant absolute increment over realized spend. Attacked energy is $1.15\times$ of budget at 4.5 and $1.62\times$ at 2.0 (five seeds per budget, goal rates under attack at least 0.99). These are increments of $+1.1$ and $+1.4$ energy over realized spends of 4.03 and 1.88. Policies trained at 3.0 hold the 4.5 threshold under the same attack in five of five seeds, at a frontier price of 1.7 points of goal rate. The reclaimed margin must exceed the absolute extraction of the attacker rather than a fixed fraction of the budget. Backing off 3.5 to 2.5 holds in three of five seeds and 3.0 to 2.0 in none. The mitigation is thus available at moderate thresholds and closes out at extreme scarcity, where within the tested range no back-off reclaims more than the attacker extracts.

D.5 The duals hold on cooperative coverage (discovery)

To confirm the navigation findings in a task with different mechanics, we repeat the protocol of Section 5.2 on the VMAS *discovery* scenario. Five agents must jointly cover five distinct targets (one agent per target, cover-once semantics). The same $J{=}2$ per-agent constraints (collision, control energy) and the same adversarial nature apply. This environment passes the admission gate. The unconstrained calibration run reaches a coverage rate of 0.94–0.98 without the adversary and 0.61–0.63 with it, a $\sim 35\%$ drop that leaves the task solvable. Budgets are sized at $\{0.6, 1.0, 1.5\}\times$ the measured unconstrained cost per dimension (diagonal cells, five seeds each).

Across the three budget levels the achieved energy cost tracks the budget monotonically (4.1 ± 0.4, 7.2 ± 0.6, 8.8 ± 0.7 against budgets 8, 13.5, 20). Every agent in every seed remains feasible in mean cost on both dimensions (the per-episode worst-agent, worst-dimension satisfaction rate is 0.76–0.78 across the cells). Coverage stays at 0.52–0.58 against the adversarial ceiling of ~ 0.62. As on navigation at budgets above the natural scales, the dual acts transiently. The energy multiplier rises early while the policy is still inefficient and peaks higher at tighter budgets; individual agents' multipliers reach up to 1.9 at the tight cell against 1.0 at the slack cell. The multiplier shapes an efficient operating mode and returns to its floor once the shaped policy is feasible. The grid therefore reproduces the slack regime of the navigation frontier (Appendix D.3) in a coverage task with different reward mechanics.

The on-policy instantiation (Section 5.2) at the same cells reaches a coverage rate of 0.83–0.85 clean and attacked with every seed feasible. Its energy stays at natural spend (4.7–5.1) inside every budget and its multipliers at zero. These slack cells are thus a transfer control rather than a test of constrained stabilization. An anchored constraint-only variant retrained at the tight cell (five seeds, RC-MAT without the adversary) reaches a coverage rate of 0.79 ± 0.09 against 0.83 ± 0.09 for RC-MAT. That is a difference of -0.04 with a 95% Welch interval of ± 0.13 around it. At that cell the energy cost stays at 3.9 against a budget of 8 and the multipliers stay at zero. An energy budget of 2.5, below that natural spend, with the collision budget at 0.018 defeats both variants (five seeds each, default protocol). AC-MAT and the unanchored variant end at a coverage rate of 0.000 with four and three of five seeds feasible for every agent. An intermediate energy budget of 3.2, above the 2.5 one but still below the 3.9 the working policies spend, defeats both variants in the same way (coverage 0.000 in five of five seeds each). No tested energy budget on this task is both binding and attainable. At the three reported discovery cells both variants keep every multiplier at zero throughout, so the contrast isolates the adversary, which adds no separable margin here either. Table 17 collects the anchor effect on every task with an unanchored variant.

Table 17: The anchor helps on navigation (with and without warm-up) and on pursuit–evasion, and nowhere else. Score (goal or coverage rate; tag: catches; MuJoCo: reward) and feasible seeds of the final policy, by agent-mean cost within every budget over 50 clean episodes (discovery: every agent on the 500-episode tail; pursuit–evasion: both agents, 500 episodes, NatureActors active); n/r: not reported.

			anchored		unanchored		
task (budgets)	variants	seeds	score	feasible	score	feasible	unanchored collapse
navigation $(0.035, 2.5)$	AC-MAT / −Anchor	15	$0.947_{\pm 0.020}$	12/15	$0.123_{\pm 0.069}$	10/15	yes
navigation $(0.035, 2.5)$, no warm-up	AC-MAT / −Anchor	5 / 15	$0.954_{\pm 0.014}$	2/5	$0.743_{\pm 0.097}$	15/15	no
pursuit–evasion, symmetric	ζ=10 / 0	5	n/r	4/5	n/r	1/5	yes
discovery (0.6× scale)	AC-MAT / none	5	$0.79_{\pm 0.09}$	n/r	n/r	n/r	n/r
discovery $(0.018, 2.5)$	AC-MAT / −Anchor	5	0.000	4/5	0.000	3/5	both
dispersion $(135, 35)$	AC-MAT / −Anchor	5	$0.136_{\pm 0.056}$	5/5	$0.168_{\pm 0.053}$	5/5	no
dispersion $(135, 35)$	RC-MAT (anchor and adversary)	5	$0.298_{\pm 0.131}$	3/5	n/r	n/r	n/r
tag (energy 35)	RC-MATD3, ζ=10 / 0	5	$36.5_{\pm 25.6}$	en. 1.03×	17–50	en. 1.00–1.06×	no
Walker2d 2×3 (4.86)	constraint-only	5	n/r	1/5	n/r	3/5	no
HalfCheetah 2×3 (1.94)	constraint-only	5	2414	4/5	2354	4/5	no

The last column records whether the unanchored constrained variant collapses or enters the runaway basin once the dual is released. Where it does (navigation with warm-up, pursuit–evasion) the anchor restores training. Where it trains through the release the two variants are within seed noise on every task but navigation without warm-up. There the unanchored variant reaches 0.743 and the anchored one 0.954 (five seeds, two of five feasible in agent mean).

D.6 The duals pool over the energy budget on competitive tag (MPE)

On competitive tag, where the opponent rather than the task pressures the budget, realized energy tracks budgets swept from 0.4× to 0.8× the unconstrained scale (off-policy; details below). To complement the pursuit–evasion domain with a competitive benchmark, we run the MPE *simple tag* scenario as ported by VMAS. Three cooperating pursuers chase one evader in an open arena, with a zero-sum contact reward and distance shaping for both teams. We score play by catches (pursuer–evader contact steps) per episode. Each of the four agents has a single control-energy constraint (J=1) under the same adversarial nature. This game passes the admission gate. Engagement calibration runs sustain 25–53 catches per episode without the adversary and 20–45 with it. Under the adversary the unconstrained energy scale is ≈59 per agent and episode.

We train five seeds at an energy budget of 35, that is 0.6× the unconstrained scale. The budget binds and stays contested (Figure 7). Across the five seeds the achieved energy averages 35.9 against the budget of 35. Means over agents within a seed range from 32.5 to 42.1, and means for a single agent in a single seed from 26 to 44. Two seeds are taken at their extended horizon, below. The episodic cost distribution therefore straddles the threshold (per-episode satisfaction rate 0.31–0.82). In the cooperative tasks the multipliers act transiently and return to their floor once an efficient policy has been shaped. Here, by contrast, the endpoint multipliers remain engaged: in 16 of the 20 agent–seed pairs the final multiplier is interior (between 0.6 and 18.7). All four remaining are at the ceiling of 20, in one seed. A zero-sum contact incentive keeps pushing energy expenditure up. The dual therefore holds the policies at the constraint boundary rather than retiring after a transient. Under the drift check

of Appendix B.7 two of the five seeds are flagged on a multiplier component at 6000 episodes. Both consume the full two-extension resume budget (12,000 episodes) and remain flagged. We therefore report their dual allocation as non-stationary within 12,000 episodes and take endpoint statistics from the extended runs. In one of the two flagged seeds, the cost and catch series are stationary throughout while the split of the shared energy load across the four multipliers keeps migrating. In the other flagged seed, all four multipliers are at the ceiling in the endpoint tally above. There the extensions drive all four multipliers to the ceiling with energy 20% above budget (42.1 against 35). Play remains engaged under this budget, ∼40% below the unconstrained energy scale. Catches average 36.5 ± 25.6 across seeds against 20–53 for the unconstrained calibration runs, with a large seed spread in both protocols. Our test does not separate RC-MATD3 from AC-MATD3 trained under the same protocol, extensions included. Four of its five seeds remain flagged after two extensions. It gives catches 28.4 ± 6.5 (Welch t=0.7), energy means over agents within a seed 35.2–40.0, and multipliers at the ceiling in several agent–seed pairs. This saturation thus comes from the dual at this budget, not from training against the adversary.

A budget sweep tests whether the engaged dual responds to the threshold rather than to one setting. Two single-seed runs at budgets 24 and 47 ($0.4\times$ and $0.8\times$ the unconstrained scale) bracket the cell run with five seeds, enter no statistic, and mirror the edge points of Appendix D.3. Realized energy tracks the budget across the sweep. Per-agent means are 20.7–26.0, agent mean 23.7, at budget 24 and 40.0–47.7, agent mean 42.9, at 47, against 35.9 at 35. Per-episode satisfaction straddles the threshold throughout (0.40–0.75). Multipliers follow the contest for the budget. At 24 all four finish interior (2.2–8.8). At 47 the one agent riding its threshold (1.5% over) has a multiplier of 8.3 and the next-closest agent (8% inside) one of 2.6. Both agents well inside budget rest at the floor. Play stays engaged at both budgets (100 and 71 catches per episode against the $36.5_{\pm 25.6}$ average over five seeds at 35). Given that spread, catch counts are an engagement check rather than a graded score.

This engaged-dual regime is itself independent of the anchor. Retraining the cell's five seeds with the anchor removed (ζ=0) and every other setting identical produces no runaway. Realized energy stays within 6% of budget (agent means per seed 1.00–$1.06\times$). The largest final multiplier component across the twenty agent–seed pairs is 5.7 against the ceiling of 20 (one rests at the floor). Play stays engaged at 17–50 catches per episode. At the shared 6000-episode budget, then, the multiplier runaway that motivates the anchor on pursuit–evasion (Appendix C.7) does not transfer to the tag benchmark. Over that horizon the dual holds the contested boundary with and without the anchor, apart from the seed whose multipliers reach the ceiling under extension. The competitive-stability evidence for the anchor is thus specific to pursuit–evasion rather than generic to competitive pressure. Without the anchor, the variant extended to the same doubled horizon on its three drift-flagged seeds does not saturate: final multiplier components lie in 3.6–17.9 against the ceiling of 20, though their means still rise over the extension ($5.2 \to 7.9$, $5.3 \to 10.1$, $5.4 \to 9.3$). Per-agent energy there is 31–34 on one seed, 32–37 on another, and 32–40 on the third. Feasibility at the doubled horizon is therefore seed dependent without the anchor as well. At this horizon the single anchored saturation thus lacks an anchor-free counterpart.

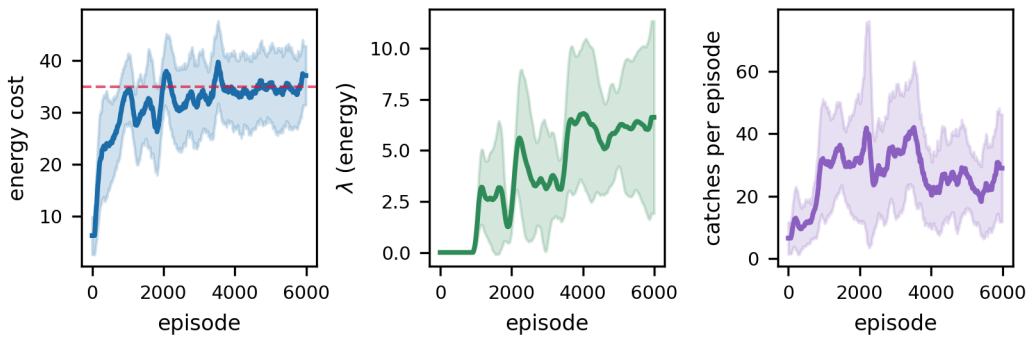

Figure 7: Training dynamics on the tag benchmark (N=4, J=1, five seeds, mean±s.d. over the shared 6000-episode budget, agent-averaged, smoothed over a 150-episode window). Left to right: energy cost (budget dashed), energy multiplier, catches per episode.

D.7 THE DUALS HOLD ON SPATIAL DISPERSION AND THE ANCHOR GIVES NO ADVANTAGE

In the VMAS *dispersion* scenario four agents spawn together at the arena center and must spread to four scattered food targets. The shared spawn keeps the collision cost high, and budgets follow the unconstrained calibration run's scale of $(135, 35)$ per agent. At the $1.0\times$ budget (five seeds, 20,000 episodes) AC-MAT collects $0.136_{\pm 0.056}$ of the food and the unanchored constraint-only variant $0.168_{\pm 0.053}$, both feasible in every seed. Duals thus hold and the anchor gives no advantage (Table 17). The robust variant RC-MAT collects $0.298_{\pm 0.131}$ (Welch $t = 2.5$ against AC-MAT) but is within budget in agent-mean cost in three of five seeds and in every agent in one, with collision of 154 pooled over seeds against the budget of 135. On this task the adversary rather than the anchor raises coverage, at the price of feasibility.

D.8 ADMISSION GATE

Beyond the domains of Section 5 and Appendices D.5 and D.6, we evaluate three further VMAS scenarios (two corridor give-way variants and narrow-passage traversal) under the same protocol. A task enters the pool when an unconstrained calibration run reaches a stable task-success rate of at least 0.5 on the off-policy backbone or at least 0.3 on the transformer, on which the later pool was formed; a rate pinned by the task structure at a fixed fraction of agents does not count. Both bars were set before the tasks of this section were tried, and all three fail. (The multi-agent give-way variant plateaus with exactly half the agents ever completing their crossing, invariant to the success metric and to a doubled horizon.) Three open-arena candidates fail the same gate. Package transport reaches task success 0.00 across package masses, with and without an added potential-based approach-shaping term. On the MPE spread benchmark the plain off-policy backbone reaches a goal rate of 0.10–0.12 unconstrained in the four-agent configuration, and the transformer completes no goals in the three-agent configuration of Zhang et al. (2025) at $1\times$ or $50\times$ our interaction budget. This task thus leaves the pool. Dynamic ball balancing has an unconstrained success rate of 0.10 or below on every calibration run despite healthy shaped returns. A fourth candidate is spatial dispersion. It fails the gate for the off-policy backbone (unconstrained ceiling 0.34–0.41 with an unstable return) but passes it on the transformer. That calibration run holds a stable 0.33–0.36. Its constrained results are in Appendix D.7. We use open-arena tasks for the reported action-perturbation experiments. Inside a corridor an adversary that perturbs the executed action degenerates to the single direction opposing travel.

D.9 FULL CONSTRAINT-GRID RESULTS

Collision levels $\{0.030, 0.035, 0.039\}$ approach the $\approx$0.03 collision cost measured for goal-reaching policies (an operating scale, not a bound). Energy levels $\{2.5, 3.0, 3.5\}$ lie below the $\approx$4 energy an unconstrained goal-reaching calibration run spends. In Table 3 the plain backbone spends 27 because nothing limits its motion. Min-dim sat. in Table 20 is defined in Appendix C.2. For the grid of the robust variant (Table 19), the 500-episode tails give a clean count of nine of nine when seeds are pooled. At $(0.030, 3.5)$ the pooled collision is 0.029 against 0.031 at 50 episodes. Per-seed counts move by up to two seeds in either direction. Table 19 keeps the 50-episode protocol's eight of nine. The mid cell $(0.035, 3.0)$ repeats the pattern of the tight cell in clean goal rate: RC-MAT $0.983_{\pm 0.010}$, unanchored 0.186 (0.213 in the protocol of the retrained attack at zero size), PID $0.812_{\pm 0.054}$, and plain 0.938 with budgets violated. In Table 19 the training-time attack leaves costs and goal rate within one s.d. of clean in every cell. The breaches are the attacked collision of 0.033 at $(0.030, 3.0)$ and the clean collision of 0.031 at $(0.030, 3.5)$, both daggered there. At the collision-0.030 tier, where the counts over five seeds are lowest, ten further seeds per cell give mean-feasible counts of 11, 8, and 6 of fifteen at energy 2.5, 3.0, and 3.5. Their collision pooled over seeds is 0.0249, 0.0291, and 0.0325 against 0.030, and Table 19 keeps the first five seeds. For the ablation at the tight cell, the clean-goal differences have Welch 95% intervals: RC-MAT minus the unanchored constraint-only variant 0.862 $[0.829, 0.895]$, anchored minus unanchored constraint-only 0.820 $[0.780, 0.859]$, and anchored minus PID-stabilized constraint-only 0.192 $[0.133, 0.250]$, all in clean goal rate. Table 20 reports the numbers per cell of the off-policy budget grid: achieved cost per dimension, min-dimension satisfaction rate, and return for every cell of the 3×3 budget grid, averaged over five seeds.

For the grid of the robust variant (Table 19), a harsher criterion counts a seed just when its one-sided 95% upper confidence bound on both agent-mean costs is within budget. It gives 3–5 of five at the 0.035/0.039 collision levels and 0–2 at the floor-riding 0.030 level (the anchored constraint-only variant alike). The per-agent criterion requires each agent's own 500-episode mean with its one-sided 95% upper confidence bound within budget on both dimensions. Every bound is a marginal normal bound on one dimension (mean plus 1.645 standard errors over the episodes). A pair passes when both marginal bounds are within budget, a 95% criterion per dimension. Under it the feasible agent and seed pairs per cell are 15/14/4 of twenty across the 0.035 level's energy cells, 13/12/11 at the 0.039 level, and 7/3/1 at the 0.030 level. Threshold-riding play leaves many individual agents' bounds straddling the budget even where pooled means are interior. The collision dimension accounts for nearly every exceedance except at $(0.039, 2.5)$, where energy fails in seven of twenty. Under a criterion requiring every agent and seed bound within budget, no cell passes. The best cell reaches fifteen of twenty, the ceiling of the per-agent high-probability reading. Per trained policy instead of per pair, 32 of the 45 policies keep every agent's 500-episode mean cost within both budgets. The seed counts of five are 3/2/2, 4/5/3, and 4/5/4 across the energy levels 2.5/3.0/3.5 at the collision levels 0.030, 0.035, and 0.039 respectively. By the strict criterion 11 of 45 keep every agent's one-sided 95% upper bound within budget (0/0/0, 3/1/0, 3/2/2). A practitioner selecting a seed without evaluation data can use the dual's own training-time rollouts. Over the last 100 deterministic dual evaluations (two episodes each, the last 1000 training episodes), every agent within budget at training time agrees with the 500-episode evaluation's all-agent verdict on 35 of the 45 grid policies. Five pass at training time but fail in evaluation, five the reverse. Across policies the correlation between training-time and evaluated cost is 0.97 for energy and 0.52 for collision. Per policy we take the worst agent's mean cost relative to its budget. Across the 45 grid policies its median is 0.93 on collision and 0.92 on energy. Its 90th percentiles are 1.10 and 0.97 and its maxima 1.21 and 1.02. Twelve policies have at least one agent over the collision budget and one policy an agent over the energy budget. A seed passes just when every agent's 500-episode mean cost is within budget on both dimensions. Under it the 0.035 and 0.039 collision levels keep 3–5 of five seeds feasible and the 0.030 level 2–3. Also requiring the worst agent's one-sided 95% upper bound within budget drops this to 0–3 of five, the strictest criterion the data supports. For AC-MAT, pooled over the five seeds of each cell (2,500 episodes per agent), every per-agent one-sided 95% upper bound is within budget except the collision bounds at the 0.030 level and at $(0.035, 3.0)$. Among these the largest exceeds budget by 14%. Per trained policy, 30 of the 45 AC-MAT grid policies pass when every agent of a seed is within both budgets on its 500-episode tail: 3/1/2 of five across the energy levels 2.5/3.0/3.5 at the 0.030 level, 5/3/2 at the 0.035 level, and 5/5/4 at 0.039. The largest single-agent mean exceedance is 1.23× the collision budget at the 0.030 level. No agent exceeds 0.98× at the tight cell or at $(0.039, 2.5)$ and $(0.039, 3.0)$.

Selecting a deployable policy. Two selection rules on the same 45 AC-MAT policies quantify what a practitioner can do about per-agent feasibility. A training-time rule accepts a policy when every agent's mean cost over the last 20 deterministic dual evaluations (two episodes each, spanning the last 200 training episodes) is within budget on both dimensions. It accepts 20 of 45, of which 14 are all-agent feasible on the 500-episode tail (0.70, against a base rate of 0.67). The training-time rule therefore barely improves on the base rate. Tightening the training-time margin to 0.95 or 0.90 of the budget leaves the precision unchanged. An evaluation-based rule with 50 held-out clean episodes per candidate accepts 18 of 45, of which 16 are all-agent feasible on the remaining 450 episodes (0.89). Their clean goal rate is 0.970 against 0.974 for all candidates. At 100 held-out episodes it accepts 29 with 23 feasible (0.79). A third procedure trains with a margin. AC-MAT at 0.91× and 0.94× of the tight cell's budgets (0.032, 2.35), judged at the paper's budgets, is feasible for every agent on the 500-episode tail in five of five seeds. Its worst single-agent mean is at 0.93× its budget (against 0.98× for the untightened pool), at a clean goal rate of $0.910_{\pm 0.027}$ against 0.943. This margin costs about 0.03 of goal rate and gains 0.05 of budget headroom per agent.

Performance at equal all-agent feasibility. The same all-agent criterion applied to the competing variants at the tight cell (fifteen seeds each, 500-episode held-out tails) separates reliability from goal rate. AC-MAT is feasible for every agent in 14 of 15 seeds and its feasible seeds complete 0.949 of goals; the PID variant is feasible in 13 of 15 at 0.731; the unanchored variant trained with no warm-up is feasible in all 15 at 0.743. Counting an infeasible policy as a failure (feasible fraction times the goal rate of feasible seeds) gives 0.886, 0.634, and 0.743; the training-time rule above accepts 5, 4, and 11 of the fifteen with 5, 3, and 11 of those feasible on the tail. Most reliably feasible is the no-warm-up variant. The anchored variant keeps an advantage of 0.20 in goal rate and 0.37 in

retention over it (0.645 against 0.274 under the retrained attack) at one fewer feasible seed. Across the grid the trade is the same (Table 18; five seeds per cell for the two competing variants, all judged on 500-episode tails). The no-warm-up unanchored variant is feasible for every agent in 44 of 45 policies and never exceeds a budget by more than 3%, PID in 32 of 45 with exceedances up to $1.69\times$, and AC-MAT in 30 of 45 with exceedances up to $1.23\times$. Their clean goal rates run 0.70–0.85, 0.72–0.90, and 0.94–0.99. Per-agent reliability and goal rate trade against each other across the schedule. The anchor raises the second at a cost in the first, and a training margin (above) recovers that cost at the tight cell. One predetermined margin rule trains every cell at $0.91\times$ its collision and $0.94\times$ its energy budget and judges at the paper's budgets (five seeds per cell). It recovers part of that cost without tuning per cell: AC-MAT rises from 30 to 34 of 45 policies feasible for every agent at clean goal rates of 0.91–0.99. It has five of five at the three cells where energy binds hardest. The shortfall concentrates where collision is the binding axis at a loose energy budget (1 and 3 of five at $(0.030, 3.5)$ and $(0.035, 3.5)$, worst agents at 1.09–$1.10\times$). On the heterogeneous assignments the same rule moves the first from none to three of five seeds feasible for every agent ($0.952_{\pm 0.007}$ clean) and leaves the second at none of five (0.978, worst agent $1.18\times$). It thus still trails the no-warm-up variant in reliability there.

Goal rate at matched reliability. The competing variants receive the same margin rule and AC-MAT a tighter one, so each method traces its own reliability–goal curve (five seeds per cell, judged at the paper's budgets on 500-episode tails). Trained at $0.91\times$ and $0.94\times$ of each cell's budgets, the PID variant rises from 32 to 38 of 45 policies feasible for every agent at a mean clean goal rate of 0.797 (from 0.819). Trained with no warm-up, the unanchored variant rises from 44 to 45 of 45 at 0.781 (from 0.798). Five fresh training seeds at the same margin give 40 of 45 at 0.774 for the PID variant (78 of 90 over both seed sets) and 44 of 45 at 0.762 for the no-warm-up variant (89 of 90). Trained at $0.85\times$ and $0.88\times$, a second margin chosen after the $0.91\times$/$0.94\times$ rule stopped at 34, AC-MAT reaches 39 of 45 at 0.954 (from 30 at 0.974 and 34 at 0.962) with five of five in six of the nine cells. At $0.80\times$ and $0.83\times$, the third margin tried, it reaches 45 of 45 at 0.933 with five of five in every cell. Five fresh training seeds at the same margin give 41 of 45 at 0.932 (86 of 90 over both seed sets); the four misses are at the 0.030 collision level, with a worst agent at 1.01–$1.27\times$ its collision budget. Two reliability levels fix the comparison: 44 of 45 and 40 of 45, the fresh-seed levels of the no-warm-up and PID variants. By linear interpolation between its margin points (curve positions, not trained policies) AC-MAT completes 0.937 of goals at 44 of 45 and 0.950 at 40 of 45, against 0.798 for the no-warm-up variant at 44 of 45. On its first seed set the PID variant reaches neither level (its 38 of 45 point has goal rate 0.797). At full reliability AC-MAT completes 0.933 and the no-warm-up variant 0.781 (fresh seeds: 0.932 at 41 and 0.762 at 44 of 45). At that margin point the anchored policies thus complete about 0.15 more goals. Both other curves stay within 0.78–0.82 at every margin. This margin costs retention. At the tight cell the $0.80\times$/$0.83\times$ AC-MAT policies (ten seeds, all ten feasible for every agent) complete $0.887_{\pm 0.024}$ of goals clean and retain $0.501_{\pm 0.055}$ under the retrained attack at b=0.1, against 0.645 at the paper's budgets. The no-warm-up variant at its $0.91\times$/$0.94\times$ margin retains $0.275_{\pm 0.048}$ (five seeds, all feasible), the same as at the paper's budgets. On the heterogeneous assignments the $0.85\times$/$0.88\times$ margin takes AC-MAT to five of five seeds on the first (0.946) and two of five on the second (0.965). A $0.80\times$/$0.83\times$ margin takes it to five of five (0.907) and four of five (0.944). The PID variant reaches four and two of five (0.694 and 0.863) and the no-warm-up variant four and five of five (0.679 and 0.819). A selection rule that accepts a trained policy just when every agent is within budget on its held-out tail gives the same goal-rate gap without retraining. Over the untightened grid policies it accepts 30 AC-MAT policies at a mean clean goal rate of 0.971, 34 of the $0.91\times$/$0.94\times$ AC-MAT policies at 0.956, 32 PID policies at 0.825, and 44 no-warm-up policies at 0.798: the same 0.15 of goals, paid in acceptance rate.

Table 18: Across the budget grid the unanchored no-warm-up variant is the most reliably feasible per agent and AC-MAT the highest in goal rate. Clean goal rate, seeds with every agent within budget on the 500-episode tail, and the largest single-agent mean cost over its budget; five seeds per cell.

	AC-MAT			−Anchor+PID			−Anchor, no warm-up		
budgets	goal	all agents	worst	goal	all agents	worst	goal	all agents	worst
(0.030, 2.5)	$0.954_{\pm 0.020}$	3/5	1.02×	$0.732_{\pm 0.081}$	4/5	1.07×	$0.699_{\pm 0.129}$	5/5	0.96×
(0.030, 3.0)	$0.980_{\pm 0.009}$	1/5	1.23×	$0.844_{\pm 0.055}$	2/5	1.69×	$0.847_{\pm 0.064}$	4/5	1.03×
(0.030, 3.5)	$0.994_{\pm 0.005}$	2/5	1.23×	$0.846_{\pm 0.050}$	3/5	1.41×	$0.836_{\pm 0.083}$	5/5	0.96×
(0.035, 2.5)	$0.943_{\pm 0.022}$	5/5	0.96×	$0.757_{\pm 0.082}$	4/5	1.02×	$0.803_{\pm 0.087}$	5/5	0.96×
(0.035, 3.0)	$0.984_{\pm 0.011}$	3/5	1.13×	$0.828_{\pm 0.067}$	4/5	1.05×	$0.736_{\pm 0.072}$	5/5	0.88×
(0.035, 3.5)	$0.990_{\pm 0.008}$	2/5	1.09×	$0.879_{\pm 0.024}$	5/5	0.98×	$0.827_{\pm 0.053}$	5/5	0.98×
(0.039, 2.5)	$0.955_{\pm 0.016}$	5/5	0.98×	$0.717_{\pm 0.126}$	2/5	1.04×	$0.768_{\pm 0.112}$	5/5	0.97×
(0.039, 3.0)	$0.984_{\pm 0.009}$	5/5	0.97×	$0.875_{\pm 0.045}$	4/5	1.02×	$0.832_{\pm 0.057}$	5/5	0.97×
(0.039, 3.5)	$0.986_{\pm 0.012}$	4/5	1.17×	$0.895_{\pm 0.063}$	4/5	1.02×	$0.836_{\pm 0.103}$	5/5	0.91×

RC-MAT, the robust variant with a training-time NatureActor, tracks the same grid. Table 19 lists it at every cell. Its clean goal rate is within 0.01 of AC-MAT in all nine cells: slightly below in three, and 0.004/0.009 above at the two loosest-collision high-energy cells, where its seed spread is 0.002–0.004.

Table 19: The robust variant RC-MAT across the same grid: seed-pooled clean mean cost stays inside both budgets in eight of nine cells, and a different single cell fails under the training-time attack (five seeds per cell, mean$_{\pm \text{s.d.}}$). Att.: the training-time adversary at $1\times\epsilon_\delta$. Feasible and UCB: as in Table 4 (seeds of five). †: over budget.

	clean		attacked			feasible		
budgets (coll, en)	collision	energy	collision	energy	goal (att.)	clean	att.	UCB
(0.030, 2.5)	$0.024_{\pm 0.008}$	$2.25_{\pm 0.07}$	$0.023_{\pm 0.003}$	$2.25_{\pm 0.11}$	$0.944_{\pm 0.024}$	3/5	5/5	7/20
(0.030, 3.0)	$0.029_{\pm 0.009}$	$2.63_{\pm 0.11}$	$0.033^{\dagger}_{\pm 0.005}$	$2.63_{\pm 0.16}$	$0.975_{\pm 0.010}$	2/5	2/5	3/20
(0.030, 3.5)	$0.031^{\dagger}_{\pm 0.009}$	$3.19_{\pm 0.27}$	$0.023_{\pm 0.005}$	$3.10_{\pm 0.12}$	$0.992_{\pm 0.008}$	2/5	5/5	1/20
(0.035, 2.5)	$0.023_{\pm 0.006}$	$2.26_{\pm 0.19}$	$0.025_{\pm 0.006}$	$2.23_{\pm 0.10}$	$0.938_{\pm 0.038}$	5/5	5/5	15/20
(0.035, 3.0)	$0.023_{\pm 0.005}$	$2.69_{\pm 0.16}$	$0.027_{\pm 0.011}$	$2.70_{\pm 0.14}$	$0.991_{\pm 0.010}$	5/5	4/5	14/20
(0.035, 3.5)	$0.025_{\pm 0.005}$	$3.02_{\pm 0.10}$	$0.033_{\pm 0.009}$	$3.08_{\pm 0.09}$	$0.994_{\pm 0.002}$	5/5	3/5	4/20
(0.039, 2.5)	$0.026_{\pm 0.007}$	$2.32_{\pm 0.20}$	$0.027_{\pm 0.007}$	$2.32_{\pm 0.18}$	$0.943_{\pm 0.037}$	4/5	5/5	13/20
(0.039, 3.0)	$0.030_{\pm 0.009}$	$2.72_{\pm 0.09}$	$0.030_{\pm 0.008}$	$2.77_{\pm 0.05}$	$0.987_{\pm 0.006}$	4/5	4/5	12/20
(0.039, 3.5)	$0.035_{\pm 0.005}$	$3.16_{\pm 0.11}$	$0.035_{\pm 0.008}$	$3.27_{\pm 0.10}$	$0.995_{\pm 0.004}$	3/5	4/5	11/20

Paired values per seed over shared seeds (500-episode tails, both variants) compare the two variants cost by cost. Across all nine cells the difference in collision cost per seed between the variants stays within ± 0.004 and its 95% interval straddles zero everywhere (worst interval edge 0.011). The energy difference stays within ± 0.12 with every interval straddling zero (worst edge 0.29, under 12% of the tightest budget). The mean-criterion feasibility verdicts agree in four or five of five seeds in eight cells and in two of five at $(0.030, 3.5)$, with the disagreements split between the variants. One interval excludes zero: a $+0.006$ edge in goal rate for RC-MAT at $(0.039, 3.0)$. The tabulated RC-MAT goal is its training-time attack value (within one s.d. of clean throughout); the constraint-only variant trains no adversary and is evaluated clean.

Table 20: RC-MATD3, cooperative navigation, full budget grid (N=4, five seeds per cell, mean$_{\pm\text{s.d.}}$, training-time attack). First row: the same five runs as Table 7 and the 6000-episode RC-MATD3 row of Table 8, in a separate evaluation pass.

coll. budget	energy budget	collision cost	energy cost	min-dim sat	goal rate	return
0.039	5.2	$0.024_{\pm 0.004}$	$3.80_{\pm 0.17}$	$0.72_{\pm 0.03}$	$0.29_{\pm 0.08}$	$2.45_{\pm 0.07}$
0.039	8.6	$0.041^{\dagger}_{\pm 0.010}$	$5.91_{\pm 0.47}$	$0.71_{\pm 0.03}$	$0.34_{\pm 0.09}$	$2.77_{\pm 0.09}$
0.039	13	$0.045^{\dagger}_{\pm 0.011}$	$8.35_{\pm 0.32}$	$0.77_{\pm 0.02}$	$0.46_{\pm 0.05}$	$3.00_{\pm 0.08}$
0.064	5.2	$0.025_{\pm 0.004}$	$3.39_{\pm 0.16}$	$0.77_{\pm 0.03}$	$0.29_{\pm 0.05}$	$2.35_{\pm 0.10}$
0.064	8.6	$0.035_{\pm 0.006}$	$5.41_{\pm 0.43}$	$0.78_{\pm 0.04}$	$0.29_{\pm 0.05}$	$2.66_{\pm 0.06}$
0.064	13	$0.046_{\pm 0.013}$	$8.06_{\pm 0.47}$	$0.77_{\pm 0.03}$	$0.51_{\pm 0.06}$	$2.97_{\pm 0.11}$
0.097	5.2	$0.028_{\pm 0.004}$	$3.64_{\pm 0.15}$	$0.74_{\pm 0.02}$	$0.32_{\pm 0.12}$	$2.45_{\pm 0.07}$
0.097	8.6	$0.037_{\pm 0.007}$	$5.90_{\pm 0.33}$	$0.73_{\pm 0.04}$	$0.22_{\pm 0.10}$	$2.72_{\pm 0.13}$
0.097	13	$0.058_{\pm 0.013}$	$8.46_{\pm 0.43}$	$0.79_{\pm 0.04}$	$0.49_{\pm 0.08}$	$3.03_{\pm 0.15}$

D.10 Training dynamics

Figure 1 in the body shows goal rate, energy cost, and energy multiplier of the primary backbone at the tight cell for the anchored variant, the unanchored variant under the default warm-up, and the unanchored variant with no warm-up (five seeds each). Figure 8 adds the collision cost and its multiplier. Dual release is at episode 1000 (dotted line). Without the anchor, the release drives the energy multiplier to about 7 within 200 episodes. The energy cost undershoots its budget, and the goal rate falls from about 0.9 to about 0.1 and does not recover. With no warm-up the same variant trains through to about 0.8. The anchored variant reaches 0.96 with both multipliers staying below 0.2 and both costs riding just under budget.

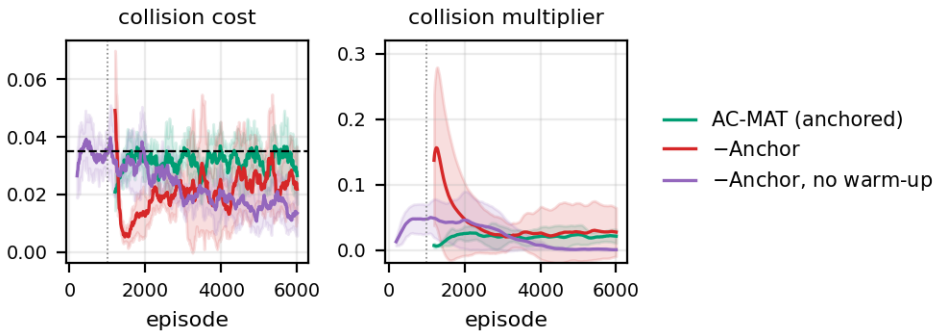

Figure 8: Collision cost and collision multiplier at the tight cell for the variants of Figure 1: AC-MAT (green), −Anchor (red), −Anchor, no warm-up (purple); five seeds, mean±s.d.; values from the dual, smoothed over 200 episodes; dashed: collision budget.

D.11 Per-seed dispersion of the contested cells

The five individual seeds behind the RC-MATD3 and AC-MATD3 cells of attacked goal rate of Table 7 are as follows. Under the training-time attack at the two tight cells, the full-method seeds lie in a narrow band (0.17–0.21 in the problem with a single constraint). The AC-MATD3 seeds split into survivors and near-zero collapses (three of five at 0.01 there; 0.005–0.415 at the cell with two constraints). Large attacked-column deviations measure this split rather than a uniform degradation. At the mid budget both variants cluster under attack, and in the clean condition the two variants disperse comparably. Paired over fifteen seeds, the comparison of training against the adversary in Section 5.3 is, in full (mean gap, 95% interval): clean $+0.002$ $[-0.016, 0.021]$ tight and $+0.006$ $[-0.004, 0.016]$ mid; best response $+0.019$ $[-0.017, 0.055]$ tight and $+0.025$ $[-0.006, 0.055]$ mid. The 95% intervals allow gains from training against the adversary of up to 0.055 in attacked goal rate and 0.021 in clean goal rate. The training-size frontier of Appendix D.4 (Table 15) is the positive counterpart: a stronger training attacker does gain about 0.10 of retention at a cost in constraint slack.

D.12 HETEROGENEOUS PER-AGENT BUDGETS

Each agent has its own dual, and budgets may differ across agents. This experiment uses the off-policy backbone. We start from the central grid cell (collision 0.064, energy 8.6, five seeds). We halve one agent's energy budget to 4.3 and leave its three teammates unchanged. The targeted agent cuts its energy expenditure from 5.16 to 3.20 while the teammates' expenditure rises slightly (5.50 to 5.74). Goal rate moves little (0.288 uniform, 0.273 heterogeneous). All four agents are feasible in mean in four of five seeds, against five of five under the uniform budget (the remaining seed has two agents over the collision budget). A single mechanism thus supports agent-specific budget assignments. Each agent responds to its own constraint and the collective reallocates effort around it (Figure 9). On the transformer backbone at binding budgets the duals separate the agents' spend as well, but per-agent feasibility is weaker than at the uniform cells. With agents 1–2 at the tight pair $(0.035, 2.5)$ and agents 3–4 at $(0.035, 3.5)$, RC-MAT's tight-budget agents spend $0.97\times$ and $1.04\times$ of their energy budget against $0.77\times$ and $0.78\times$ for the loose-budget agents, at a clean goal rate of 0.980 (five seeds, 500-episode tails). With one agent at $(0.030, 2.5)$ and three at $(0.039, 3.5)$ the singled-out agent is at $1.13\times$ its collision budget and $0.98\times$ its energy budget against 0.83–$0.91\times$ for its teammates, at 0.985. Each agent responds to its own budget, but the tighter-budget agents ride at or just over them. All four agents are thus within budget in no seed of the first assignment and one of five of the second, against five of five at the uniform tight cell. The competing variants on the same two assignments (five seeds each, 500-episode tails) show the same trade as the uniform cells. On the first assignment the PID variant reaches $0.761_{\pm 0.053}$ clean with one of five seeds feasible for every agent and the unanchored variant trained with no warm-up $0.730_{\pm 0.108}$ with four of five, against 0.980 for RC-MAT with none. On the second they reach $0.838_{\pm 0.077}$ with one of five and $0.797_{\pm 0.083}$ with two of five, against 0.985 for RC-MAT with one. AC-MAT at the same budgets reaches $0.953_{\pm 0.027}$ with three of five on the first assignment and $0.987_{\pm 0.010}$ with none of five on the second, where the singled-out agent runs at $1.16\times$ its collision budget in seed mean. A no-training control separates what the per-agent multipliers contribute from what the budgets impose. The fifteen AC-MAT policies trained at the uniform tight cell are judged agent by agent against the heterogeneous budgets on their 500-episode tails. They are feasible for every agent in 14 of 15 on the first assignment, which loosens two agents' energy budgets, and in 10 of 15 on the second, which tightens one agent's collision budget (worst agent $1.18\times$), at a clean goal rate of 0.947. Training on the heterogeneous budgets thus raises the goal rate (0.953–0.987 against 0.947) and does not by itself bring the singled-out agent within budget. The training margin of Appendix D.9 brings every agent within budget in five of five seeds on the first assignment and four of five on the second ($0.80\times$/$0.83\times$). The unanchored no-warm-up variant is the most reliably feasible on heterogeneous budgets, at a goal rate 0.19–0.25 lower. Appendix D.9 gives the effect of the training margin on these assignments.

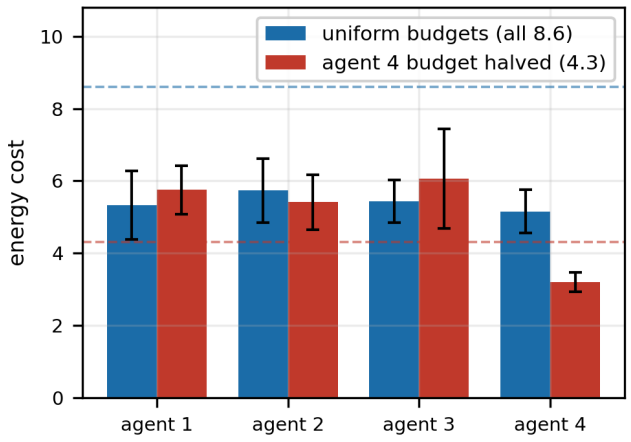

Figure 9: Off-policy per-agent energy under uniform (blue) and heterogeneous (red) budgets (five seeds, mean±s.d.; dashed lines mark the two budget levels). The targeted agent shows the largest reduction.

D.13 External baseline training dynamics

Port fidelity. Both constraint-only baselines run the baseline authors' maintained official implementation with its algorithm files unmodified. Our additions are an environment adapter and the authors' own benchmark-tuned hyperparameter block. We override the problem-matching parameters alone: episode length, discount, network width, and the cost limit. We convert the cost limit between discounted and undiscounted accounting by ratios measured on evaluation rollouts, 1.249 on the problem with two constraints and 1.287 on the native task. By the baselines' own undiscounted internal accounting (tail-window training episode cost), the collapsed MAPPO-Lagrangian policies are two orders of magnitude under their native limits. They are 0.01–0.02 against limits of 7.98–15.96 across the three levels. MACPO tracks its native limit at ratios 0.98–1.05.

Portability control: the anchor inside the official learner. We also test portability in the direction opposite to the ports above, by moving the anchor into an official learner. The anchor of (4) is added to the official MAPPO-Lagrangian actor loss as the same action-mean penalty, with the reference an exponential moving average of the actor's parameters at the same rate. An adapter-side subclass leaves the algorithm files unmodified. The port is retrained at our interaction budget with anchor weight $\zeta \in \{10, 2\}$ (five seeds per cell). Neither weight rescues the port. At the tight budget the clean goal rate moves from $0.03_{\pm 0.01}$ to $0.05_{\pm 0.01}$ (ζ=10) and $0.06_{\pm 0.01}$ (ζ=2); at the mid budget from 0.04 to 0.04 and 0.06. Energy spent falls further below the limit (0.53 and 0.71 against 0.88 at tight). A transferred attack leaves every variant at 0.01. Adding the anchor to official MAPPO-Lagrangian does not reproduce the trainability gain of Table 3. That learner also differs in dual rule, stochastic rollout evaluation, and undiscounted cost accounting. A factorial inside the same official learner separates them (five seeds per variant at the tight and mid cells with a single constraint). Feeding its dual a deterministic discounted cost measured against the discounted budget, with or without the anchor, leaves the port at clean goal rates of $0.038_{\pm 0.020}$/$0.034_{\pm 0.008}$ and $0.043_{\pm 0.010}$/$0.046_{\pm 0.003}$ (tight/mid). Replacing its multiplier rule by the per-agent subgradient rule of Section 4.3 on top of both raises them merely to $0.115_{\pm 0.012}$/$0.118_{\pm 0.016}$. Every variant spends well under its energy budget. No tested combination of the anchor, the deterministic dual signal, and the multiplier rule transfers the trainability effect into that learner at its released multiplier initialization and rate.

Home-benchmark control (official code). To complement the diagnostic ports with the baseline's own benchmark, we run the official MACPO implementation on its Safe MAMuJoCo home task (HalfCheetah 2×3, five seeds, the released training-script hyperparameters verbatim). By its own train-rollout reporting over the final 8×10^5 steps it reaches reward $2646_{\pm 111}$ at cost $17.9_{\pm 6.2}$ against its bound of 10, every seed over the bound. Our construction trained on HalfCheetah 2×3 of the Safety-Gymnasium velocity suite, the same robot and partition with a velocity-threshold cost in place of the moving-obstacle proximity cost, under its own discounted cost convention, with budget 1.94 against an unconstrained natural spend of 76.6, gives reward 2414 with four of five seeds within budget on the final deterministic evaluation for the anchored constraint-only variant, 2380 with one of five for RC-MAT (mean cost 2.70, $1.4\times$ the budget), and 2354 with four of five for the unanchored constrained variant. Cost definitions and units differ between the two tasks, so this is a co-reported record rather than a ranking. The anchor trainability effect does not appear on this task (Table 17, Appendix C.6).

Figure 10 shows the collapse noted in Appendix C.4. MAPPO-Lagrangian's return peaks near one million steps as its cost approaches the limit. Its dual keeps pressing and drives the cost far below the limit toward zero. Return decays with it until the policy is trivially feasible and static. Policy entropy collapses alongside (from 0.8 at initialization to -7 at ten million steps). Once frozen, the policy collects no moving on-policy data and cannot re-explore even as the multiplier relaxes. The interaction is harshest when the constrained quantity is control energy, where the cheapest feasible policy is to stand still. MACPO recovers late in the opposite way. By ten million steps its training return climbs back to positive values with the training cost at the limit, but the stochastic exploration noise drives that movement. The deterministic policy spends roughly a fifth of the behavior policy's energy and completes almost no goals. That is what the deterministic evaluation protocol reports. Table 21 re-evaluates every method of Table 6 under its own stochastic execution as well, with the energy cost in both the discounted accounting the budgets use and the undiscounted accounting the ports regulate (200 episodes per run, five seeds). Stochastic execution changes the goal rate of no port (MACPO 0.02–0.12, PID-Lag 0.28–0.93, MAPPO-Lagrangian 0 throughout) while raising their energy two- to fourfold. The port order in Table 6 is thus independent of execution convention.

Stochastic execution also shows what our own budgets cover. The sampled RC-MAT policy completes goals as well as its mean policy (0.995–0.997) but spends 15–16 in discounted energy, three times the 4.6–4.8 of the mean policy and above every budget, because the dual uses deterministic rollouts. Our budgets are measured on, and hold for, the deterministic deployment. Making the dual read the sampled policy instead does not turn that into a certificate for stochastic execution (its evaluation rollouts sample actions; rule, box, step, and cadence unchanged; AC-MAT recipe, tight cell, five seeds). By the noise alone, the sampled policy's energy exceeds the budget at the start of training. The multipliers rise until the policy stops moving, and training ends at a clean goal rate of $0.022_{\pm 0.020}$ with deterministic and sampled energy at 0.21 and 0.36. The deterministic evaluation in the default recipe is thus deliberate: under the tested exploration and optimization settings the mean policy alone meets the budgets, and the anchor benefit is measured for that policy.

Table 21: Stochastic execution leaves the goal rate of every port unchanged and raises energy for all methods; our budgets hold only for the deterministic policy. Native single-cost task, 200 episodes per run, five seeds, mean$_{\pm \text{s.d.}}$; RC-MATD3 has no stochastic mode.

		deterministic execution			stochastic execution		
Method	budget	goal rate	energy (γ=0.99)	energy (undisc.)	goal rate	energy (γ=0.99)	energy (undisc.)
RC-MAT	6.2	$0.992_{\pm 0.006}$	$4.59_{\pm 0.14}$	$6.33_{\pm 0.20}$	$0.997_{\pm 0.003}$	$15.47_{\pm 0.37}$	$25.03_{\pm 0.82}$
RC-MAT	8.7	$0.993_{\pm 0.007}$	$4.75_{\pm 0.27}$	$6.51_{\pm 0.33}$	$0.995_{\pm 0.004}$	$16.19_{\pm 0.14}$	$26.10_{\pm 0.31}$
RC-MAT	12.4	$0.992_{\pm 0.006}$	$4.68_{\pm 0.13}$	$6.44_{\pm 0.19}$	$0.995_{\pm 0.004}$	$16.14_{\pm 0.27}$	$26.11_{\pm 0.56}$
−Anchor	6.2	$0.760_{\pm 0.301}$	$4.63_{\pm 0.23}$	$7.18_{\pm 0.50}$	$0.764_{\pm 0.288}$	$9.13_{\pm 1.19}$	$16.65_{\pm 4.25}$
−Anchor	8.7	$0.949_{\pm 0.032}$	$6.46_{\pm 0.47}$	$9.13_{\pm 0.86}$	$0.963_{\pm 0.033}$	$12.63_{\pm 1.28}$	$20.22_{\pm 2.85}$
−Anchor	12.4	$0.971_{\pm 0.022}$	$9.39_{\pm 0.62}$	$13.34_{\pm 1.83}$	$0.980_{\pm 0.022}$	$18.75_{\pm 0.92}$	$29.54_{\pm 2.47}$
RC-MATD3	6.2	$0.089_{\pm 0.029}$	$3.66_{\pm 0.15}$	$5.02_{\pm 0.21}$	n/a	n/a	n/a
RC-MATD3	8.7	$0.208_{\pm 0.050}$	$5.96_{\pm 0.41}$	$7.69_{\pm 0.43}$	n/a	n/a	n/a
RC-MATD3	12.4	$0.302_{\pm 0.032}$	$6.69_{\pm 0.51}$	$8.50_{\pm 0.58}$	n/a	n/a	n/a
MAPPO-Lag (8.3×)	6.2	$0.000_{\pm 0.000}$	$0.00_{\pm 0.00}$	$0.01_{\pm 0.00}$	$0.000_{\pm 0.000}$	$0.01_{\pm 0.00}$	$0.01_{\pm 0.00}$
MAPPO-Lag (8.3×)	8.7	$0.000_{\pm 0.000}$	$0.00_{\pm 0.00}$	$0.01_{\pm 0.00}$	$0.000_{\pm 0.000}$	$0.01_{\pm 0.00}$	$0.01_{\pm 0.00}$
MAPPO-Lag (8.3×)	12.4	$0.000_{\pm 0.000}$	$0.00_{\pm 0.00}$	$0.01_{\pm 0.00}$	$0.000_{\pm 0.000}$	$0.01_{\pm 0.00}$	$0.01_{\pm 0.00}$
MACPO (8.3×)	6.2	$0.023_{\pm 0.011}$	$1.08_{\pm 0.07}$	$2.16_{\pm 0.24}$	$0.033_{\pm 0.021}$	$3.67_{\pm 0.06}$	$8.16_{\pm 0.30}$
MACPO (8.3×)	8.7	$0.065_{\pm 0.048}$	$1.94_{\pm 0.37}$	$3.43_{\pm 0.22}$	$0.063_{\pm 0.051}$	$5.30_{\pm 0.29}$	$11.19_{\pm 0.07}$
MACPO (8.3×)	12.4	$0.118_{\pm 0.075}$	$3.46_{\pm 0.49}$	$5.21_{\pm 0.27}$	$0.115_{\pm 0.079}$	$8.06_{\pm 0.44}$	$15.91_{\pm 0.16}$
PID-Lag (8.3×)	6.2	$0.281_{\pm 0.108}$	$3.87_{\pm 0.17}$	$5.37_{\pm 0.20}$	$0.276_{\pm 0.123}$	$4.80_{\pm 0.37}$	$7.45_{\pm 0.55}$
PID-Lag (8.3×)	8.7	$0.839_{\pm 0.078}$	$5.79_{\pm 0.30}$	$7.69_{\pm 0.32}$	$0.829_{\pm 0.078}$	$7.55_{\pm 0.41}$	$11.31_{\pm 0.51}$
PID-Lag (8.3×)	12.4	$0.915_{\pm 0.030}$	$8.40_{\pm 0.62}$	$10.40_{\pm 0.73}$	$0.931_{\pm 0.029}$	$11.09_{\pm 0.35}$	$15.40_{\pm 0.19}$

The unconstrained control run cited in Section C.4 (five seeds) shows the same multiplier dynamics in reverse. Initialized at 0.78 with an unreachable cost limit, the multiplier decreases linearly to its zero floor at 32% of training. The remaining two thirds are vanilla MAPPO learning that reaches a goal rate of $0.82_{\pm 0.04}$ while spending 13.6 energy. That leaves the dual as the most consistent explanation of the constrained runs' collapse.

Three remedies applied to the port (validation-feasibility early stopping with snapshots every ∼400k steps, a cost-limit curriculum annealed from three times the target, and a pinned entropy coefficient of 0.05; five seeds each) lift the deterministic goal rate to at most $0.109_{\pm 0.019}$. None reaches a working point.

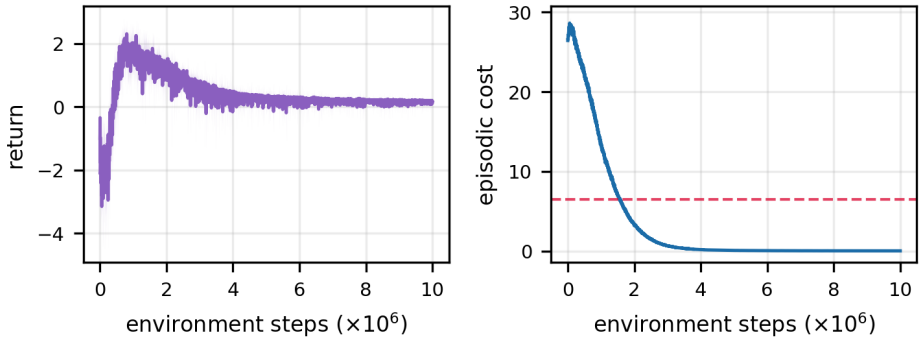

Figure 10: MAPPO-Lagrangian at its native interaction budget and the tight energy budget (five seeds, mean±s.d.). Left: training return. Right: episodic training cost against its limit (dashed).