Rethinking Ratio-Based Trust Regions for Policy Optimization in Multi-Agent Reinforcement Learning

Chulabhaya Wijesundara
Northeastern University; STR

Andrea Baisero
University of California, Irvine

Zhongheng Li
STR

Gregory Castañón
STR

Alan Carlin
STR

Christopher Amato
Northeastern University

Abstract

Centralized training with decentralized execution (CTDE) is a standard framework for cooperative multi-agent policy-gradient reinforcement learning, allowing agents to learn from joint information while acting from local observations. Ratio-based trust-region methods such as Multi-Agent Proximal Policy Optimization (MAPPO) and Multi-Agent Simple Policy Optimization (MASPO) update decentralized actors using per-agent probability ratios weighted by joint advantage estimates. Teammate non-stationarity increases the variance of these advantages, which in turn increases the variance in the local ratio updates. This exposes two method-specific failure modes: MAPPO's additive clipping removes gradients for outlier samples and weakens recovery from policy drift, while MASPO's soft quadratic penalty can allow probability collapse. We introduce Multi-Agent Ratio Symmetry (MARS), a novel policy optimization objective that replaces these additive ratio-based trust-region mechanisms with a multiplicatively symmetric geometric barrier. MARS preserves corrective gradients while assigning unbounded cost as probability ratios approach zero. Across 47 tasks spanning eight multi-agent environments, including novel JAX benchmarks PaxMen and AeroJAX, MARS matches or exceeds MAPPO and MASPO in aggregate environment-level performance. Ablations show that these gains arise from the geometry of the symmetric barrier rather than from flexible trust-region boundaries alone.

1 Introduction

Deep Multi-Agent Reinforcement Learning (MARL) has enabled agents to learn coordinated behavior in domains ranging from games (Ellis et al., 2023) to swarm robotics (Liu et al., 2024) and transportation infrastructure (Saifullah et al., 2024). A common framework for cooperative MARL is centralized training with decentralized execution (CTDE), where actors are trained using centralized information but execute from local observations. Within CTDE, ratio-based trust-region policy-gradient methods such as Multi-Agent Proximal Policy Optimization (MAPPO) (Yu et al., 2022) and Multi-Agent Simple Policy Optimization (MASPO) (Selmonaj et al., 2025) are widely used because they are simple and empirically robust.

CTDE creates a distinct challenge for ratio-based policy updates. In MAPPO and MASPO, each decentralized actor is updated using a per-agent probability ratio, while the update direction is weighted by an advantage estimate from a centralized critic. Because this critic evaluates outcomes under simultaneously changing teammate policies, its advantage estimates can have high variance (Lyu et al., 2023; Foerster et al., 2018). These high-variance joint advantages are then applied directly to local ratio updates, producing large increases or decreases in individual action probabilities.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

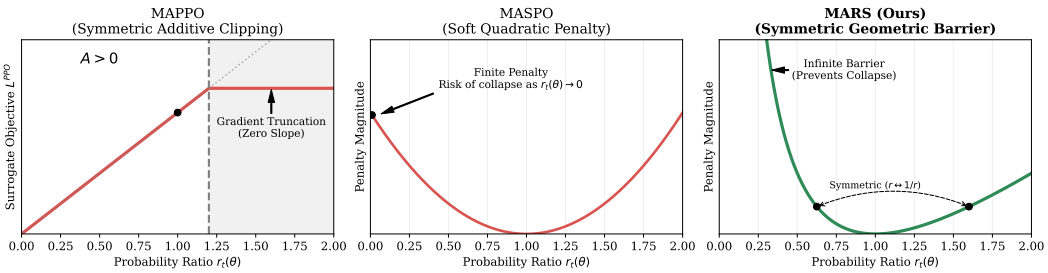


Figure 1: *Comparison of ratio-based trust region mechanisms.* MAPPO's additive clipping creates zero-gradient regions after ratio outliers cross the clipping bounds. MASPO restores gradient flow with a soft quadratic penalty, but its cost remains finite as $r \to 0$. MARS replaces additive ratio control with a multiplicatively symmetric geometric barrier that preserves gradients on $r > 0$ and diverges as $r \to 0$, making probability extinction non-optimal for the per-sample ratio surrogate.

Existing additive trust-region mechanisms are brittle in this regime. MAPPO clips per-agent ratios to a fixed interval around the old policy. When a sample crosses the clipping boundary in the improving direction, its policy-gradient contribution is zeroed. In CTDE, this can remove corrective information from precisely the samples most affected by high-variance joint advantages, allowing policy drift to accumulate across minibatch updates.

MASPO addresses this truncation by replacing hard clipping with a soft quadratic penalty, so samples outside the nominal trust region continue to provide gradients. However, because the penalty is defined by additive distance from the old policy, it remains finite as the probability ratio approaches zero, corresponding to probability extinction for the sampled action. A large negative joint advantage can therefore dominate the penalty and continue decreasing the probability of a sampled action, even when that negative signal reflects teammate exploration or a transient coordination failure. This can reduce exploration over actions that remain useful for coordination (Agarwal et al., 2021a).

The shared limitation is that MAPPO and MASPO measure ratio changes additively around one, even though probability ratios represent relative changes in action probability. Additive bounds treat equal numerical changes above and below one as symmetric, but expansion and contraction are naturally symmetric only as multiplicative inverses. In MAPPO, this makes the clipping thresholds geometrically asymmetric: contraction can proceed farther than expansion before truncation. In MASPO, the same additive geometry appears as a quadratic penalty whose cost remains finite as the probability ratio approaches zero. Under high-variance joint advantages, noisy negative signals can suppress useful actions more aggressively than intended, while the objective provides either no corrective gradient after clipping or only finite resistance near probability collapse.

We introduce **Multi-Agent Ratio Symmetry (MARS)**, a CTDE policy optimization objective that replaces additive ratio control with a multiplicatively symmetric geometric barrier. MARS preserves smooth gradients over the valid ratio domain and assigns unbounded cost as probability ratios approach zero. It therefore retains gradient information from outlier samples while allowing negative-advantage actions to be down-weighted without making ratio extinction optimal.

Contributions. We contribute: (i) MARS, a CTDE-focused full-gradient trust-region objective for cooperative MARL based on a multiplicatively symmetric geometric barrier; (ii) theoretical results showing that MARS preserves corrective gradients and imposes an infinite barrier as probability ratios approach zero; (iii) two JAX-native benchmarks, PaxMen for coordinated exploration and AeroJAX for continuous 3D aerial combat; and (iv) an eight-environment, 47-task evaluation showing that MARS matches or exceeds MAPPO and MASPO in aggregate environment-level performance, with ablations attributing the gains to the curvature of the symmetric barrier rather than flexible trust-region boundaries alone.

2 Background

2.1 Problem Formulation

We model cooperative MARL as a decentralized partially observable Markov decision process (Dec-POMDP) (Oliehoek & Amato, 2016) $\mathcal{M} = \langle \mathcal{I}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{I}}, \{\mathcal{O}_i\}_{i\in\mathcal{I}}, T, O, R, b_0, \gamma \rangle$. Here, $\mathcal{I} = \{1, \ldots, n\}$ is the set of agents, $\mathcal{S}$ is the state space, and each agent i has action and observation spaces $\mathcal{A}_i$ and $\mathcal{O}_i$. At timestep t, agents take a joint action $\mathbf{a}_t = (a_{1,t}, \ldots, a_{n,t})$, the environment transitions according to $s_{t+1} \sim T(s_t, \mathbf{a}_t)$, and each agent receives a local observation $o_{i,t+1}$. All agents share the reward $r_t = R(s_t, \mathbf{a}_t)$.

The objective is to learn decentralized policies $\pi_i(a_{i,t} \mid h_{i,t})$, where $h_{i,t}$ is agent i's local action-observation history, that maximize $J = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t]$. In centralized training with decentralized execution (CTDE), actors condition only on local histories at execution time, while training may use centralized information to estimate value functions or advantages.

2.2 Multi-Agent Proximal Policy Optimization

Multi-Agent Proximal Policy Optimization (MAPPO) (Yu et al., 2022) adapts the clipped ratio objective of PPO (Schulman et al., 2017) to CTDE by combining decentralized actors with a centralized critic. For each agent i, MAPPO forms the per-agent probability ratio

$$r_{i,t}(\theta) = \frac{\pi_\theta(a_{i,t} \mid h_{i,t})}{\pi_{\theta_{\text{old}}}(a_{i,t} \mid h_{i,t})}, \tag{1}$$

and weights the policy update by a joint advantage estimate $\widehat{\mathbf{A}}_t$. The per-agent MAPPO objective is

$$\mathcal{L}_{i,t}^{\text{MAPPO}}(\theta) = \min\Big(r_{i,t}(\theta)\widehat{\mathbf{A}}_t,\ \text{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon)\widehat{\mathbf{A}}_t\Big), \tag{2}$$

where ϵ controls the clipping range.

2.3 Multi-Agent Simple Policy Optimization

Multi-Agent Simple Policy Optimization (MASPO) (Selmonaj et al., 2025) extends the soft-penalty objective of Simple Policy Optimization (SPO) (Xie et al., 2025) to CTDE MARL. Using the same per-agent ratio $r_{i,t}(\theta)$ and joint advantage estimate $\widehat{\mathbf{A}}_t$, MASPO replaces clipping with an additive quadratic penalty:

$$\mathcal{L}_{i,t}^{\text{MASPO}}(\theta) = r_{i,t}(\theta)\widehat{\mathbf{A}}_t - \frac{|\widehat{\mathbf{A}}_t|}{2\epsilon}\left(r_{i,t}(\theta) - 1\right)^2. \tag{3}$$

The quadratic term softly penalizes deviations of the per-agent ratio from one while retaining gradient flow outside the nominal trust region.

3 Multi-Agent Ratio Symmetry

3.1 Motivation: Joint Advantage Variance Exposes Ratio-Control Failures

The preceding section defined MAPPO and MASPO as per-agent ratio objectives weighted by a joint advantage estimate. This combination is central to CTDE policy optimization: each decentralized actor is updated through a local probability ratio, while the learning signal reflects the joint behavior of all agents. As teammates adapt, the same local action can receive different advantage estimates across updates because the critic evaluates outcomes under a non-stationary joint policy (Lyu et al., 2023; Foerster et al., 2018). Thus, high-variance joint advantage estimates are applied directly to per-agent probability ratios.

3.1.1 Gradient Truncation in MAPPO

In MAPPO, the additive trust region is enforced through hard clipping bounds $[1-\epsilon, 1+\epsilon]$, as shown in Figure 1. The clipped objective removes gradients when a sample crosses the clipping boundary in

the improving direction. For the per-agent, single-step clipped objective, the ratio-gradient is

$$\frac{\partial \mathcal{L}_{i,t}^{\mathrm{MAPPO}}}{\partial r_{i,t}} = \begin{cases} \widehat{\mathbf{A}}_t, & \text{if } |r_{i,t}(\theta) - 1| < \epsilon \text{ or } \operatorname{sign}(\widehat{\mathbf{A}}_t)(r_{i,t}(\theta) - 1) < 0, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

See Appendix C.1 for the derivation.

In CTDE, this truncation is consequential because ratio outliers are often produced by high-variance joint advantage estimates. Once a transition is clipped, its objective term no longer provides a gradient that pulls the ratio back toward the trust region. Since policy parameters are shared across transitions in a minibatch, updates from other active samples can still move the clipped ratio further from the old policy. Without a smooth restoring signal for clipped outliers to counteract this drift, MAPPO allows the policy to drift arbitrarily far from the constraint, undermining the monotonic improvement guarantee (Wang et al., 2020; Xie et al., 2025; Engstrom et al., 2019).

3.1.2 Finite Penalty in MASPO

MASPO removes MAPPO's hard clipping and restores gradient flow by replacing the clipped objective with a soft quadratic penalty,

$$\psi_{\mathrm{MASPO}}(r) = (r-1)^2. \tag{5}$$

This penalty is defined using additive distance from one. As a result, it assigns finite cost to probability extinction:

$$\lim_{r \to 0^+} (r-1)^2 = 1. \tag{6}$$

Thus, MASPO maintains gradients outside the nominal trust region, but its penalty remains bounded as the probability ratio approaches zero.

This bounded cost matters under joint advantage noise. A large negative joint advantage can dominate the finite penalty and continue decreasing the probability of the sampled action. Because the joint advantage may reflect teammate exploration or temporary coordination failure, this update can prematurely suppress actions that remain useful for future coordination, contributing to near-deterministic policies (Agarwal et al., 2021a).

3.1.3 Additive Ratio Geometry as a Common Structure

MAPPO and MASPO fail through different mechanisms, but both regulate probability-ratio change using additive distance from $r = 1$. In MAPPO, the additive interval $[1-\epsilon, 1+\epsilon]$ determines where clipping occurs. This interval is symmetric as a numerical distance from one, but it is not symmetric as a relative change in probability. For example, with $\epsilon = 0.2$, the lower bound 0.8 corresponds to a 20% contraction, whose inverse expansion is $1/0.8 = 1.25$, while the upper bound permits expansion only to 1.2. Thus, before clipping occurs, MAPPO permits a stronger multiplicative contraction than expansion. Once the lower boundary is crossed for a negative-advantage sample, the clipped objective removes the gradient, so the same additive boundary both permits geometrically uneven contraction and then removes the corrective signal after the sample becomes an outlier.

MASPO uses the same additive parameterization in soft form: its quadratic penalty depends on distance from $r = 1$. This restores gradient flow outside the nominal trust region, but the penalty remains finite as $r \to 0$. As a result, a sufficiently large negative joint advantage can continue pushing the sampled-action ratio toward zero. Thus, MAPPO's truncation and MASPO's collapse are distinct failures, but both arise from controlling relative probability changes through additive distance around $r = 1$.

The geometric issue is clearer in log-ratio space. Probability ratios encode relative change: expansion by r is naturally inverted by contraction by $1/r$. These inverse changes are symmetric in $u = \log r$, since r and $1/r$ become u and $-u$. Additive neighborhoods around one do not preserve this symmetry, while the collapse limit $r \to 0$ corresponds to $u \to -\infty$. This motivates replacing additive ratio control with a regularizer defined by multiplicative symmetry rather than numerical distance from one.

3.2 The Symmetric Ratio Objective

The preceding analysis suggests three criteria for a CTDE-compatible ratio regularizer:

1. **Smooth corrective gradients:** The objective should remain differentiable across the valid ratio domain so that outlier samples continue to provide gradient information rather than being removed by clipping.
2. **Unbounded extinction barrier:** The penalty should diverge as $r \to 0^+$, ensuring that no finite negative joint advantage can make probability extinction optimal.
3. **Multiplicative symmetry:** The penalty should satisfy $\psi(r) = \psi(1/r)$, treating expansion and contraction as inverse changes in probability rather than as unequal additive deviations from one.

3.2.1 The Multi-Agent Ratio Symmetry Penalty

Many ratio penalties could satisfy these criteria. We choose a geometric symmetrization of MASPO's additive quadratic penalty because it aligns with the multiplicative geometry of probability ratios while keeping the construction directly comparable to MASPO. In particular, the geometric mean treats a ratio and its inverse as paired deviations, preserves local quadratic behavior near $r = 1$, and changes the penalty primarily by replacing additive distance with multiplicative symmetry.

Concretely, we construct the Multi-Agent Ratio Symmetry (MARS) penalty by symmetrizing the MASPO quadratic penalty under the transformation $r \mapsto 1/r$. Applying geometric symmetrization to $(r-1)^2$ yields

$$\psi_{\mathrm{MARS}}(r) = r + \frac{1}{r} - 2. \tag{7}$$

See Appendix C.2 for the derivation. This penalty is zero at $r = 1$, smooth for all $r > 0$, and symmetric under inversion by design. The linear term penalizes unbounded expansion as $r \to \infty$, while the reciprocal term creates an unbounded barrier as $r \to 0^+$.

The derivative of the penalty with respect to the ratio is

$$\frac{d}{dr}\psi_{\mathrm{MARS}}(r) = 1 - \frac{1}{r^2}. \tag{8}$$

As $r \to 0^+$, the reciprocal-gradient term dominates. Therefore, when a large negative joint advantage pushes a sampled-action ratio downward, the MARS penalty contributes an increasingly strong opposing term in ratio space.

Now consider the per-agent, one-step ratio surrogate

$$\mathcal{L}_{i,t}^{\mathrm{MARS}}(\theta) = r_{i,t}(\theta)\widehat{\mathbf{A}}_t - \alpha_t \psi_{\mathrm{MARS}}\big(r_{i,t}(\theta)\big), \tag{9}$$

where $\alpha_t > 0$ controls the penalty strength. The ratio-gradient is

$$\frac{\partial \mathcal{L}_{i,t}^{\mathrm{MARS}}}{\partial r_{i,t}} = \widehat{\mathbf{A}}_t - \alpha_t \left(1 - \frac{1}{r_{i,t}(\theta)^2}\right). \tag{10}$$

Equation 10 shows the key barrier mechanism. For negative joint advantages, the linear surrogate can favor reducing the sampled-action ratio, but as $r \to 0^+$ the reciprocal term in the MARS penalty dominates any finite advantage. Thus, MARS can reduce ratios for negative-advantage samples while preventing ratio extinction from becoming optimal. Proposition 3.1 formalizes this limit behavior.

Proposition 3.1 (Barrier Against Probability Extinction)**.** *Let $r > 0$ and consider the per-agent, one-step MARS surrogate*

$$\mathcal{L}(r) = r\widehat{\mathbf{A}} - \alpha\psi_{\mathrm{MARS}}(r),$$

with $\alpha > 0$ and $\psi_{\mathrm{MARS}}(r) = r + \frac{1}{r} - 2$. Then

$$\lim_{r\to 0^+} \mathcal{L}(r) = -\infty, \qquad \lim_{r\to 0^+} \frac{\partial \mathcal{L}}{\partial r} = +\infty.$$

Consequently, probability extinction is never an optimizer of the per-sample ratio surrogate, and the ratio-gradient points away from zero in a neighborhood of $r = 0$. Full proof in Appendix C.3.

Having established the barrier behavior near $r = 0$, we now consider the second desired property: preserving gradient information for ratio outliers rather than creating clipped zero-gradient regions.

Proposition 3.2 (No Gradient Truncation)**.** *Let $r > 0$, $\alpha > 0$, and $\widehat{\mathbf{A}} \in \mathbb{R}$ be finite. For the MARS surrogate*

$$\mathcal{L}(r) = r\widehat{\mathbf{A}} - \alpha\psi_{\mathrm{MARS}}(r), \qquad \psi_{\mathrm{MARS}}(r) = r + \frac{1}{r} - 2,$$

the ratio-gradient

$$\frac{\partial \mathcal{L}}{\partial r} = \widehat{\mathbf{A}} - \alpha\left(1 - \frac{1}{r^2}\right)$$

has no zero-gradient interval on $(0, \infty)$. Full proof in Appendix C.4.

Proposition 3.2 shows that the MARS trust-region mechanism does not create MAPPO-style gradient truncation. Gradients may vanish at isolated stationary points, but there is no interval of ratios for which outlier samples stop contributing gradient information solely because they crossed a clipping boundary.

Together, Propositions 3.1 and 3.2 show that MARS assigns unbounded cost as probability ratios approach zero while avoiding gradient truncation.

3.2.2 Trust Region Alignment

The MARS penalty defines the shape of the ratio regularizer, while the scalar penalty weight α_t determines how strongly that penalty is applied. We choose α_t so that the advantage term and penalty term balance at the prescribed target ratio, making that ratio the stationary point of the one-step surrogate. This target plays the same role as the clipping boundary in MAPPO or the ϵ-aligned optimum in MASPO, but is specified directly in multiplicative ratio space rather than through additive bounds around one. This distinction matters because additive bounds of the form $(1 \pm \epsilon)$ are not multiplicatively symmetric: the lower and upper bounds are not reciprocal. We therefore define separate target ratios for expansion and contraction:

- **Expansion target ($B_{\mathrm{upper}} > 1$):** the intended boundary for up-weighting actions with $\widehat{\mathbf{A}}_t \geq 0$.
- **Contraction target ($B_{\mathrm{lower}} < 1$):** the intended boundary for down-weighting actions with $\widehat{\mathbf{A}}_t < 0$.

Target-ratio alignment for α_t. Let $r_{\mathrm{target}} = B_{\mathrm{upper}}$ for $\widehat{\mathbf{A}}_t \geq 0$ and $r_{\mathrm{target}} = B_{\mathrm{lower}}$ otherwise. We set α_t by requiring the ratio-gradient of the one-step surrogate to vanish at r_{target}. Since

$$\frac{\partial \mathcal{L}_{i,t}^{\mathrm{MARS}}}{\partial r_{i,t}} = \widehat{\mathbf{A}}_t - \alpha_t\left(1 - \frac{1}{r_{i,t}^2}\right), \qquad \left.\frac{\partial \mathcal{L}_{i,t}^{\mathrm{MARS}}}{\partial r_{i,t}}\right|_{r_{i,t}=r_{\mathrm{target}}} = 0,$$

we obtain

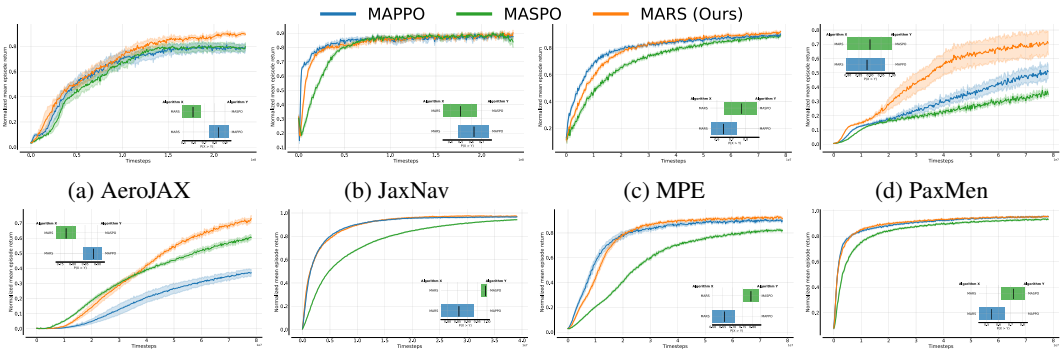
$$\alpha_t = \frac{\widehat{\mathbf{A}}_t}{1 - r_{\mathrm{target}}^{-2}} = \begin{cases} \frac{\widehat{\mathbf{A}}_t}{1 - B_{\mathrm{upper}}^{-2}} & \text{if } \widehat{\mathbf{A}}_t \geq 0, \\ \frac{\widehat{\mathbf{A}}_t}{1 - B_{\mathrm{lower}}^{-2}} & \text{if } \widehat{\mathbf{A}}_t < 0. \end{cases} \tag{11}$$

When $B_{\mathrm{upper}} > 1$ and $B_{\mathrm{lower}} < 1$, this yields $\alpha_t \geq 0$ in both regimes. The standard ϵ-aligned MASPO scaling is recovered as the additive special case $r_{\mathrm{target}} = 1 + \mathrm{sign}(\widehat{\mathbf{A}}_t)\epsilon$; see Appendix C.5.

Final objective. We now have the two ingredients needed to instantiate the MARS surrogate: a multiplicatively symmetric penalty shape and a penalty weight α_t calibrated to the desired expansion or contraction target. Combining the penalty in Equation (7) and the aligned weight in Equation (11) with the base ratio surrogate in Equation (9) gives the final per-agent MARS objective:

$$\mathcal{L}_{i,t}^{\mathrm{MARS}}(\theta) = r_{i,t}(\theta)\widehat{\mathbf{A}}_t - \alpha_t\left(r_{i,t}(\theta) + \frac{1}{r_{i,t}(\theta)} - 2\right). \tag{12}$$

Thus, MARS keeps the standard ratio-weighted policy-gradient term but replaces additive clipping or additive quadratic penalties with a calibrated multiplicatively symmetric barrier. The full actor objective averages this surrogate over agents and timesteps in the same CTDE training loop as MAPPO and MASPO.



(a) AeroJAX (b) JaxNav (c) MPE (d) PaxMen

(e) RWARE (f) Search and Rescue (g) SMAX (h) Connector

Figure 2: *Aggregated learning performance and probability of improvement.* Results are aggregated across all tasks within each environment using per-task min-max normalization. Main plots depict the mean across 10 independent random seeds (shaded regions denote 95% confidence intervals). Inset plots show the aggregate probability that MARS improves upon baselines; if the probability of improvement is higher than 0.5 and the confidence intervals do not contain 0.5, then the results are statistically significant (Agarwal et al., 2021b).

4 Results

4.1 Experimental Setup

We use the JAX-based Mava library (de Kock et al., 2021) to compare MARS against standard MAPPO and MASPO. To isolate whether MARS gains arise from the symmetric barrier itself rather than from allowing separate expansion and contraction limits, we include four boundary-flexibility ablations. The asymmetric MAPPO ablation replaces the standard clipping interval $[1-\epsilon, 1+\epsilon]$ with $[1-\epsilon_{\text{lower}}, 1+\epsilon_{\text{upper}}]$. The asymmetric MASPO ablation keeps the MASPO quadratic penalty but scales it using a sign-dependent target, $r_{\text{target}} = 1+\epsilon_{\text{upper}}$ for $\widehat{\mathbf{A}}_t \geq 0$ and $r_{\text{target}} = 1-\epsilon_{\text{lower}}$ otherwise. The MARS ablations keep the MARS barrier fixed but constrain the target ratios: multiplicatively symmetric MARS uses $B_{\text{lower}} = 1/B_{\text{upper}}$, while additively symmetric MARS uses $B_{\text{upper}} = 1+\epsilon$ and $B_{\text{lower}} = 1-\epsilon$. Exact objectives and search spaces are given in Appendix E.3.2.

All comparisons therefore keep the simultaneous per-agent actor-update structure fixed and vary only the ratio objective or its trust-region targets. Sequential-update methods such as HATRPO and HAPPO (Kuba et al., 2022) modify the agent update order and factorize the multi-agent policy improvement step, making them complementary to the ratio-geometry question studied here rather than direct controls for this objective-level comparison.

All methods share the same common hyperparameter search space, with trust-region parameters tuned separately using Optuna's Tree-structured Parzen Estimator (Akiba et al., 2019). Methods with one tunable trust-region parameter receive 40 trials per task, while methods with separate upper and lower ratio bounds receive 80 trials per task to account for the larger search space. Agents share policy parameters and observe their agent ID; full hyperparameter ranges and training lengths are given in Appendix E.

We evaluate across 47 cooperative MARL tasks spanning eight environments. Six are drawn from established JAX-based benchmark suites: SMAX, RWARE, Connector, MPE, JaxNav, and Search and Rescue (Bonnet et al., 2024; Rutherford et al., 2024b). We additionally introduce two JAX-native benchmarks designed to stress different multi-agent failure modes: **PaxMen**, a coordinated exploration benchmark with variable team sizes and stochastic maze layouts, and **AeroJAX**, a continuous-control aerial combat benchmark with attitude-dependent sensing and nonlinear 3D flight dynamics. Appendix D details each environment and how PaxMen and AeroJAX differ from the prior benchmarks on which they build.

Table 1: *Aggregate performance by environment.* We report the interquartile mean (IQM) of per-task min-max normalized absolute episode returns (Gorsane et al., 2022), bounded by 95% stratified bootstrap confidence intervals. Scores are aggregated across all tasks within each environment using per-task min-max normalization, then IQM is computed over 10 independent seeds. Bold entries denote the top score for each environment. An asterisk (*) marks results where the confidence interval overlaps with that of the highest score.

Environment	MAPPO	MASPO	MARS
AeroJAX	$0.82_{(0.79,0.85)}$	$0.82_{(0.79,0.84)}$	$\mathbf{0.91}_{(0.90,0.91)}$
Connector	$\mathbf{0.95}^{*}_{(0.94,0.95)}$	$0.93_{(0.93,0.93)}$	$\mathbf{0.95}^{*}_{(0.95,0.95)}$
JaxNav	$0.88_{(0.87,0.89)}$	$\mathbf{0.89}^{*}_{(0.89,0.89)}$	$\mathbf{0.89}^{*}_{(0.88,0.90)}$
MPE	$0.89_{(0.88,0.89)}$	$0.88_{(0.88,0.89)}$	$\mathbf{0.91}_{(0.91,0.92)}$
PaxMen	$0.49_{(0.43,0.54)}$	$0.34_{(0.32,0.36)}$	$\mathbf{0.72}_{(0.59,0.79)}$
RWARE	$0.36_{(0.34,0.40)}$	$0.60_{(0.58,0.61)}$	$\mathbf{0.71}_{(0.69,0.72)}$
S & R	$0.97_{(0.97,0.97)}$	$0.94_{(0.94,0.94)}$	$\mathbf{0.98}_{(0.98,0.98)}$
SMAX	$0.89_{(0.89,0.90)}$	$0.82_{(0.81,0.82)}$	$\mathbf{0.92}_{(0.92,0.92)}$

4.2 Multi-Agent Benchmark Performance

Across 47 tasks in eight cooperative MARL environments, MARS matches or exceeds MAPPO and MASPO in aggregate performance, as shown in Table 1. Figure 2 summarizes environment-level learning curves and probability-of-improvement estimates, while Appendix G provides per-task learning curves for all 47 tasks. The probability-of-improvement insets in Figure 2 indicate statistically significant improvements over both baselines in six of the eight environments. These gains are most pronounced in two regimes: high-variance continuous control and discrete coordination tasks requiring sustained exploration.

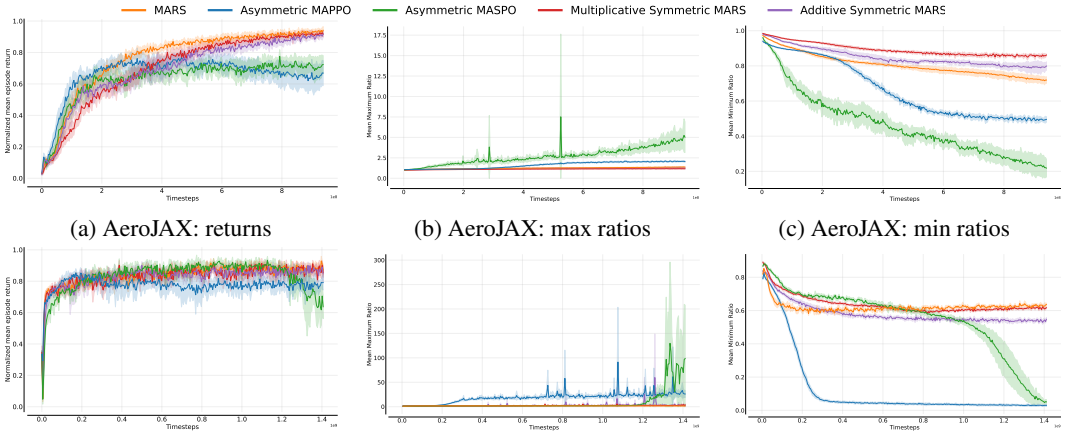
Stability in continuous control. In continuous-control environments, MARS shows improved training stability, particularly in AeroJAX and JaxNav, where policy updates interact with high-variance dynamics and dense ratio changes. In Figure 2a–b, MAPPO and MASPO exhibit late-training degradation or plateauing, while MARS maintains higher returns through the end of training. Similar but smaller gains appear in MPE and Search and Rescue. These results are consistent with the mechanism proposed in Section 3: preserving ratio-space gradients while imposing a barrier near probability extinction can reduce instability from high-variance joint advantages.

Coordination in discrete control. In discrete coordination tasks, MARS performs especially well in sparser-reward environments such as PaxMen and RWARE, where agents must discover coordinated action sequences before receiving useful feedback. Because the MARS penalty assigns unbounded cost near probability extinction, it discourages premature removal of sampled actions from the policy support. This is consistent with the stronger performance observed in Figure 2d–e. In denser tasks such as SMAX and Connector, MARS also matches or improves over the baselines, suggesting that the same ratio-control mechanism remains useful even when rewards provide more frequent feedback.

4.3 Ablating Boundary Flexibility and Barrier Geometry

To separate objective geometry from boundary flexibility, we run long-horizon ablations on two high-variance continuous-control domains: AeroJAX 8v8 and JaxNav 11x11x4a. These tasks provide complementary diagnostics: AeroJAX combines continuous flight dynamics with multi-agent combat coordination, while JaxNav requires decentralized collision avoidance under local sensing. Figure 3 reports learning curves together with maximum- and minimum-ratio diagnostics for the ablation variants defined in Section 4.1.

Objective geometry. We first test whether MARS gains come from the geometric barrier rather than simply from using separately tuned expansion and contraction limits. In both domains, MARS maintains higher or more stable returns than the asymmetric baselines (Figures 3a and 3d). The ratio diagnostics show that asymmetric MASPO develops declining minimum ratios and large maximum-ratio spikes, especially in AeroJAX (Figures 3b–3c), while asymmetric MAPPO exhibits degradation



(a) AeroJAX: returns (b) AeroJAX: max ratios (c) AeroJAX: min ratios

(d) JaxNav: returns (e) JaxNav: max ratios (f) JaxNav: min ratios

Figure 3: *Ablation analysis across two continuous-control domains.* We compare MARS against boundary-flexibility controls: asymmetric MAPPO, asymmetric MASPO, and two MARS target-parameterization variants. Across AeroJAX 8v8 and JaxNav 11x11x4a, MARS variants maintain more stable ratio dynamics than the asymmetric baselines: minimum and maximum ratios spike in asymmetric MASPO, while asymmetric MAPPO's minimum ratios degrade. These failures are consistent with asymmetric MASPO's finite resistance to vanishing ratios and asymmetric MAPPO's clipped-gradient recovery failure. The stability of the MARS variants indicates that the geometric barrier is the main source of ratio stability, while flexible expansion and contraction targets primarily affect learning efficiency.

in minimum ratios despite avoiding the largest upward spikes. In JaxNav, asymmetric MAPPO shows a sharper early minimum-ratio drop, while MASPO develops later ratio instability (Figures 3e–3f). In contrast, the MARS variants maintain more stable ratio dynamics across both environments. These results support the claim that the MARS penalty shape, not only boundary flexibility, contributes to stability.

Domain-specific failure modes. The two environments show different versions of the same ratio-control problem. In AeroJAX, asymmetric MASPO shows maximum-ratio spikes together with decreasing minimum ratios, consistent with finite resistance as ratios approach zero. Asymmetric MAPPO degrades more gradually, consistent with clipped samples losing corrective gradients after leaving the nominal trust region. In JaxNav, the MAPPO failure is sharper: minimum ratios drop early in training, while MASPO learns initially but has unstable ratios. Thus, the MARS barrier stabilizes ratio dynamics across environments even when the baseline failure mode differs.

Boundary symmetry and parameterization. Finally, we compare MARS with target-parameterization ablations that keep the MARS penalty fixed while restricting how expansion and contraction targets are chosen. Multiplicatively symmetric MARS maintains stable ratio dynamics in both domains, especially in the minimum-ratio panels (Figures 3c and 3f), indicating that the symmetric barrier is sufficient to avoid the most severe ratio pathologies. Additively symmetric MARS is also stable but learns more slowly, especially in AeroJAX (Figure 3a), suggesting that additive target choices are less efficient even with the MARS barrier. MARS with independently tuned expansion and contraction targets can learn faster or reach slightly higher returns, particularly in AeroJAX (Figure 3a), so target-ratio flexibility mainly affects learning speed rather than ratio stability.

5 Conclusion

We introduced Multi-Agent Ratio Symmetry (MARS), a policy optimization objective for CTDE cooperative MARL. MARS addresses two limitations of additive ratio-based trust-region methods: clipped objectives can remove corrective gradients for ratio outliers, while finite additive penalties can allow probability collapse under large negative joint advantages. By replacing additive ratio

control with a multiplicatively symmetric geometric barrier, MARS preserves ratio-space gradients and makes probability extinction non-optimal for the per-sample surrogate. Aggregating performance across 47 tasks in eight environments, MARS matches or outperforms MAPPO and MASPO, with ablations indicating that the curvature of the symmetric barrier is the primary source of stability. These results suggest that aligning ratio regularization with multiplicative probability geometry is a useful direction for robust multi-agent policy optimization.

Acknowledgments and Disclosure of Funding

This work was partially supported by the National Science Foundation (NSF) under Grant No. 2525087. The authors declare no competing interests.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98): 1–76, 2021a.

Agarwal, R., Schwarzer, M., Castro, P. S., Courville, A. C., and Bellemare, M. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems*, 34:29304–29320, 2021b.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2623–2631, 2019.

Belousov, B. and Peters, J. f-divergence constrained policy improvement, 2018. URL `https://arxiv.org/abs/1801.00056`.

Bonnet, C., Luo, D., Byrne, D., Surana, S., Abramowitz, S., Duckworth, P., et al. Jumanji: a diverse suite of scalable reinforcement learning environments in JAX. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

de Kock, R., Mahjoub, O., Abramowitz, S., Khlifi, W., Tilbury, C. R., Formanek, C., Smit, A., and Pretorius, A. Mava: a research library for distributed multi-agent reinforcement learning in jax. *arXiv preprint arXiv:2107.01460*, 2021.

Ellis, B., Cook, J., Moalla, S., Samvelyan, M., Sun, M., Mahajan, A., Foerster, J., and Whiteson, S. Smacv2: An improved benchmark for cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 36:37567–37593, 2023.

Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Janoos, F., Rudolph, L., and Madry, A. Implementation matters in deep rl: A case study on ppo and trpo. In *International conference on learning representations*, 2019.

Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., and Whiteson, S. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Gorsane, R., Mahjoub, O., de Kock, R. J., Dubb, R., Singh, S., and Pretorius, A. Towards a standardised performance evaluation protocol for cooperative marl. *Advances in Neural Information Processing Systems*, 35:5510–5521, 2022.

Kobayashi, T. Proximal policy optimization with relative pearson divergence*. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8416–8421. IEEE Press, 2021. doi: 10.1109/ICRA48506.2021.9560856. URL `https://doi.org/10.1109/ICRA48506.2021.9560856`.

Kobayashi, T. Proximal policy optimization with adaptive threshold for symmetric relative density ratio. *Results in Control and Optimization*, 10:100192, 2023.

Kuba, J. G., Chen, R., Wen, M., Wen, Y., Sun, F., Wang, J., and Yang, Y. Trust region policy optimisation in multi-agent reinforcement learning. In *International Conference on Learning Representations*, 2022.

Lechner, M., Seyde, T., Wang, T.-H. J., Xiao, W., Hasani, R., Rountree, J., Rus, D., et al. Gigastep-one billion steps per second multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 36:155–170, 2023.

Li, C., Wang, T., Wu, C., Zhao, Q., Yang, J., and Zhang, C. Celebrating diversity in shared multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:3991–4002, 2021.

Liu, H.-S., Kuroki, S., Kozuno, T., Sun, W.-F., and Lee, C.-Y. Language-guided pattern formation for swarm robotics with multi-agent reinforcement learning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8998–9005. IEEE, 2024.

Lyu, X., Baisero, A., Xiao, Y., Daley, B., and Amato, C. On centralized critics in multi-agent reinforcement learning. *Journal of Artificial Intelligence Research*, 77:295–354, 2023.

Mahjoub, O., Abramowitz, S., de Kock, R. J., Khlifi, W., Toit, S. V. D., Daniel, J., Nessir, L. B., Beyers, L., Formanek, J. C., Clark, L., and Pretorius, A. Sable: a performant, efficient and scalable sequence model for MARL. In *Forty-second International Conference on Machine Learning*, 2025.

Oliehoek, F. A. and Amato, C. *A concise introduction to decentralized POMDPs*, volume 1. Springer, 2016.

Rutherford, A., Beukman, M., Willi, T., Lacerda, B., Hawes, N., and Foerster, J. No regrets: Investigating and improving regret approximations for curriculum discovery. *Advances in Neural Information Processing Systems*, 37:16071–16101, 2024a.

Rutherford, A., Ellis, B., Gallici, M., Cook, J., Lupu, A., Ingvarsson Juto, G., Willi, T., Hammond, R., Khan, A., Schroeder de Witt, C., et al. Jaxmarl: Multi-agent rl environments and algorithms in jax. *Advances in Neural Information Processing Systems*, 37:50925–50951, 2024b.

Saifullah, M., Papakonstantinou, K., Andriotis, C., and Stoffels, S. Multi-agent deep reinforcement learning with centralized training and decentralized execution for transportation infrastructure management. *arXiv preprint arXiv:2401.12455*, 2024.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Selmonaj, A., Del Rio, G., Schneider, A., and Antonucci, A. Coordinated strategies in realistic air combat by hierarchical multi-agent reinforcement learning. *arXiv preprint arXiv:2510.11474*, 2025.

Wang, Y., He, H., and Tan, X. Truly proximal policy optimization. In *Uncertainty in artificial intelligence*, pp. 113–122. PMLR, 2020.

Xie, Z., Zhang, Q., Yang, F., Hutter, M., and Xu, R. Simple policy optimization. In *Forty-second International Conference on Machine Learning*, 2025.

Xu, H., Yan, Z., Xuan, J., Zhang, G., and Lu, J. Improving proximal policy optimization with alpha divergence. *Neurocomput.*, 534(C):94–105, May 2023. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.02.008. URL `https://doi.org/10.1016/j.neucom.2023.02.008`.

Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., and Wu, Y. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 35: 24611–24624, 2022.

A Limitations

MARS is designed for cooperative CTDE policy-gradient reinforcement learning with decentralized actors and joint advantage estimates. Its theoretical results characterize the per-sample ratio surrogate in ratio space, showing that the objective avoids clipped gradient plateaus and makes probability-ratio extinction non-optimal. These results do not constitute a global convergence guarantee for nonlinear neural-network training or a monotonic improvement guarantee for full multi-agent policy updates.

Our experiments focus on cooperative MARL settings with shared rewards and centralized critics. Although the benchmark suite spans 47 tasks across eight environments, it does not cover competitive, mixed-motive, or open-ended multi-agent settings. The behavior of MARS in settings with independent rewards, learned communication, population-based training, or very large heterogeneous agent populations remains an open question.

MARS also introduces target ratio parameters for expansion and contraction. We tune these parameters using the same protocol as the baselines and include ablations to separate objective geometry from boundary flexibility, but the method still requires selecting appropriate target ranges. Future work could study adaptive target selection, extensions to other multi-agent training paradigms, and stronger theoretical guarantees for full policy updates.

B Related Work: Divergence-Based Policy Optimization

Divergence-based regularization has been explored in single-agent policy optimization. Belousov & Peters (2018) study policy improvement under general (f)-divergence regularization, including a generator proportional to the MARS penalty ($\psi(r) = r + r^{-1} - 2$). Similarly, AlphaPPO (Xu et al., 2023) replaces the KL regularizer in PPO-Penalty with an alpha-divergence family encompassing reverse-Pearson/Neyman geometry. These approaches establish precedents for the underlying divergence geometry but optimize policy-level objectives rather than advantage-conditioned, target-aligned sampled-ratio surrogates.

PPO-RPE (Kobayashi, 2021, 2023) instead regularizes a transformed probability ratio, yielding a penalty that remains finite as the original ratio approaches zero and is not invariant under ratio inversion. In contrast, MARS applies an inversion-symmetric barrier directly to per-agent ratios within simultaneous CTDE updates, using centralized joint advantages and sample-dependent coefficients to align the surrogate with separate expansion and contraction targets.

C Proofs and Derivations

C.1 MAPPO Clipped Objective Gradient

The per-agent MAPPO clipped surrogate is

$$\mathcal{L}_{i,t}^{\mathrm{MAPPO}}(\theta) = \min\left(r_{i,t}(\theta)\widehat{\mathbf{A}}_t,\ \mathrm{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon)\widehat{\mathbf{A}}_t\right), \tag{13}$$

where

$$r_{i,t}(\theta) = \frac{\pi_\theta(a_{i,t} \mid h_{i,t})}{\pi_{\theta_{\mathrm{old}}}(a_{i,t} \mid h_{i,t})}$$

is the per-agent probability ratio and $\widehat{\mathbf{A}}_t$ is the joint advantage estimate. We derive the derivative of this objective with respect to $r_{i,t}$ away from the clipping kink points. At the clipping boundaries, the derivative can be interpreted using subgradients; these measure-zero cases do not affect the truncation argument.

Inside the trust region. If $|r_{i,t}(\theta) - 1| \le \epsilon$, then

$$\mathrm{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon) = r_{i,t}(\theta).$$

Therefore,

$$\mathcal{L}_{i,t}^{\mathrm{MAPPO}} = r_{i,t}(\theta)\widehat{\mathbf{A}}_t, \qquad \frac{\partial \mathcal{L}_{i,t}^{\mathrm{MAPPO}}}{\partial r_{i,t}} = \widehat{\mathbf{A}}_t. \tag{14}$$

Above the trust region. If $r_{i,t}(\theta) > 1+\epsilon$, then the clipped ratio is $1+\epsilon$, and

$$\mathcal{L}_{i,t}^{\mathrm{MAPPO}} = \min\left(r_{i,t}\widehat{\mathbf{A}}_t,\, (1+\epsilon)\widehat{\mathbf{A}}_t\right).$$

If $\widehat{\mathbf{A}}_t > 0$, the clipped term is smaller, so the objective is constant with respect to $r_{i,t}$ and the derivative is zero. If $\widehat{\mathbf{A}}_t < 0$, the unclipped linear term is smaller, so the derivative is $\widehat{\mathbf{A}}_t$.

Below the trust region. If $r_{i,t}(\theta) < 1-\epsilon$, then the clipped ratio is $1-\epsilon$, and

$$\mathcal{L}_{i,t}^{\mathrm{MAPPO}} = \min\left(r_{i,t}\widehat{\mathbf{A}}_t,\, (1-\epsilon)\widehat{\mathbf{A}}_t\right).$$

If $\widehat{\mathbf{A}}_t > 0$, the unclipped linear term is smaller, so the derivative is $\widehat{\mathbf{A}}_t$. If $\widehat{\mathbf{A}}_t < 0$, the clipped term is smaller, so the objective is constant with respect to $r_{i,t}$ and the derivative is zero.

Combining these cases gives

$$\frac{\partial \mathcal{L}_{i,t}^{\mathrm{MAPPO}}}{\partial r_{i,t}} = \begin{cases} \widehat{\mathbf{A}}_t, & \text{if } |r_{i,t}(\theta)-1| < \epsilon \text{ or } \operatorname{sign}(\widehat{\mathbf{A}}_t)(r_{i,t}(\theta)-1) < 0, \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

Thus, MAPPO removes the ratio-gradient exactly for samples that cross the clipping boundary in the advantage-improving direction.

C.2 Symmetric Penalty Derivation

MASPO uses an additive quadratic penalty on the probability ratio,

$$\psi_{\mathrm{MASPO}}(r) = (r-1)^2. \tag{16}$$

For probability ratios, the natural inverse of expansion by r is contraction by $1/r$. To construct a penalty invariant under ratio inversion, we symmetrize the MASPO penalty by taking the geometric mean of the penalty evaluated at r and at $1/r$:

$$\psi_{\mathrm{MARS}}(r) = \sqrt{\psi_{\mathrm{MASPO}}(r)\psi_{\mathrm{MASPO}}(1/r)}, \qquad r > 0. \tag{17}$$

Now,

$$\psi_{\mathrm{MASPO}}(1/r) = \left(\frac{1}{r}-1\right)^2 = \left(\frac{1-r}{r}\right)^2 = \frac{(r-1)^2}{r^2}. \tag{18}$$

Substituting,

$$\psi_{\mathrm{MARS}}(r) = \sqrt{(r-1)^2\frac{(r-1)^2}{r^2}} = \sqrt{\frac{(r-1)^4}{r^2}}. \tag{19}$$

Since $r > 0$ and $(r-1)^2 \geq 0$,

$$\sqrt{\frac{(r-1)^4}{r^2}} = \frac{(r-1)^2}{r}. \tag{20}$$

Expanding yields

$$\psi_{\mathrm{MARS}}(r) = \frac{(r-1)^2}{r} = \frac{r^2-2r+1}{r} = r + \frac{1}{r} - 2. \tag{21}$$

This penalty is zero at $r = 1$, symmetric under inversion, and diverges as $r \to 0^+$.

C.3 Proof of Proposition 3.1: Barrier Against Probability Extinction

Let $r \in (0,\infty)$, let $\widehat{\mathbf{A}} \in \mathbb{R}$ be finite, and let $\alpha > 0$. Consider

$$\mathcal{L}(r) = r\widehat{\mathbf{A}} - \alpha\psi_{\mathrm{MARS}}(r), \qquad \psi_{\mathrm{MARS}}(r) = r + \frac{1}{r} - 2. \tag{22}$$

Substituting the penalty gives

$$\mathcal{L}(r) = r\widehat{\mathbf{A}} - \alpha r - \frac{\alpha}{r} + 2\alpha. \tag{23}$$

Objective limit. As $r \to 0^+$, the terms $r\widehat{\mathbf{A}}$ and $-\alpha r$ vanish, and the constant 2α remains finite. The dominant term is $-\alpha/r$. Since $\alpha > 0$,

$$\lim_{r\to 0^+} -\frac{\alpha}{r} = -\infty. \tag{24}$$

Therefore,

$$\lim_{r\to 0^+} \mathcal{L}(r) = -\infty. \tag{25}$$

Gradient limit. Differentiating with respect to r,

$$\frac{\partial \mathcal{L}}{\partial r} = \widehat{\mathbf{A}} - \alpha\left(1 - \frac{1}{r^2}\right) = \widehat{\mathbf{A}} - \alpha + \frac{\alpha}{r^2}. \tag{26}$$

As $r \to 0^+$, the finite terms $\widehat{\mathbf{A}} - \alpha$ are dominated by α/r^2. Since $\alpha > 0$,

$$\lim_{r\to 0^+} \frac{\alpha}{r^2} = +\infty. \tag{27}$$

Hence,

$$\lim_{r\to 0^+} \frac{\partial \mathcal{L}}{\partial r} = +\infty. \tag{28}$$

Thus, probability extinction is not an optimizer of the per-sample ratio surrogate, and in a neighborhood of $r = 0$ the ratio-gradient points toward increasing r.

C.4 Proof of Proposition 3.2: No Gradient Truncation

Let $r \in (0, \infty)$, let $\widehat{\mathbf{A}} \in \mathbb{R}$ be finite, and let $\alpha > 0$. The MARS surrogate is

$$\mathcal{L}(r) = r\widehat{\mathbf{A}} - \alpha\left(r + \frac{1}{r} - 2\right). \tag{29}$$

Smoothness. The function $\mathcal{L}$ is a linear combination of polynomial terms and the rational term $1/r$. The only singularity of $1/r$ occurs at $r = 0$, which is outside the domain $(0, \infty)$. Therefore, $\mathcal{L}$ is C^∞ on $(0, \infty)$.

No gradient truncation. The ratio-gradient is

$$\frac{\partial \mathcal{L}}{\partial r} = \widehat{\mathbf{A}} - \alpha + \frac{\alpha}{r^2}. \tag{30}$$

Suppose the gradient vanishes. Then

$$\widehat{\mathbf{A}} - \alpha + \frac{\alpha}{r^2} = 0, \tag{31}$$

which implies

$$\frac{\alpha}{r^2} = \alpha - \widehat{\mathbf{A}}. \tag{32}$$

If $\alpha - \widehat{\mathbf{A}} \leq 0$, this equation has no solution on $r > 0$. If $\alpha - \widehat{\mathbf{A}} > 0$, then

$$r = \sqrt{\frac{\alpha}{\alpha - \widehat{\mathbf{A}}}}, \tag{33}$$

which gives exactly one positive stationary point.

Therefore, the zero-gradient set is either empty or a singleton. In particular, there is no nontrivial interval on which $\partial\mathcal{L}/\partial r = 0$. Hence, unlike a clipped objective, the MARS trust-region mechanism introduces no gradient truncation in ratio space.

C.5 Trust Region Boundaries

The MARS ratio-gradient for agent i at timestep t is

$$\frac{\partial \mathcal{L}_{i,t}^{\mathrm{MARS}}}{\partial r_{i,t}} = \widehat{\mathbf{A}}_t - \alpha_t \left(1 - \frac{1}{r_{i,t}^2}\right). \tag{34}$$

To align the stationary point with an intended trust-region boundary, we choose a target ratio $r_{\mathrm{target}} > 0$ with $r_{\mathrm{target}} \neq 1$ and enforce

$$\left.\frac{\partial \mathcal{L}_{i,t}^{\mathrm{MARS}}}{\partial r_{i,t}}\right|_{r_{i,t}=r_{\mathrm{target}}} = 0. \tag{35}$$

This gives

$$0 = \widehat{\mathbf{A}}_t - \alpha_t \left(1 - r_{\mathrm{target}}^{-2}\right), \tag{36}$$

and therefore

$$\alpha_t = \frac{\widehat{\mathbf{A}}_t}{1 - r_{\mathrm{target}}^{-2}}. \tag{37}$$

Additive ϵ-aligned target. If the target is expressed using the standard additive boundary,

$$r_{\mathrm{target}} = 1 + \mathrm{sign}(\widehat{\mathbf{A}}_t)\epsilon, \tag{38}$$

then

$$\alpha_t = \frac{\widehat{\mathbf{A}}_t}{1 - \left(1 + \mathrm{sign}(\widehat{\mathbf{A}}_t)\epsilon\right)^{-2}}. \tag{39}$$

Multiplicative target ratios. For MARS, we specify the expansion and contraction targets directly:

$$r_{\mathrm{target}} = \begin{cases} B_{\mathrm{upper}}, & \widehat{\mathbf{A}}_t \geq 0, \\ B_{\mathrm{lower}}, & \widehat{\mathbf{A}}_t < 0. \end{cases} \tag{40}$$

This yields

$$\alpha_t = \begin{cases} \dfrac{\widehat{\mathbf{A}}_t}{1 - B_{\mathrm{upper}}^{-2}}, & \widehat{\mathbf{A}}_t > 0, \\ \dfrac{\widehat{\mathbf{A}}_t}{1 - B_{\mathrm{lower}}^{-2}}, & \widehat{\mathbf{A}}_t < 0. \end{cases} \tag{41}$$

When $B_{\mathrm{upper}} > 1$ and $B_{\mathrm{lower}} < 1$, $\alpha_t > 0$ in both nonzero-advantage regimes. For $\widehat{\mathbf{A}}_t > 0$, the numerator and denominator are both positive. For $\widehat{\mathbf{A}}_t < 0$, the numerator is negative and the denominator is negative because $B_{\mathrm{lower}}^{-2} > 1$. If $\widehat{\mathbf{A}}_t = 0$, the corresponding policy-gradient term is zero; in implementation, this sample can be skipped for the actor update or assigned $\alpha_t = 0$.

D Environments

D.1 AeroJAX

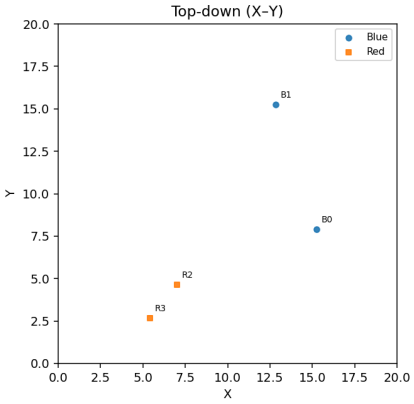


Figure 4: Rendering of a two versus two dogfighting AeroJAX scenario.

AeroJAX is a JAX-native multi-agent aerial combat environment built from the Gigastep framework (Lechner et al., 2023). Whereas Gigastep emphasizes high-throughput, arcade-style multi-agent simulation, AeroJAX increases the fidelity of the control problem by introducing decoupled 3D flight dynamics and attitude-dependent combat sensing. Agents must manage six-degree-of-freedom maneuvering, energy loss, lift-induced drag, turn coordination, and throttle control while engaging heuristic opponents in continuous 3D dogfighting scenarios.

A key distinction from the original Gigastep setting is that sensing and combat are tied to aircraft attitude. AeroJAX uses view and weapons cones that rotate with each aircraft's roll and pitch, so maintaining target lock requires coordinated control of orientation, speed, and relative position rather than planar pursuit alone. This makes the benchmark distinct from simplified kinematic environments such as MPE, JaxNav, and Search and Rescue: the learning problem combines continuous control, nonlinear flight coupling, partial observability, and multi-agent combat coordination. As a result, AeroJAX provides a high-variance stress test for policy optimization stability under dense ratio changes and complex dynamics.

Observation space. Each agent observes the environment through a limited view cone defined by specific horizontal and vertical angles and a maximum depth. The observation vector includes ego-specific flight data (roll, pitch, angle of attack, flight path angle) and a sorted list of detected entities. Observed attributes for these entities include relative position, heading, speed, health, team alignment, and unit type. Detection is stochastic based on distance, but agents can share target information with allies within a defined communication range.

Action space. The environment utilizes a continuous multi-dimensional action space representing aircraft control surfaces. Agents output a 3-dimensional vector corresponding to aileron command (roll rate), elevator command (angle of attack rate), and throttle (thrust fraction).

Reward. Agents achieve sparse rewards for killing enemies and sparse penalties for ally deaths. Winning the engagement (ending the episode with more allies alive than enemies) also provides a sparse reward.

D.2 PaxMen

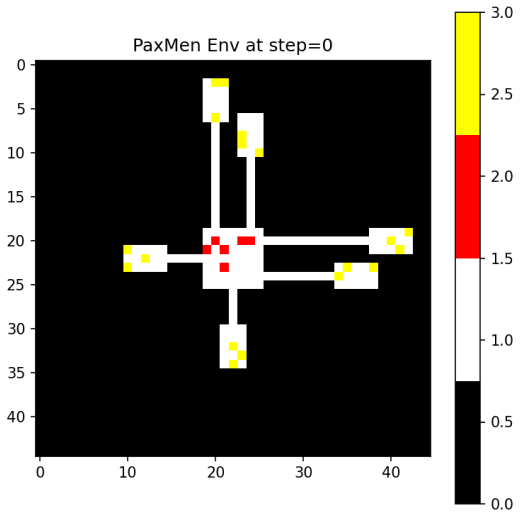


Figure 5: Rendering of an example six-agent PaxMen layout.

PaxMen is a JAX-native cooperative exploration benchmark based on the PacMen environment (Li et al., 2021). The original PacMen task uses a shared multi-agent setting in which agents coordinate to collect dots in a structured maze. PaxMen extends this setting by supporting larger team sizes, including four-, five-, and six-agent configurations, and by introducing stochastic layout variation across episodes. Rather than training on a single static map, agents encounter a fixed set of maze topologies with varying corridor orientations and lengths.

This design makes PaxMen a diagnostic benchmark for decentralized coordinated exploration. Agents begin in a central hub and must disperse into branching corridors to maximize coverage without explicit communication. Because agents are homogeneous and share rewards, successful policies must break symmetry and adopt complementary roles; policies that send multiple agents down the same corridor waste exploration capacity. This distinguishes PaxMen from combat micro-management tasks such as SMAX and routing or transport benchmarks such as RWARE and Connector. PaxMen instead isolates the challenge of sustained exploration and role differentiation under partial observability.

Observation space. Each agent observes a partial, local view of the environment defined by a 5×5 grid centered on the agent. The observation is a flattened vector containing values representing walls, empty space, other agents, and dots.

Action space. The action space is discrete and consists of five actions: four cardinal movement directions (Up, Down, Left, Right) and a dedicated Eat action. Movement into walls is invalid and results in no position change.

Reward. Agents receive a global reward calculated as the sum of all individual agent contributions at each timestep. A positive reward of +1.0 is generated for consuming a dot; if multiple agents consume the same dot simultaneously, the reward is split such that the total team gain remains +1.0. A small step penalty of -0.025 is applied to encourage efficient exploration. The dots respawn once all have been consumed.

D.3 JaxNav

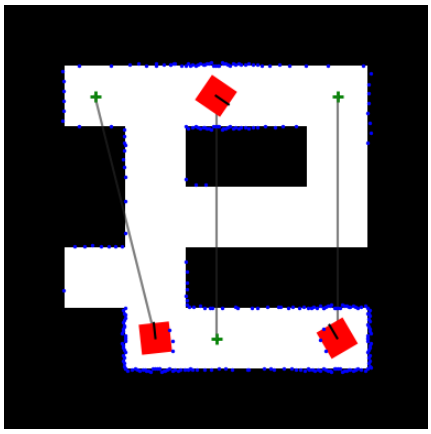


Figure 6: Rendering of an example three-agent JaxNav navigation scenario.

JaxNav is a JAX-based continuous 2D navigation environment for differential-drive robots, originally introduced by Rutherford et al. (2024a) and included in the JaxMARL suite (Rutherford et al., 2024b). The environment tasks agents with navigating to specific goal regions while avoiding collisions with static obstacles and other dynamic agents. It relies exclusively on local range sensing without inter-agent communication, requiring agents to learn decentralized collision avoidance strategies.

Observation space. Each agent observes a local vector comprising three primary components: a set of 200 LiDAR range readings from a 360° scan (normalized by the maximum range of 6m), the agent's current linear and angular velocities, and the relative goal direction in polar coordinates. The goal distance is clipped to the maximum LiDAR range to preserve local-only sensing constraints.

Action space. The environment features a continuous 2D action space governing differential-drive dynamics. Agents output a vector specifying the target linear velocity (v) and angular velocity (ω), which are executed subject to acceleration limits and kinematic constraints.

Reward. Agents receive a dense shaped reward based on their progress toward the goal (weighted at 0.25) combined with a sparse bonus of +4.0 for successfully reaching the target radius (0.3m). To ensure safety, penalties are applied for collisions (-4.0) and for breaching a minimum safety distance (LiDAR proximity < 0.3m). A small step penalty of -0.01 is applied to encourage time-optimal paths. In the multi-agent setting, the final reward is a convex combination of the agent's individual reward and the collective team reward ($\lambda = 0.5$).

D.4 Search and Rescue

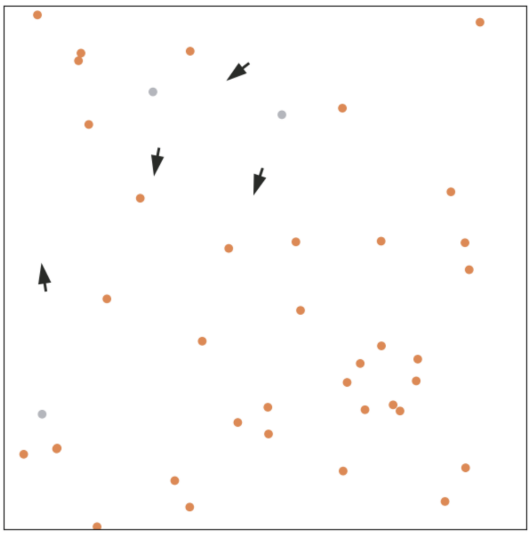


Figure 7: Rendering of an example four-agent Search and Rescue scenario.

Search and Rescue is a continuous-space multi-agent coordination task from the Jumanji suite (Bonnet et al., 2024). A team of homogeneous searchers operates within a 2D square arena featuring wrap-around (toroidal) boundaries. The objective is to locate a set of potentially moving targets within a fixed time horizon. A target is considered "found" when a searcher navigates within a specific contact radius while having the target within its view cone.

Observation space. Agents perceive the environment through a segmented, ray-based local view. The primary observation is a multi-channel ray array where each ray encodes the normalized distance to the nearest entity along that vector. Channels separate entity types into other searchers, found targets, and unfound targets. Additionally, the observation vector includes lightweight scalar context, such as the fraction of targets remaining and the current timestep fraction.

Action space. Agents operate in a continuous multi-dimensional action space. Each agent outputs a 2-dimensional vector determining its change in heading ($\Delta\theta$) and velocity (Δv). These commands are scaled by environment parameters to update the searcher's position and orientation, with speed clipped to a fixed operational interval.

Reward. Rewards are strictly event-based and issued for *new* detections. An agent receives a reward of +1.0 when it detects a target that was previously unfound. If multiple agents discover the same target simultaneously, the reward is split evenly to maintain a constant team utility. By default, the magnitude of these rewards decays linearly over the episode duration to encourage rapid discovery.

D.5 Other environments

We refer to the Appendix in Mahjoub et al. (2025) for thorough descriptions of the other environments we use (Connector, MPE, RWARE, and SMAX) rather than re-writing them here.

E Hyperparameters

E.1 Overview

We use Optuna (Akiba et al., 2019), a Bayesian optimization framework, to systematically search for optimal hyperparameters across all algorithms. Our tuning procedure covers MARS, MAPPO, MASPO, and their ablation variants.

We employ the Tree-structured Parzen Estimator (TPE) sampler. MARS, asymmetric MAPPO, and asymmetric MASPO are allocated 80 trials for their tuning budget, while MAPPO, MASPO, and additive symmetric MARS and multiplicative symmetric MARS are allocated 40 trials for their tuning budget. This larger tuning budget accounts for the additional trust-region parameter in methods with independently tuned upper and lower bounds.

E.2 Evaluation Protocol

Each trial follows a rigorous multi-seed evaluation protocol to ensure statistical reliability:

1. **Seeds:** Five fixed seeds are used for each trial
2. **Sequential evaluation:** Seeds are evaluated one at a time with intermediate pruning checks
3. **Performance metric:** Mean evaluation return averaged over the final 20% of evaluation steps, which smooths end-of-training noise while focusing on final convergence quality
4. **Trial objective:** Mean performance across all 5 seeds
5. **Optimization direction:** Maximize evaluation return

E.3 Hyperparameter Search Spaces

E.3.1 Common Hyperparameters

Table 2 shows the common hyperparameters tuned across all algorithms, inspired by the hyperparameter search space used by Mahjoub et al. (2025).

Table 2: Common hyperparameters tuned across all algorithms.

Hyperparameter	Search Space
Actor learning rate	$\{10^{-4}, 2.5 \times 10^{-4}, 5 \times 10^{-4}\}$
Critic learning rate	$\{10^{-4}, 2.5 \times 10^{-4}, 5 \times 10^{-4}\}$
Entropy coefficient	$\{0.0, 10^{-5}, 10^{-2}\}$
Value clipping epsilon	$\{0.05, 0.1, 0.2\}$
Number of minibatches	$\{2, 4, 8\}$
PPO epochs	$\{2, 4, 8\}$

E.3.2 Algorithm-Specific Hyperparameters

All ablation variants use the same rollout, critic, optimizer, and shared-policy training loop as their corresponding base method. They differ only in the ratio objective or in how the trust-region target ratios are parameterized.

MARS. MARS uses the per-agent surrogate

$$\mathcal{L}_{i,t}^{\mathrm{MARS}}(\theta) = r_{i,t}(\theta)\widehat{\mathbf{A}}_t - \alpha_t \left(r_{i,t}(\theta) + \frac{1}{r_{i,t}(\theta)} - 2 \right),$$

with independently tuned expansion and contraction targets

$$r_{\mathrm{target}} = \begin{cases} B_{\mathrm{upper}}, & \widehat{\mathbf{A}}_t \geq 0, \\ B_{\mathrm{lower}}, & \widehat{\mathbf{A}}_t < 0, \end{cases} \qquad \alpha_t = \frac{\widehat{\mathbf{A}}_t}{1 - r_{\mathrm{target}}^{-2}}.$$

We tune

$$B_{\mathrm{upper}} \in \{1.05, 1.1, 1.2, 2, 4, 6, 8, 10\}, \quad B_{\mathrm{lower}} \in \{0.95, 0.9, 0.8, 0.7, 0.5, 0.3, 0.1\}.$$

Multiplicatively symmetric MARS. This ablation keeps the MARS loss unchanged but constrains the target ratios to be reciprocal. For a single tuned target $b > 1$,

$$B_{\mathrm{upper}} = b, \qquad B_{\mathrm{lower}} = \frac{1}{b}.$$

Thus expansion and contraction targets are multiplicative inverses. We tune

$$b \in \{1.05, 1.1, 1.2, 1.5, 1.9\}.$$

Additively symmetric MARS. This ablation also keeps the MARS loss unchanged but constrains the target ratios using MAPPO-style additive symmetry. For a single tuned target $b > 1$, equivalently $\epsilon = b - 1$,

$$B_{\text{upper}} = b = 1 + \epsilon, \qquad B_{\text{lower}} = 2 - b = 1 - \epsilon.$$

Thus the only difference from MARS is the target-ratio parameterization used to compute α_t. We tune

$$b \in \{1.05, 1.1, 1.2, 1.5, 1.9\}.$$

MAPPO. Standard MAPPO uses the clipping interval $[1 - \epsilon, 1 + \epsilon]$:

$$\mathcal{L}_{i,t}^{\text{MAPPO}}(\theta) = \min\left(r_{i,t}(\theta)\widehat{\mathbf{A}}_t, \text{clip}(r_{i,t}(\theta), 1 - \epsilon, 1 + \epsilon)\widehat{\mathbf{A}}_t\right).$$

We tune

$$\epsilon \in \{0.05, 0.1, 0.2\}.$$

Asymmetric MAPPO. For the asymmetric MAPPO ablation, we replace the standard symmetric clipping interval with independently tuned lower and upper clipping bounds:

$$\text{clip}\left(r_{i,t}(\theta), 1 - \epsilon_{\text{lower}}, 1 + \epsilon_{\text{upper}}\right).$$

The objective is otherwise identical to MAPPO. We tune

$$\epsilon_{\text{upper}} \in \{0.05, 0.1, 0.2, 1.0, 3.0, 5.0, 7.0, 9.0\}, \quad \epsilon_{\text{lower}} \in \{0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 0.9\}.$$

MASPO. Standard MASPO uses the quadratic penalty

$$\mathcal{L}_{i,t}^{\text{MASPO}}(\theta) = r_{i,t}(\theta)\widehat{\mathbf{A}}_t - \frac{|\widehat{\mathbf{A}}_t|}{2\epsilon}\left(r_{i,t}(\theta) - 1\right)^2.$$

We tune the same single trust-region parameter as standard MAPPO:

$$\epsilon \in \{0.05, 0.1, 0.2\}.$$

Asymmetric MASPO. For the asymmetric MASPO ablation, we keep the MASPO quadratic penalty but choose its scale using a sign-dependent target ratio:

$$r_{\text{target}} = \begin{cases} 1 + \epsilon_{\text{upper}}, & \widehat{\mathbf{A}}_t \geq 0, \\ 1 - \epsilon_{\text{lower}}, & \widehat{\mathbf{A}}_t < 0. \end{cases}$$

Equivalently, the quadratic penalty coefficient is chosen so that the one-step MASPO surrogate has its stationary point at the corresponding upper or lower target:

$$\mathcal{L}_{i,t}^{\text{asym-MASPO}}(\theta) = r_{i,t}(\theta)\widehat{\mathbf{A}}_t - \frac{\widehat{\mathbf{A}}_t}{2(r_{\text{target}} - 1)}\left(r_{i,t}(\theta) - 1\right)^2.$$

This coefficient is positive for both $\widehat{\mathbf{A}}_t > 0$ and $\widehat{\mathbf{A}}_t < 0$ because $r_{\text{target}} > 1$ in the positive-advantage case and $r_{\text{target}} < 1$ in the negative-advantage case. We tune the same independently selected upper and lower values used for asymmetric MAPPO.

Ablation Fairness. For fairness in our ablation study in Section 4.3, we re-tune MARS using the exact same asymmetric clipping bounds (converted to their target ratio forms as described in Appendix Section C.5) as asymmetric MAPPO and asymmetric MASPO; in other words, all algorithms share an **identical** hyperparameter search space. This ensures that, within the boundary-flexibility ablation, the differentiating factor between the three objectives is the objective itself.

E.4 Fixed Hyperparameters

To focus the hyperparameter search, we fix the following parameters across all trials (Table 3).

Table 3: Fixed hyperparameters across all tuning experiments.

Hyperparameter	Value
Agent ID encoding	True
Update batch size	2
Rollout length	128
Max gradient norm	0.5
GAE λ	0.95
Discount factor (γ)	0.99
Value function coefficient	0.5
Learning rate decay	False

E.5 Training/Tuning Lengths

We run each seed of each trial of each task for every environment except AeroJAX and JaxNav for approximately 40 million timesteps. Due to the higher complexity of AeroJAX and JaxNav, we allow for additional training time for those seeds, approximately 120 million timesteps.

Final evaluation runs using tuned hyperparameters are run for twice the tuning length: approximately 240 million timesteps for AeroJAX and JaxNav and approximately 80 million timesteps for all other environments. Ablation runs in Section 4.3 are run for approximately one billion timesteps to evaluate long-horizon stability.

F Algorithm Pseudocode

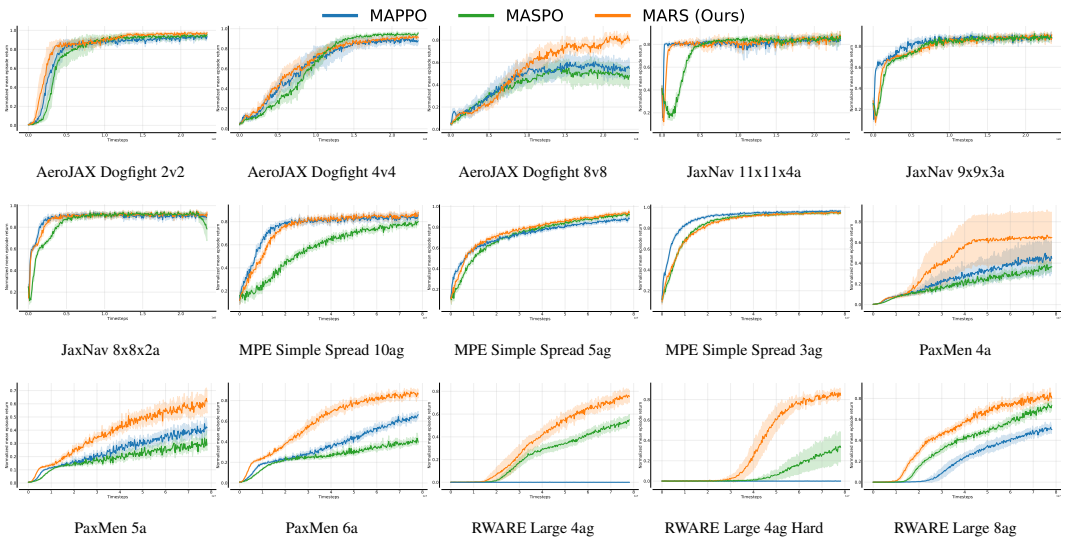
Below is pseudocode for MARS in the same CTDE training loop used by MAPPO and MASPO, replacing only the actor objective. Although the pseudocode denotes the centralized critic as state-based, history-dependent critics such as $V(h)$ and $V(h, s)$ can offer theoretical advantages (Lyu et al., 2023). However, computational constraints in large-scale tasks often necessitate state-based critics $V(s)$. Our work utilizes a variety of critic inputs, depending on what was provided by the environment and the scale of the task. Additionally, the pseudocode writes the actor objective as a loss to be minimized, so the MARS surrogate and entropy bonus appear inside a leading negative sign.

Algorithm 1 Recurrent Multi-Agent Ratio Symmetry (MARS) with CTDE

Input: Initial policy parameters θ, critic parameters ϕ
Hyperparameters: Expansion target $B_{\text{upper}} > 1$, Contraction target $B_{\text{lower}} < 1$, learning rates η_θ, η_ϕ
while training not converged **do**
  Data Collection (Decentralized Rollout):
  Collect batch $\mathcal{D}$ by running policies $\pi_\theta(a_{i,t}|h_{i,t})$ for all agents $i \in \{1, \dots, N\}$.
  Store trajectories $\tau = \{(s_t, \mathbf{h}_t, \mathbf{a}_t, r_t, \pi_{\text{old}})\}_{t=0}^{T}$.
  // Note: Actors use local histories $h_{i,t}$; Critic uses global state s_t.
  Advantage Estimation (Centralized):
  Compute joint advantages $\hat{\mathbf{A}}_t$ and value targets R_t using GAE and centralized critic $V_\phi(s_t)$.
  Optimization Loop:
  for each epoch, minibatch $b \in \mathcal{D}$ **do**
    Re-evaluate policies to get $\pi_\theta(a_{i,t}|h_{i,t})$ and ratios $r_{i,t}(\theta) = \frac{\pi_\theta(a_{i,t}|h_{i,t})}{\pi_{\text{old}}(a_{i,t}|h_{i,t})}$.
    1. Trust Region Alignment:
    $r_{\text{target}} \leftarrow \begin{cases} B_{\text{upper}} & \text{if } \hat{\mathbf{A}}_t \geq 0 \\ B_{\text{lower}} & \text{if } \hat{\mathbf{A}}_t < 0 \end{cases}$
    2. Target-aligned penalty weight and barrier:
    // Shared penalty weight derived from the joint advantage and target ratio
    $\alpha_t \leftarrow \frac{\hat{\mathbf{A}}_t}{1-(r_{\text{target}})^{-2}}$
    // Per-agent geometric barrier
    $\psi_{\text{MARS}}(r_{i,t}) \leftarrow r_{i,t} + \frac{1}{r_{i,t}} - 2$
    3. Independent Updates:
    // Actor Update (Uses History $h_{i,t}$ and Joint Advantage $\hat{\mathbf{A}}_t$)
    $\mathcal{L}_{\text{Actor}}(\theta) = -\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(r_{i,t}(\theta)\hat{\mathbf{A}}_t - \alpha_t\psi_{\text{MARS}}(r_{i,t}(\theta)) + \beta\mathcal{H}(\pi_\theta(\cdot|h_{i,t}))\right)\right]$
    $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \mathcal{L}_{\text{Actor}}$
    // Critic Update (Uses State s_t)
    $\mathcal{L}_{\text{Critic}}(\phi) = \mathbb{E}\left[(V_\phi(s_t) - R_t)^2\right]$
    $\phi \leftarrow \phi - \eta_\phi \nabla_\phi \mathcal{L}_{\text{Critic}}$
  end for
end while

G Per-Task Performance

Below we provide a learning curve breakdown at the per-task level:



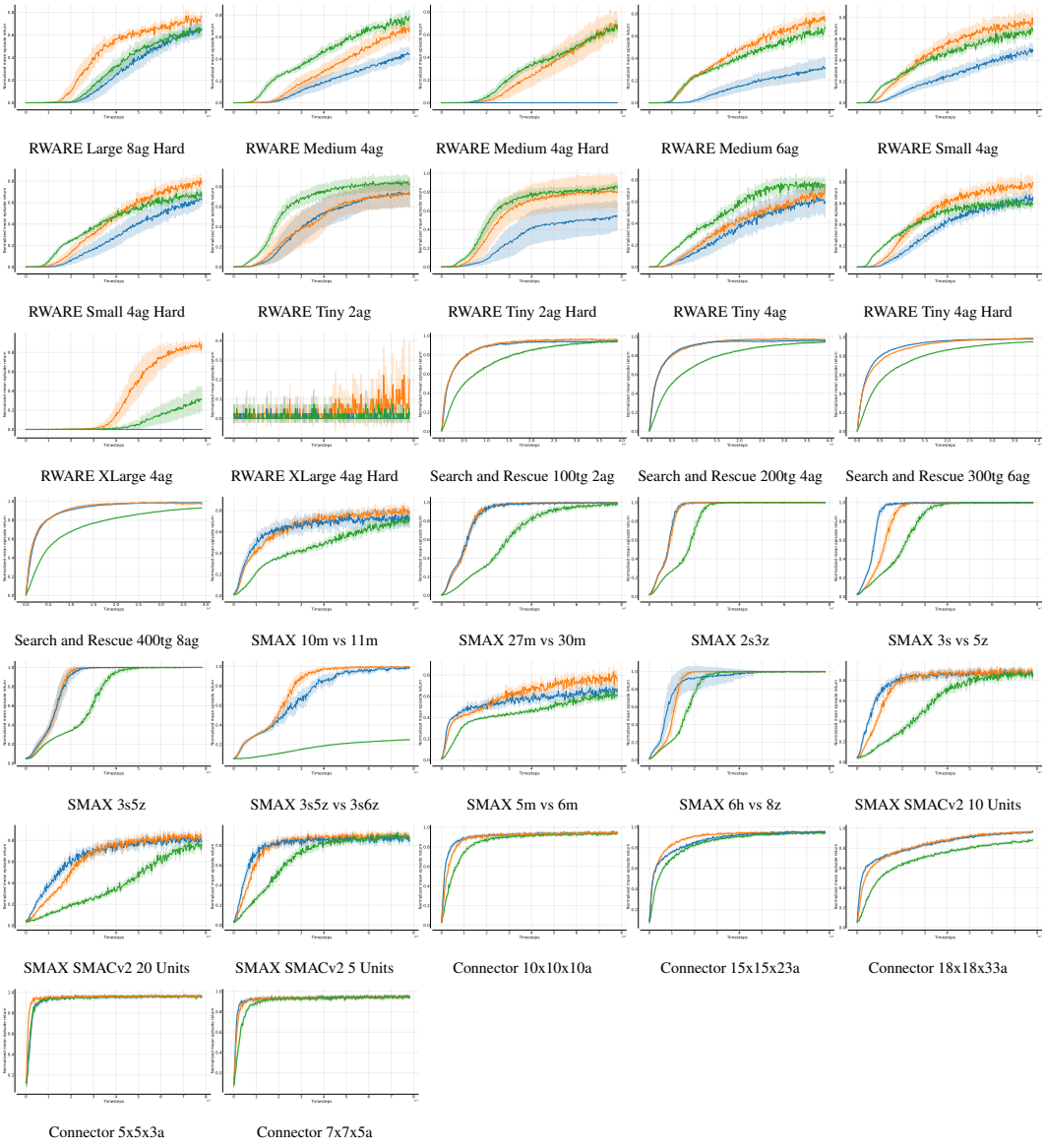


Figure 9: *Per-task learning performance.* Plots depict the mean across 10 independent random seeds (shaded regions denote 95% confidence intervals). Per-task min-max normalization is used.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's main contributions: the MARS objective, its ratio-space properties, the PaxMen and AeroJAX benchmarks, and the 47-task empirical evaluation. The claims are scoped to cooperative CTDE multi-agent policy-gradient reinforcement learning and are supported by the formal surrogate-level results in Section 3 and the experiments in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in Appendix A, including the scope of the theoretical guarantees, benchmark coverage, target-ratio tuning, and open questions for broader multi-agent settings.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper states the assumptions for each theoretical result in the proposition statements, including the positive ratio domain $r > 0$, finite joint advantage estimate $\widehat{\mathbf{A}}$, and positive penalty strength $\alpha > 0$. Full proofs are provided in Appendix C, including the MAPPO clipped-gradient derivation, the MARS penalty derivation, the extinction-barrier result, the no-gradient-truncation result, and the target-ratio alignment derivation.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the experimental setup in Section 4, including the evaluated algorithms, benchmark suite, tuning protocol, number of Optuna trials, shared policy-parameter setting, and evaluation scope. Appendix E provides the hyperparameter search spaces, fixed hyperparameters, training lengths, and evaluation protocol, while Appendix F provides pseudocode for MARS. Full code, compute details, and exact final tuned hyperparameters for all results are provided in the Supplementary Materials to assist in reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Full code, compute details, and exact final tuned hyperparameters for all results are provided in the supplementary materials to assist in reproducibility.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 summarizes the training setup, benchmark suite, baseline algorithms, tuning protocol, and evaluation design. Appendix E specifies the hyperparameter search spaces, tuning budgets, fixed hyperparameters, optimizer choice, evaluation protocol, seed usage, and training lengths needed to understand the reported results. Final tuned hyperparameters for all experiments are provided in the supplementary materials.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Section 4 reports 95% confidence intervals for aggregate performance, learning curves, and probability-of-improvement estimates. The figure and table captions specify the uncertainty estimates used, including stratified bootstrap confidence intervals for aggregate metrics and confidence intervals across independent random seeds for learning curves, following the evaluation recommendations of Agarwal et al. (2021b). Per-task learning curve results in Appendix Section G also report 95% confidence intervals.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The supplementary material reports accelerator configurations and compute records for the main experiments, and Appendix Section E reports tuning lengths, final evaluation lengths, and long-horizon ablation lengths. The full evaluation requires substantial compute because it spans 47 tasks, multiple random seeds, Optuna tuning, and billion-step ablations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research conforms to the NeurIPS Code of Ethics. The work uses simulated reinforcement-learning environments, does not involve human subjects or private data, and the supplementary code is anonymized for review.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper studies foundational cooperative MARL optimization in simulation. Potential positive impacts include more stable training for cooperative multi-agent systems. Potential negative impacts are limited to downstream misuse of improved multi-agent coordination methods, particularly in simulated navigation or combat-inspired domains. The work does not release deployed systems, pretrained agents for real-world control, or human-subject data.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release high-risk pretrained models, scraped datasets, human-subject data, or generative models requiring controlled release safeguards. The supplementary material contains simulation environments, training code, and hyperparameter records.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Existing assets are credited in the paper and supplementary archive. The archive includes license files for the modified Mava codebase and released environment packages, and a top-level notice identifies upstream projects and licenses used or adapted.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces the MARS implementation and the PaxMen and AeroJAX environment packages. The supplementary archive provides source code, package metadata, licenses, README documentation, hyperparameter records, and representative commands for running experiments with these assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing, user studies, or human-subject experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development, theory, experiments, and environments do not use LLMs as a methodological component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.