SRC: SINKHORN RESPONSIBILITY CREDIT FOR SPATIAL MULTI-AGENT REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Credit assignment in cooperative multi-agent reinforcement learning (MARL) is commonly formulated as attributing a shared team return to individual agents through value estimates, counterfactual baselines, or decomposed utilities. In spatial and object-centric tasks, however, coordination also depends on which agents should cover which landmarks, goals, resources, or targets. Shared rewards and global shaping signals can provide useful team-level feedback while leaving this assignment pattern unresolved.

We propose Sinkhorn Responsibility Credit (SRC), a transport-based credit module that models this pattern as an entropic optimal-transport plan between agents and task objects. SRC retains the plan as a soft responsibility matrix and converts it into dense agent-specific credit. This differs from global optimal-transport shaping, which uses the same geometric objective but broadcasts a single team-level scalar.

Under a fixed policy backbone, marginal SRC improves MPE simple spread returns relative to shared rewards by 5.07 to 20.03 points and relative to global optimal-transport shaping by 2.71 to 16.50 points, with larger gains as the number of agents grows. This isolates the central mechanism: the same spatial objective is more informative when its transport assignment is returned to individual agents than when it is collapsed into a global potential. On VMAS navigation, static and temporal SRC scale more reliably than leave-one-out marginal responsibility: across $N = 4, 6, 8$ and 20 matched seeds, SRC-S wins 60/60 and SRC-T wins 59/60 seed-wise comparisons against shared rewards with paired $p < 7 \times 10^{-8}$ at every team size. Difference rewards remain the strongest MPE baseline because they evaluate task-utility counterfactuals directly; SRC instead provides an assignment signal from observed agent-object geometry. Auxiliary learner checks, hybrid experiments, BenchMARL/TorchRL MAPPO runs, and stress tests further identify the regime in which SRC is most effective: tasks where observable objects and meaningful costs expose decision-relevant assignment structure.

1 INTRODUCTION

Cooperative multi-agent reinforcement learning (MARL) is frequently trained with a shared team reward. Although this signal aligns agents with a common objective, it also erases which agent was responsible for which part of the task. In spatial and object-centric domains, agents may cover different landmarks, approach different goals, or interact with different targets while receiving identical scalar feedback. This ambiguity can slow learning and encourage redundant behavior as the number of agents and task objects increases.

Most MARL credit-assignment methods focus on contribution to return. Centralized critics estimate agent-specific advantages (Foerster et al., 2018; Yu et al., 2022), value-decomposition methods factorize a joint value into per-agent utilities (Sunehag et al., 2018; Rashid et al., 2020; Wang et al., 2021a), and game-theoretic methods estimate marginal contribution through Shapley-style or nucleolus-based credit (Wang et al., 2020a; 2022; Li et al., 2025). These approaches remain central, but they do not explicitly represent the agent-object matching that often drives spatial coordination.

We formalize this matching as agent-object responsibility and introduce Sinkhorn Responsibility Credit (SRC). SRC computes an entropic optimal-transport plan between agents and task objects,

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

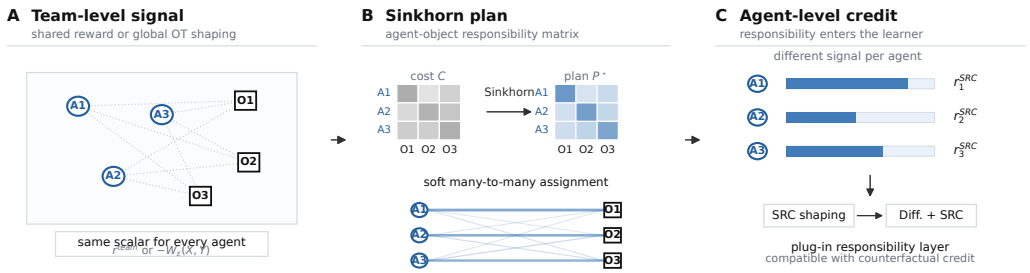

Figure 1: Main concept of Sinkhorn Responsibility Credit. Shared rewards and global geometric shaping provide team-level feedback but leave agent-object responsibility ambiguous. SRC converts spatial agent-object costs into a soft Sinkhorn responsibility plan and uses it to produce agent-specific credit. The resulting signal can be used as an independent shaping term or combined with counterfactual credit.

interprets it as a soft responsibility matrix, and converts it into agent-specific credit. Both SRC and global OT shaping use the same geometric objective; global OT reduces it to a team-level scalar, whereas SRC keeps the assignment information. Figure 1 illustrates this distinction. We therefore evaluate SRC as a plug-in responsibility layer, not as a replacement for learned critics, value decomposition, or task-utility counterfactuals. Its target setting is one in which observable objects and meaningful costs expose coordination structure that a shared scalar reward hides.

Our experiments isolate the credit signal. The main runs keep the policy backbone fixed and vary only the reward or credit construction; auxiliary A2C, DQN, and centralized-critic PPO checks then test whether the same responsibility signal can be attached beyond the primary PPO setting.

The contributions are:

- We identify agent-object responsibility as a distinct credit structure in spatial and object-centric MARL, complementing return-contribution views of credit assignment.
- We propose Sinkhorn Responsibility Credit, a modular transport-based signal that converts a Sinkhorn assignment plan into dense agent-specific rewards without changing the policy backbone or requiring a learned critic.
- We isolate responsibility decomposition by comparing SRC with shared rewards, global OT shaping, and a stronger task-utility Difference baseline under a controlled PPO pipeline.
- We provide matched-seed evidence on MPE and VMAS, and use backbone checks, hybrids, native BenchMARL/TorchRL MAPPO runs, GRF, SMACv2, and interaction-heavy stress tests to characterize the conditions under which spatial responsibility remains informative or insufficient.

2 Related Work

Credit assignment and CTDE. Credit assignment is central to cooperative MARL because shared rewards hide individual responsibility. Centralized-training decentralized-execution methods address this by using global information during learning while preserving decentralized policies. MADDPG and COMA introduced influential centralized-critic and counterfactual-advantage formulations (Lowe et al., 2017; Foerster et al., 2018); MAPPO showed that strong on-policy training remains competitive across cooperative benchmarks (Yu et al., 2022); and ICLR work such as DOP imports value decomposition into actor-critic learning to reduce centralized-decentralized mismatch and improve credit assignment (Wang et al., 2021c). Recent methods extend this line through partial reward decoupling and spatial-temporal return decomposition (Kapoor et al., 2024; Chen et al., 2024). SRC differs by constructing an explicit agent-object responsibility structure before the policy update rather than inferring responsibility only through a critic or return decomposer.

Value, role, and communication decomposition. Value decomposition methods factorize the joint action-value function into agent-level utilities. VDN, QMIX, QTRAN, and QPLEX define in-

creasingly expressive factorization classes under CTDE and IGM-style constraints (Sunehag et al., 2018; Rashid et al., 2020; Son et al., 2019; Wang et al., 2021a). ICLR work has also decomposed coordination through communication or roles: NDQ learns nearly decomposable value functions with minimized communication, and RODE discovers role action spaces from action effects (Wang et al., 2020c; 2021b). More recent mixers add attention, graph structure, locality, temporal modeling, or interpretability (Liu et al., 2023a; Zohar et al., 2022; Liu et al., 2023b; Wang et al., 2025; Hou et al., 2026). Related recent work also considers multi-level counterfactual advantages and graph-structured value decomposition (Zhao & Xie, 2025; Wang et al., 2025). These methods mainly structure values, actions, roles, subtasks, or messages; SRC structures responsibility between agents and task objects, so it can act as an additional feature, regularizer, or shaping signal for such learners.

Marginal contribution and learned reward signals. Counterfactual and game-theoretic credit methods compare observed team utility against alternatives. Difference rewards remove an agent from the team utility (Wolpert & Tumer, 2002; Agogino & Tumer, 2008); Shapley and nucleolus methods estimate fair marginal contribution over coalitions (Wang et al., 2020a; 2022; Li et al., 2025); and potential-based shaping gives invariance conditions for auxiliary rewards (Ng et al., 1999; Devlin et al., 2014). Learned-reward and implicit-credit methods such as LIIR, ICES, LICA, and influence-based exploration learn intrinsic or gradient-level signals from returns, transitions, interactions, or critics (Du et al., 2019; Li et al., 2024; Zhou et al., 2020; Wang et al., 2020b). The marginal SRC variant follows this counterfactual spirit, but changes the comparison object: it evaluates a structured transport objective over agents and objects rather than the environment return.

Spatial-temporal structure and optimal transport. Recent credit-assignment work emphasizes that contribution can be spatially and temporally structured. SAD and communication-corrected replay address how cooperative agents make actions or messages useful to teammates during centralized training or off-policy learning (Hu & Foerster, 2020; Ahilan & Dayan, 2021), while spatial-temporal redistribution methods decompose delayed return across agents and time (Xiao et al., 2022; Chen et al., 2024). Optimal transport provides the geometry used by SRC: entropic OT and Sinkhorn iterations efficiently compare spatial distributions (Cuturi, 2013; Peyré & Cuturi, 2019), and OT has been used in RL/MARL for distribution matching (Wang et al., 2023). SRC uses OT differently from global shaping: the Sinkhorn plan is retained as an agent-object responsibility matrix and then decomposed into agent-specific credit.

Benchmarks and diagnostic scope. Our experiments use benchmark families that expose different responsibility structures. MPE simple spread is a spatial coverage task from the MADDPG suite (Lowe et al., 2017); VMAS provides vectorized navigation (Bettini et al., 2022); PettingZoo and the benchmark suite of Papoudakis et al. (2021) cover Pursuit, LBF, and RWARE (Terry et al., 2021); and GRF/SMACv2 add temporally extended scoring and combat dynamics (Kurach et al., 2020; Ellis et al., 2023). We use these benchmarks to separate spatial-coverage cases from interaction-heavy settings where distance-only responsibility should fail.

Appendix A summarizes SRC relative to common credit-assignment families.

3 BACKGROUND

We consider a cooperative Markov game with N agents and a shared task objective. SRC does not assume a particular policy class or centralized critic; it assumes that, at each timestep, the environment can expose or derive a spatial abstraction of the current task state. Agent i has position $x_i(t) \in \mathbb{R}^d$, and the task contains M objects with positions $y_j(t) \in \mathbb{R}^d$. Depending on the benchmark, an object may be a landmark, goal, food item, shelf, ball, or enemy target. Static landmarks are the special case $y_j(t) = y_j$. We write

$$X_t = \{x_i(t)\}_{i=1}^{N}, \quad Y_t = \{y_j(t)\}_{j=1}^{M}. \tag{1}$$

SRC builds a cost matrix between agents and objects. In the main MPE and VMAS experiments, this cost is squared Euclidean distance,

$$C_{ij}(t) = \|x_i(t) - y_j(t)\|_2^2. \tag{2}$$

This choice matches coverage and navigation tasks, where proximity is a useful responsibility proxy; GRF, SMACv2, LBF, RWARE, and Pursuit test cases where possession, timing, loading, combat, or capture mechanics break this proxy.

Let $a_i = 1/N$ and $b_j = 1/M$ be uniform marginals. Uniform marginals give each agent and object equal transport mass, matching the symmetric coverage tasks used for the main claim. Non-uniform marginals could encode heterogeneous agent capacity or object priority, but we leave that extension outside the present work. The entropic optimal transport problem is

$$W_\epsilon(X_t, Y_t) = \min_{P \in \Pi(a,b)} \sum_{i=1}^{N} \sum_{j=1}^{M} P_{ij} C_{ij}(t) + \epsilon \sum_{i=1}^{N} \sum_{j=1}^{M} P_{ij}(\log P_{ij} - 1), \tag{3}$$

where $P \in \mathbb{R}_+^{N \times M}$ is a transport plan, $\Pi(a, b) = \{P : P\mathbf{1} = a, P^\top \mathbf{1} = b\}$, the total transported mass is one, and $0 \log 0$ is interpreted as zero. The scalar value W_ϵ can be used directly as a global OT shaping potential, but that use collapses the assignment structure. SRC instead uses the Sinkhorn plan itself as the object-responsibility structure from which agent-specific credit is computed.

The Sinkhorn algorithm computes the entropic plan efficiently for the agent and object counts used in our experiments. With K Sinkhorn iterations, static and temporal SRC require one $N \times M$ transport solve per timestep, while marginal SRC requires one full solve plus N leave-one-out solves. This extra cost is modest in MPE and VMAS, but it is a practical reason to prefer static or temporal SRC when the agent count is large.

4 Transport-Based Responsibility Assignment

SRC maps a spatial state abstraction into an agent-indexed credit signal. The computation has four steps: construct an agent-object cost matrix, solve an entropic OT problem, interpret the plan as a responsibility distribution, and inject the resulting scalar credit into an otherwise unchanged learner. Figure 2 gives a schematic view. The construction below describes the reward-shaping implementation used in the experiments; advantage shaping and critic-side auxiliary targets are left for future work.

Responsibility matrix. A global OT shaping baseline uses only the scalar transport value, for example through the team-level potential $\Phi_t = -W_\epsilon(X_t, Y_t)$. This tests whether the geometry itself is useful, but it broadcasts the same scalar to all agents and discards the assignment plan. SRC differs at this point: it keeps the Sinkhorn solution $P_t^\star$ and converts each row into a conditional responsibility distribution,

$$\rho_{ij}(t) = \frac{P_{ij}^\star(t)}{\sum_{k=1}^{M} P_{ik}^\star(t)}, \tag{4}$$

For an exact feasible plan with uniform marginals, $\sum_k P_{ik}^\star = a_i = 1/N$, so Eq. 4 rescales the transported row mass into a distribution over objects. In implementation, the denominator is clipped by a small constant only to avoid division by zero in degenerate numerical states. Thus $\rho_{ij}(t)$ should be read as the fraction of agent i's responsibility assigned to object j, not as a probability that agent i caused the team reward. The assignment is soft: an agent can split responsibility across objects, and an object can receive responsibility from multiple agents through the column constraints of the OT problem.

Static SRC. The static variant converts responsibility into the negative expected assignment cost,

$$d_i(t) = \sum_{j=1}^{M} \rho_{ij}(t) C_{ij}(t), \qquad s_i^{\mathrm{stat}}(t) = -d_i(t). \tag{5}$$

This signal gives higher credit to agents that are close to the objects assigned to them by the Sinkhorn plan. It uses assignment structure but is purely state-based: it does not distinguish an agent that is moving toward its responsibility from one that is already close or one that has just been reassigned by the transport plan.

Temporal SRC. The temporal variant adds a local progress term to the static signal. Using the same responsibility-weighted distance $d_i(t)$ from Eq. 5, define

$$\Delta_i^{\text{temp}}(t) = d_i(t-1) - d_i(t), \tag{6}$$

with $\Delta_i^{\text{temp}}(t) = 0$ at the first timestep of an episode. The temporal SRC signal used in the experiments is

$$s_i^{\text{temp}}(t) = s_i^{\text{stat}}(t) + \alpha \Delta_i^{\text{temp}}(t). \tag{7}$$

Positive temporal credit means that the responsibility-weighted distance decreased from the previous step. This is a progress-shaped estimator rather than a monotonic individual-improvement guarantee. Because ρ_t is recomputed at every timestep, changes in $d_i(t)$ can reflect both physical movement and team-level reassignment. This caveat matters in environments with rapidly changing task objects; empirically, temporal SRC is most reliable in navigation-like settings where assignment changes are smooth.

Marginal transport SRC. The marginal variant measures the state-level contribution of agent i to the transport objective. Let $X_{-i,t}$ be the agent set with agent i removed. The leave-one-out OT problem uses total mass one with row marginals $a_{-i,k} = 1/(N-1)$ over the remaining agents and the same object marginals b; we write this re-normalized objective as $W_\epsilon(X_{-i,t}, Y_t)$ for compactness. We define

$$s_i^{\text{marg}}(t) = W_\epsilon(X_{-i,t}, Y_t) - W_\epsilon(X_t, Y_t). \tag{8}$$

The sign convention makes the signal positive when removing agent i increases the transport objective, i.e., when the agent helps match the team to the object set. This is not a Shapley value and not an environment counterfactual return: it is a leave-one-out comparison of two regularized transport problems. Because the removal changes both the feasible polytope and the entropy term, s_i^{marg} contains a regularization-dependent offset as well as geometric contribution. For this reason, we compare it directly with static and temporal SRC instead of treating it as a theoretically privileged estimator.

Interaction-aware diagnostic extension. Some benchmarks require feasibility information that Euclidean distance does not encode. Loading food, moving requested shelves, surrounding evaders, and targeting enemies depend on action availability, object state, health, or group constraints. To test this boundary, we also evaluate hand-designed interaction-aware utilities of the form

$$U^{\text{IA}}(X_t, Y_t, z_t) = -\omega_{\text{ot}} W_\epsilon(X_t, Y_t) + \omega_{\text{ia}} F(X_t, Y_t, z_t), \tag{9}$$

where z_t contains domain-specific interaction state and F is a feasibility score. For LBF, RWARE, and Pursuit, the diagnostic credit is the leave-one-out utility difference $U^{\text{IA}}(X_t, Y_t, z_t) - U^{\text{IA}}(X_{-i,t}, Y_t, z_{-i,t})$. For SMACv2, where removing a unit is not the diagnostic used by the wrapper, we instead combine the Sinkhorn responsibility matrix with combat readouts such as range, health, damage, and kills. These variants are stress tests for distance-only responsibility, not a second proposed algorithm; Appendix Table 14 lists the concrete utilities.

Training integration. In the main experiments, the selected SRC signal is added to the environment reward observed by each agent,

$$\tilde{r}_i(t) = r_i^{\text{env}}(t) + \lambda s_i(t), \qquad s_i(t) \in \{s_i^{\text{stat}}(t),\ s_i^{\text{temp}}(t),\ s_i^{\text{marg}}(t)\}. \tag{10}$$

The coefficient λ controls the shaping strength and is fixed within each benchmark sweep. We denote the static, temporal, and marginal variants as SRC-S, SRC-T, and SRC-M. This integration leaves the optimizer, policy architecture, and observation interface unchanged, so differences primarily reflect the reward or credit signal. Unless explicitly labeled as shaped diagnostics, evaluation uses original task metrics. Advantage shaping, critic-side auxiliary targets, and responsibility priors for sampled Shapley-style credit are left for future work (Wang et al., 2020a; 2022).

5 Experiments

Our evaluation asks when exposing the agent-object assignment improves learning beyond a shared spatial objective. We isolate the credit signal with a fixed PPO backbone unless stated otherwise, and

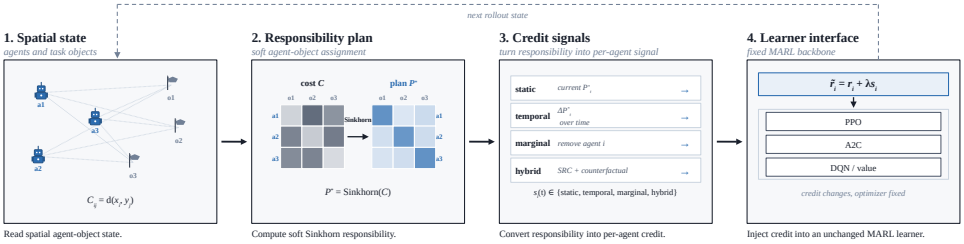

Figure 2: Detailed SRC computation pipeline. The four modules show what each stage represents: observable spatial state, a Sinkhorn responsibility plan over agents and objects, variant-specific credit signals, and the learner interface where the credit is added to a fixed MARL backbone. The diagram is schematic; the exact reward integration is given in the equations and Algorithm 1.

report original task-level metrics whenever available; shaped returns appear only in labeled diagnostics. MPE tests the mechanism directly, VMAS tests estimator choice at scale, auxiliary learner and hybrid runs test composability, and LBF/RWARE/Pursuit/GRF/SMACv2 serve as diagnostic tests for cases where distance-based responsibility is incomplete.

Benchmarks. MPE simple spread is the primary mechanism benchmark because agents must cover landmarks in a spatial domain, matching the SRC assumption that task objects and a meaningful cost metric define responsibility (Lowe et al., 2017). We evaluate $N = 3, 5, 7$ with 20 matched seeds. VMAS navigation is the main scalability benchmark: a matched $N = 4, 6, 8$ sweep isolates how static, temporal, and marginal responsibility estimators scale in a vectorized navigation setting (Bettini et al., 2022). All VMAS credit modes use 20 matched seeds, so estimator comparisons use the same seed budget. LBF, RWARE, Pursuit, and SMACv2 are stress tests for sparse, interaction-heavy, or combat-oriented settings where geometry alone may not determine task success (Terry et al., 2021; Papoudakis et al., 2021; Ellis et al., 2023). GRF is a targeted diagnostic on `academy_3_vs_1_with_keeper`, used to test whether a spatial signal can accelerate learning in a temporally extended scoring task (Kurach et al., 2020).

Credit signals and baselines. The shared baseline gives every agent the environment team reward. Global OT uses the same transport objective as SRC but broadcasts the scalar shaping signal to all agents, isolating whether transport geometry alone is sufficient. Difference rewards estimate each agent's marginal contribution to the original task utility and serve as a strong task-specific counterfactual baseline. In MPE and VMAS, Difference is computed analytically from the known coverage-style utility by evaluating the same state after removing one agent; it therefore uses privileged task-utility information and a removal oracle that SRC does not assume. SRC uses only the observed agent-object cost matrix and the Sinkhorn plan. The primary MPE comparison reports marginal SRC, while VMAS evaluates SRC as a family and compares static, temporal, and marginal readouts.

Metrics and statistics. The primary metric is original evaluation return; for MPE we also report final coverage cost, and for VMAS/GRF success rate. Unless stated otherwise, values written as mean $\pm$ uncertainty use 95% confidence intervals rather than standard deviations. GRF and SMACv2 include shaped diagnostics because sparse objectives can mask whether the spatial credit signal is being optimized, but these diagnostics are not treated as task success unless original metrics also improve. The primary MPE statistical family compares SRC against shared reward and global OT across $N = 3, 5, 7$ with Holm correction. VMAS paired tests compare each estimator against shared reward over 20 matched seeds and are reported as the main scalability analysis. Difference comparisons mark the gap to a strong task-specific counterfactual baseline. Auxiliary p-values are descriptive and uncorrected unless stated otherwise. Positive benefit favors SRC; for coverage cost and time-to-threshold, signs are reversed. Appendix C reports confidence intervals, effect sizes, Wilcoxon-Holm checks, and win counts.

Appendix A gives the full evidence map and placement of auxiliary results.

Table 1: Original evaluation return on MPE simple spread with 20 matched seeds. Values are means; paired uncertainty and significance are reported in the text, Figure 3, and Appendix C. Higher is better.

MPE	Shared	Global OT	SRC-M	Difference
$N = 3$	-25.53	-23.16	-20.45	$\mathbf{-19.54}$
$N = 5$	-46.66	-43.48	-32.45	$\mathbf{-30.99}$
$N = 7$	-64.37	-60.83	-44.33	$\mathbf{-42.70}$

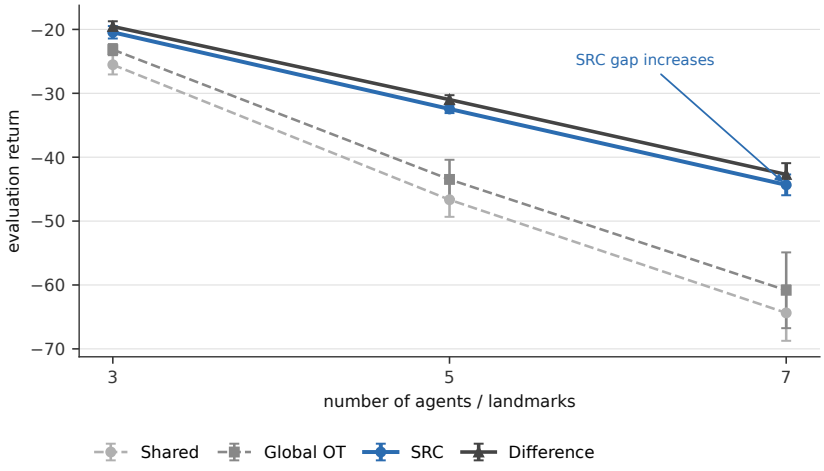

Figure 3: MPE scaling result with 20 seeds. SRC improves over shared reward and global OT shaping at all tested team sizes, and the gap grows as the number of agents and landmarks increases. Error bars show 95% confidence intervals.

6 Results and Analysis

We first present the MPE mechanism result, then the VMAS scalability result, followed by learner-dependence and stress-test analyses.

MPE simple spread. Figure 3 and Table 1 show the main result. SRC-M improves original evaluation return over shared rewards by 5.07, 14.22, and 20.03 points for $N = 3, 5, 7$, respectively. The corresponding 95% confidence intervals for the paired benefit are $[3.96, 6.21]$, $[11.79, 16.97]$, and $[15.96, 24.33]$, with Holm-corrected paired t-test p-values of 3.53×10^{-7}, 2.14×10^{-8}, and 1.75×10^{-7}. SRC-M wins all 20 matched seeds in each team size.

The comparison against global OT isolates the decomposition step. Because both methods use the same transport geometry, the gap reflects whether the matching is kept agent-indexed or collapsed into one scalar. SRC-M improves over global OT by 2.71, 11.04, and 16.50 return points for $N = 3, 5, 7$, with paired confidence intervals $[2.08, 3.39]$, $[7.91, 14.50]$, and $[10.24, 22.91]$. These differences remain significant after Holm correction (1.20×10^{-6}, 2.08×10^{-5}, and 3.28×10^{-4}), with 20/20, 19/20, and 18/20 seed-wise wins.

Difference rewards remain the strongest standalone MPE baseline: SRC-M trails by 0.92 and 1.46 points at $N = 3$ and $N = 5$ with corrected significance, and by 1.64 at $N = 7$ without significance. SRC is therefore complementary to well-aligned task counterfactuals, not a replacement for them.

Coverage cost. Coverage cost gives the same conclusion. With signs reversed so positive benefit favors SRC, SRC-M improves over shared reward by 0.31, 0.59, and 0.63 cost units for $N = 3, 5, 7$, and over global OT by 0.16, 0.45, and 0.48; all six comparisons are significant after Holm correction with at least 18/20 wins. Difference remains better than SRC-M at $N = 3, 5$ and statistically tied at $N = 7$.

Table 2: VMAS navigation scale-up with PPO and 200k environment steps. Values are original evaluation return over 20 matched seeds, mean $\pm$ 95% CI; higher is better.

Team size	Shared	Global OT	SRC-M	SRC-S	SRC-T
$N = 4$	2.67 ± 0.25	2.65 ± 0.31	2.97 ± 0.17	$\mathbf{3.77 \pm 0.10}$	3.60 ± 0.14
$N = 6$	0.96 ± 0.51	0.02 ± 1.18	0.68 ± 2.29	5.08 ± 0.12	$\mathbf{5.13 \pm 0.16}$
$N = 8$	0.27 ± 1.55	1.20 ± 2.28	2.46 ± 1.27	7.64 ± 0.23	$\mathbf{7.68 \pm 0.33}$

VMAS navigation. VMAS is the main scale-up result. It shifts the interpretation away from treating marginal SRC as the default estimator and toward a systems-level conclusion: as team size grows in vectorized navigation, static and temporal responsibility readouts provide cheaper and more stable credit than leave-one-out marginal responsibility. Table 2 reports original evaluation return over 20 matched seeds. SRC-S and SRC-T improve over shared rewards at every tested team size, with paired t-test $p < 7 \times 10^{-8}$ throughout. SRC-S wins 20/20 seeds at all three sizes; SRC-T wins 19/20 at $N = 4$ and 20/20 at $N = 6, 8$. SRC-M remains positive at $N = 4, 8$, but the 20-seed tests are only marginally significant or non-significant and the $N = 6$ mean is lower than shared reward. Thus VMAS supports the plug-in responsibility-layer view: the useful estimator can be selected without changing the learner.

Backbone robustness. Auxiliary MPE $N = 7$ checks with A2C, DQN, and a lightweight centralized-critic PPO implementation (Appendix D) indicate that SRC is not tied exclusively to PPO. Under A2C, SRC-M improves over shared reward by 3.98 points ($p = 6.52 \times 10^{-4}$) and over global OT by 6.31 ($p = 4.11 \times 10^{-3}$), while matching Difference within statistical noise. Under DQN, SRC-M significantly improves over global OT by 1.44 points ($p = 2.10 \times 10^{-3}$), but its gains over shared reward and Difference are not significant. The centralized-critic PPO check improves over shared reward and Difference but not over global OT. These runs check learner dependence rather than rank optimizers or full CTDE systems.

Stress-test environments. LBF, RWARE, Pursuit, and SMACv2 test settings where distance-based responsibility should be incomplete. In 10-seed LBF/RWARE/Pursuit sweeps, SRC-M does not improve the original metric over shared reward: LBF 8x8 favors shared reward (0.20 versus 0.00), LBF 10x10 and RWARE remain near zero, and Pursuit favors shared reward (-15.46 versus -17.81; higher is better). On SMACv2 `10gen_terran` with 5 allied units against 6 enemies, geometry-only SRC-M significantly underperforms shared reward (0.13 ± 0.14 versus 1.03 ± 0.32, $p = 5.39 \times 10^{-4}$). Combat-aware variants recover much of this loss but do not significantly outperform shared reward, confirming that proximity alone misses loading, logistics, capture, health, range, and damage events.

Native BenchMARL/TorchRL diagnostic. To test framework dependence, we implemented a BenchMARL/TorchRL VMAS reward adapter and evaluated MAPPO on a harder $N = 8$ navigation setting for 100k frames. This experiment tests whether the reward-level SRC interface transfers to a modern CTDE stack, while recognizing that a tuned centralized-critic learner may already encode much of the relevant structure. Over 20 matched seeds, native MAPPO obtains final unshaped return 0.2803 ± 0.0038 and potential-shaped SRC obtains 0.2785 ± 0.0034; the paired difference is -0.0017 ± 0.0020 with 7/20 wins ($p = 0.388$). Success is near ceiling for both methods (0.973 versus 0.976), and learning-curve AUC is also statistically indistinguishable. The result supports portability of the interface and identifies a limiting case: reward-level geometric shaping adds little when a strong centralized critic already solves a near-saturated navigation task.

GRF diagnostic. GRF tests whether spatial credit can accelerate a temporally extended scoring task. On `academy_3_vs_1_with_keeper` with weak goal-progress shaping shared by all modes, SRC-M with ball-and-goal targets improves the shaped learning signal: final shaped return rises from 0.0003 to 0.393 ($p = 2.00 \times 10^{-5}$), shaped-return AUC improves by 0.424 with 20/20 wins ($p = 1.39 \times 10^{-8}$), and reaching shaped return 0.1 is about 508k steps faster ($p = 1.32 \times 10^{-12}$). The effect does not transfer cleanly to sparse scoring metrics: original success rises from 0.66 to 0.82 but is not significant ($p = 0.159$), and original return does not improve. We therefore use GRF

as evidence that spatial responsibility can shape intermediate learning progress, while sparse task success may require a richer soccer-specific credit model.

7 Ablation Studies

The ablations address three questions. First, the MPE gap against global OT attributes the gain to returning the Sinkhorn assignment to individual agents. Second, the best scalar readout depends on the domain: marginal SRC is strongest at MPE $N = 3$, whereas static and temporal SRC are stronger at larger MPE sizes and in VMAS. Third, Difference+SRC hybrids are compositional but domain-dependent: the focused MPE $N = 7$ rerun improves original return by 1.43 points with 15/20 seed-wise wins ($p = 0.026$) and lowers final coverage cost, whereas the VMAS $N = 8$ hybrid is positive in mean but not significant on original return. Appendix Table 3 reports the full numbers.

Interaction-aware SRC variants clarify distance-only failures but do not solve them: RWARE remains at zero return, Pursuit only narrows the gap, and SMACv2 combat-aware utilities recover part of the geometry-only failure without significantly outperforming shared reward. These ablations refine the main interpretation: SRC is useful when the chosen cost captures responsibility, whereas tasks that are not reducible to proximity require richer feasibility models or phase-aware costs.

8 Discussion and Limitations

Taken together, the results identify the regime in which SRC is most effective. In coverage-like MARL tasks, preserving the spatial assignment is more informative than broadcasting the same geometric potential to every agent. This explains the consistent MPE gains over shared reward and global OT, while also explaining why Difference rewards remain competitive: task-utility counterfactuals use stronger supervision than SRC's geometry-only responsibility signal. SRC is therefore best understood as a complementary responsibility layer rather than a substitute for counterfactual or critic-based credit.

VMAS extends the mechanism from MPE coverage to vectorized navigation and shows that larger spatial teams favor static or temporal readouts over leave-one-out marginal solves. The auxiliary learner checks assess learner dependence: A2C, DQN, and centralized-critic PPO reduce PPO-specific concerns, but their mixed outcomes show that SRC should not be interpreted as a universal optimizer improvement. The native BenchMARL/TorchRL MAPPO diagnostic provides a complementary check: the adapter works in a modern CTDE framework, yet potential-shaped SRC is statistically indistinguishable from native MAPPO on $N = 8$ VMAS. This is consistent with a setting in which a centralized critic already captures most responsibility-relevant structure and the task is near success saturation.

The remaining diagnostics specify where this assumption breaks. Difference+SRC hybrids are positive in mean but not robustly significant on original return. LBF, RWARE, Pursuit, GRF, and SMACv2 show that distance alone omits loading constraints, logistics, capture rules, combat range, health, possession, and timing. SRC therefore requires observable task objects and a meaningful cost metric; interaction-heavy domains likely require learned feasibility models, phase-aware costs, or integration inside the critic rather than reward-level Euclidean shaping alone.

9 Conclusion

Sinkhorn Responsibility Credit converts spatial agent-object structure into agent-specific credit by preserving the assignment information that global OT shaping collapses into a team-level scalar. With a fixed PPO backbone, SRC-M improves over shared reward and global OT on MPE simple spread, while VMAS shows that static and temporal SRC scale reliably in vectorized navigation. The broader experimental suite further shows how learner choice, task alignment, framework interface, and interaction mechanics govern the value of this responsibility signal. SRC is therefore a reusable credit primitive for object-centric tasks with meaningful costs, and a starting point for richer responsibility models in domains where geometric proximity is not sufficient.

AI USE STATEMENT

Generative AI tools were used to assist with translation, English writing, language polishing, related-work discovery, LaTeX editing, drafting support, and initial figure/schematic drafts. The authors reviewed and are responsible for all text, equations, claims, code, analyses, and reported experimental results.

ETHICS STATEMENT

This work studies credit assignment algorithms in simulated cooperative multi-agent reinforcement learning environments. We do not use human-subject data, private data, or deployed autonomous systems. The main ethical risks are indirect: improved multi-agent coordination can be used in both beneficial and harmful automation contexts. We therefore report limitations and failure cases, and we avoid claims of general deployment readiness.

REPRODUCIBILITY STATEMENT

We report the credit modes, benchmarks, seeds, and evaluation metrics used in the experiments. The repository contains training scripts, resume-capable sweeps, aggregation utilities, tests, and compact CSV summaries for the reported analyses. An anonymized code package will be submitted with the paper, and the full public repository will be released after the review anonymity period. Main MPE comparisons use 20 matched seeds. VMAS comparisons use 20 matched seeds. Auxiliary backbone comparisons use matched 20-seed sweeps where reported; focused MPE/VMAS Difference+SRC reruns use 20 matched seeds and are auxiliary. BenchMARL/TorchRL MAPPO diagnostics use 20 matched seeds. LBF, RWARE, Pursuit, and SMACv2 use 10-seed scope aggregates; GRF uses 20 matched seeds. Calibration probes, smoke tests, and interrupted runs are excluded. Appendix B gives scripts, hyperparameters, and execution patterns.

REFERENCES

Adrian K. Agogino and Kagan Tumer. Analyzing and visualizing multiagent rewards in dynamic and stochastic domains. *Autonomous Agents and Multi-Agent Systems*, 17(2):320–338, 2008. doi: 10.1007/s10458-008-9046-9.

Sanjeevan Ahilan and Peter Dayan. Correcting experience replay for multi-agent communication. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=xvxPuCkCNPO.

Matteo Bettini, Ryan Kortvelesy, Jan Blumenkamp, and Amanda Prorok. VMAS: A vectorized multi-agent simulator for collective robot learning. In *Distributed Autonomous Robotic Systems*, pp. 42–56, 2022. doi: 10.1007/978-3-031-51497-5_4. URL https://doi.org/10.1007/978-3-031-51497-5_4.

Sirui Chen, Zhaowei Zhang, Yaodong Yang, and Yali Du. STAS: Spatial-temporal return decomposition for solving sparse rewards problems in multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17337–17345, 2024. doi: 10.1609/aaai.v38i16.29681. URL https://ojs.aaai.org/index.php/AAAI/article/view/29681.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Sam Devlin, Logan Yliniemi, Daniel Kudenko, and Kagan Tumer. Potential-based difference rewards for multiagent reinforcement learning. In *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems*, pp. 165–172. IFAAMAS, 2014.

Yali Du, Lei Han, Meng Fang, Ji Liu, Tianhong Dai, and Dacheng Tao. LIIR: Learning individual intrinsic reward in multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 32, pp. 4405–4416, 2019.

Benjamin Ellis, Jonathan Cook, Skander Moalla, Mikayel Samvelyan, Mingfei Sun, Anuj Mahajan, Jakob Nicolaus Foerster, and Shimon Whiteson. SMACv2: An improved benchmark for cooperative multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://openreview.net/forum?id=5OjLGiJW3u.

Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 2974–2982. AAAI Press, 2018. doi: 10.1609/aaai.v32i1.11794. URL https://ojs.aaai.org/index.php/AAAI/article/view/11794.

Jie Hou, Haowen Dou, Lujuan Dang, Liangjun Chen, and Chenyang Ge. Gradient-protected value decomposition for cooperative multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21779–21787, 2026. doi: 10.1609/aaai.v40i26.39329. URL https://ojs.aaai.org/index.php/AAAI/article/view/39329.

Hengyuan Hu and Jakob N. Foerster. Simplified action decoder for deep multi-agent reinforcement learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1xm3RVtwB.

Aditya Kapoor, Benjamin Freed, Jeff Schneider, and Howie Choset. Assigning credit with partial reward decoupling in multi-agent proximal policy optimization. *Reinforcement Learning Journal*, 1:380–399, 2024. URL https://rlj.cs.umass.edu/2024/papers/Paper45.html.

Karol Kurach, Anton Raichuk, Piotr Stanczyk, Michal Zajac, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, and Sylvain Gelly. Google research football: A novel reinforcement learning environment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 4501–4510, 2020. doi: 10.1609/aaai.v34i04.5878. URL https://ojs.aaai.org/index.php/AAAI/article/view/5878.

Xinran Li, Zifan Liu, Shibo Chen, and Jun Zhang. Individual contributions as intrinsic exploration scaffolds for multi-agent reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28387–28402. PMLR, 2024. URL https://proceedings.mlr.press/v235/li24aw.html.

Yugu Li, Zehong Cao, Jianglin Qiao, and Siyi Hu. Nucleolus credit assignment for effective coalitions in multi-agent reinforcement learning. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pp. 1318–1326. International Foundation for Autonomous Agents and Multiagent Systems, 2025. URL https://www.ifaamas.org/Proceedings/aamas2025/pdfs/p1318.pdf.

Shunyu Liu, Yihe Zhou, Jie Song, Tongya Zheng, Kaixuan Chen, Tongtian Zhu, Zunlei Feng, and Mingli Song. Contrastive identity-aware learning for multi-agent value decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11595–11603, 2023a. doi: 10.1609/aaai.v37i10.26370. URL https://ojs.aaai.org/index.php/AAAI/article/view/26370.

Zichuan Liu, Yuanyang Zhu, and Chunlin Chen. NA2Q: Neural attention additive model for interpretable multi-agent q-learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22539–22558. PMLR, 2023b. URL https://proceedings.mlr.press/v202/liu23be.html.

Ryan Lowe, Yi I. Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6379–6390, 2017.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 278–287. Morgan Kaufmann, 1999.

Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://openreview.net/forum?id=cIrPX-Sn5n.

Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*. Now Publishers, 2019. doi: 10.1561/2200000073.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL https://jmlr.org/papers/v22/20-1364.html.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51, 2020. URL https://jmlr.org/papers/v21/20-081.html.

Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5887–5896. PMLR, 2019. URL https://proceedings.mlr.press/v97/son19a.html.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2085–2087. International Foundation for Autonomous Agents and Multiagent Systems, 2018. URL https://www.ifaamas.org/Proceedings/aamas2018/.

J. K. Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis S. Santos, Clemens Dieffendahl, Caroline Horsch, Rodrigo Perez-Vicente, Niall L. Williams, Yashas Lokesh, and Praveen Ravi. Pettingzoo: Gym for multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15032–15043, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/803f7c4c3ff61b71be53a0c803bfb57f-Abstract.html.

Caroline Wang, Ishan Durugkar, Elad Liebman, and Peter Stone. DM^2: Decentralized multi-agent reinforcement learning via distribution matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11699–11707, 2023. doi: 10.1609/aaai.v37i10.26382. URL https://ojs.aaai.org/index.php/AAAI/article/view/26382.

Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. QPLEX: Duplex dueling multi-agent q-learning. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=Rcmk0xxIQV.

Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, and Yunjie Gu. Shapley q-value: A local reward approach to solve global reward games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7285–7292, 2020a. doi: 10.1609/aaai.v34i05.6220. URL https://ojs.aaai.org/index.php/AAAI/article/view/6220.

Jianhong Wang, Yuan Zhang, Yunjie Gu, and Tae-Kyun Kim. SHAQ: Incorporating shapley value theory into multi-agent q-learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Shudong Wang, Wenhao Ji, Haiyuan Gui, Kuijie Zhang, Luqi Wang, and Shanchen Pang. XMIX: Graph-based temporal credit assignment and attention-augmented value decomposition for multi-agent cooperative reinforcement learning. *Neurocomputing*, 655:131471, 2025. doi: 10.1016/j.neucom.2025.131471. URL https://doi.org/10.1016/j.neucom.2025.131471.

Tonghan Wang, Jianhao Wang, Yi Wu, and Chongjie Zhang. Influence-based multi-agent exploration. In *International Conference on Learning Representations*, 2020b. URL https://iclr.cc/virtual_2020/poster_BJgy96EYvr.html.

Tonghan Wang, Jianhao Wang, Chongyi Zheng, and Chongjie Zhang. Learning nearly decomposable value functions via communication minimization. In *International Conference on Learning Representations*, 2020c. URL https://iclr.cc/virtual_2020/poster_HJx-3grYDB.html.

Tonghan Wang, Tarun Gupta, Anuj Mahajan, Bei Peng, Shimon Whiteson, and Chongjie Zhang. RODE: Learning roles to decompose multi-agent tasks. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=TTUVg6vkNjK.

Yihan Wang, Beining Han, Tonghan Wang, Heng Dong, and Chongjie Zhang. DOP: Off-policy multi-agent decomposed policy gradients. In *International Conference on Learning Representations*, 2021c. URL https://openreview.net/forum?id=6FqKiVAdI3Y.

David H. Wolpert and Kagan Tumer. Optimal payoff functions for members of collectives. In *Modeling Complexity in Economic and Social Systems*, pp. 355–369. World Scientific, 2002.

Baicen Xiao, Bhaskar Ramasubramanian, and Radha Poovendran. Agent-temporal attention for reward redistribution in episodic multi-agent reinforcement learning. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 1391–1399. International Foundation for Autonomous Agents and Multiagent Systems, 2022. URL https://www.ifaamas.org/Proceedings/aamas2022/.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre M. Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24611–24624, 2022. doi: 10.52202/068431-1787. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9c1535a02f0ce079433344e14d910597-Abstract.html.

Xutong Zhao and Yaqi Xie. Multi-level advantage credit assignment for cooperative multi-agent reinforcement learning. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2971–2979. PMLR, 2025. URL https://proceedings.mlr.press/v258/zhao25c.html.

Meng Zhou, Ziyu Liu, Pengwei Sui, Yixuan Li, and Yuk Ying Chung. Learning implicit credit assignment for cooperative multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.nips.cc/paper/2020/hash/8977ecbb8cb82d77fb091c7a7f186163-Abstract.html.

Roy Zohar, Shie Mannor, and Guy Tennenholtz. Locality matters: A scalable value decomposition approach for cooperative multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9278–9285, 2022. doi: 10.1609/aaai.v36i8.20915. URL https://ojs.aaai.org/index.php/AAAI/article/view/20915.

A Additional Tables, Figures, and Algorithm

This appendix collects larger tables, schematic figures, and algorithmic details that support the main-text argument but are not required to follow the core result.

Algorithm 1 Sinkhorn Responsibility Credit

Observe agent states X_t, task objects Y_t, and environment reward $r_i^{\mathrm{env}}(t)$
Build cost matrix $C_{ij}(t)$ and compute $P_t^\star$, $W_\epsilon(X_t, Y_t)$
Row-normalize $P_t^\star$ to obtain responsibility matrix ρ_t using Eq. 4
for each agent i **do**
  Compute $d_i(t) = \sum_j \rho_{ij}(t) C_{ij}(t)$ and $s_i^{\mathrm{stat}}(t) = -d_i(t)$
  if using static SRC **then**
    Set $s_i(t) = s_i^{\mathrm{stat}}(t)$
  else if using temporal SRC **then**
    Set $\Delta_i^{\mathrm{temp}}(t) = 0$ if t is the first episode step, else $d_i(t-1) - d_i(t)$
    Set $s_i(t) = s_i^{\mathrm{stat}}(t) + \alpha \Delta_i^{\mathrm{temp}}(t)$
  else if using marginal SRC **then**
    Remove agent i to form $X_{-i,t}$ and compute $W_\epsilon(X_{-i,t}, Y_t)$ with re-normalized row marginals
    Set $s_i(t) = W_\epsilon(X_{-i,t}, Y_t) - W_\epsilon(X_t, Y_t)$
  end if
  Form shaped reward $\tilde{r}_i(t) = r_i^{\mathrm{env}}(t) + \lambda s_i(t)$
end for
Update the fixed MARL learner using $\tilde{r}_i(t)$

Table 3: Ablation results. Standalone SRC rows report original evaluation return for the reported seed set; higher is better. Hybrid rows are auxiliary because they test whether SRC can be composed with a task-specific Difference reward, not whether SRC alone beats Difference. The archived MPE hybrid sweep is shown for context, while the focused MPE $N = 7$ and VMAS $N = 8$ hybrid reruns are the stricter matched checks discussed in the text.

Setting	SRC-S	SRC-T	SRC-M	Diff.	Diff.+S	Diff.+T	Diff.+M
MPE $N = 3$	-20.73	-20.72	-20.45	$\mathbf{-19.54}$	-20.61	-20.78	-20.20
MPE $N = 5$	-30.83	-31.48	-32.45	-30.99	-30.31	-30.61	$\mathbf{-30.11}$
MPE $N = 7$	-41.26	-40.98	-44.33	-42.70	-39.54	-40.26	$\mathbf{-38.67}$
MPE $N = 7$ rerun	–	–	–	-40.95	–	–	$\mathbf{-39.52}$
VMAS $N = 8$ hybrid	–	–	–	5.34	–	–	**5.45**
VMAS $N = 8$	7.64	**7.68**	2.46	5.34	–	–	–

Table 4: Position of SRC relative to common MARL credit-assignment families. Existing methods primarily explain contribution to return, value, or action choice. SRC instead exposes a spatial agent-object responsibility plan that can be used alone or combined with these methods.

Family	Representative methods	Credit object	Relation to SRC
Centralized critics and counterfactual advantages	MADDPG, COMA, MAPPO, PRD-MAPPO	Agent action or advantage under a learned critic	Learns contribution from value estimates; SRC supplies an explicit agent-object responsibility signal before the policy update.
Value decomposition	VDN, QMIX, QTRAN, QPLEX and recent mixers	Factorization of joint action value into agent utilities	Decomposes value; SRC decomposes spatial task structure and can be used as a responsibility feature or regularizer.
Marginal contribution and game-theoretic credit	Difference rewards, Shapley-Q, SHAQ, nucleolus credit	Agent or coalition contribution to team utility	Compares agents or coalitions; SRC compares responsibility for task objects through transport geometry.
Learned intrinsic rewards and return redistribution	LIIR, ICES, AREL, STAS	Learned agent/time credit from trajectories	Learns credit from data; SRC computes a state-level responsibility matrix from observed geometry.
Graph and sub-team decomposition	XMIX, MACA	Agent graph, correlated set, or sub-team structure	Learns interaction or sub-team structure inside the learner; SRC provides an explicit object-responsibility signal that can be attached to such learners.
OT and geometric shaping	Distribution matching and global OT shaping	Team-level geometric or distributional objective	Uses OT as a global scalar; SRC uses the Sinkhorn plan as an agent-object responsibility matrix.
Reward shaping theory	Potential-based shaping and potential-based difference rewards	Invariance-preserving auxiliary reward	Gives conditions for shaping; SRC is a concrete spatial responsibility construction that can be injected into fixed learners.

Table 5: Evidence map used in the paper. The core claim is based on MPE mechanism tests and VMAS scalability tests under a fixed learner. Additional experiments are assigned explicit roles and are not interpreted as independent primary wins.

Experiment	Seeds	Evaluation purpose	Interpretation
MPE $N = 3, 5, 7$	20	SRC-M vs. shared/global OT	main mechanism, significant
VMAS $N = 4, 6, 8$	20	SRC estimator scaling	main scale-up, SRC-S/T significant
SRC variants	20	Static, temporal, marginal estimators	ablation
Diff.+SRC hybrid	20 matched	Credit composability	mixed; MPE significant, VMAS inconclusive
A2C/DQN backbones	20	Learning-backbone sensitivity	auxiliary
CC-PPO critic	20	Centralized-critic plug-in check	auxiliary, learner-dependent
LBF/RWARE/Pursuit	10	Interaction-heavy boundary cases	scope diagnostic
SMACv2 Terran 5v6	10	Combat-object diagnostic	scope diagnostic
GRF keeper	20	Spatial-objective alignment diagnostic	shaped-signal positive
BenchMARL/TorchRL MAPPO $N = 8$	20	Modern framework adapter	compatible, no significant gain

Table 6: VMAS estimator scaling with PPO and 200k environment steps. Values are original evaluation return, mean $\pm$ 95% CI. All five modes use 20 matched seeds; p-values compare against shared reward. Static and temporal SRC scale reliably; marginal SRC is more variable.

Team size	Shared	SRC-M	SRC-S	SRC-T
$N = 4$	2.67 ± 0.25	2.97 ± 0.17 ($p = 0.073$)	$\mathbf{3.77 \pm 0.10}$ ($p = 5.57\times10^{-8}$)	3.60 ± 0.14 ($p = 6.94\times10^{-8}$)
$N = 6$	0.96 ± 0.51	0.68 ± 2.29 ($p = 0.789$)	5.08 ± 0.12 ($p = 6.89\times10^{-13}$)	$\mathbf{5.13 \pm 0.16}$ ($p = 7.29\times10^{-13}$)
$N = 8$	0.27 ± 1.55	2.46 ± 1.27 ($p = 0.066$)	7.64 ± 0.23 ($p = 4.99\times10^{-9}$)	$\mathbf{7.68 \pm 0.33}$ ($p = 4.54\times10^{-9}$)

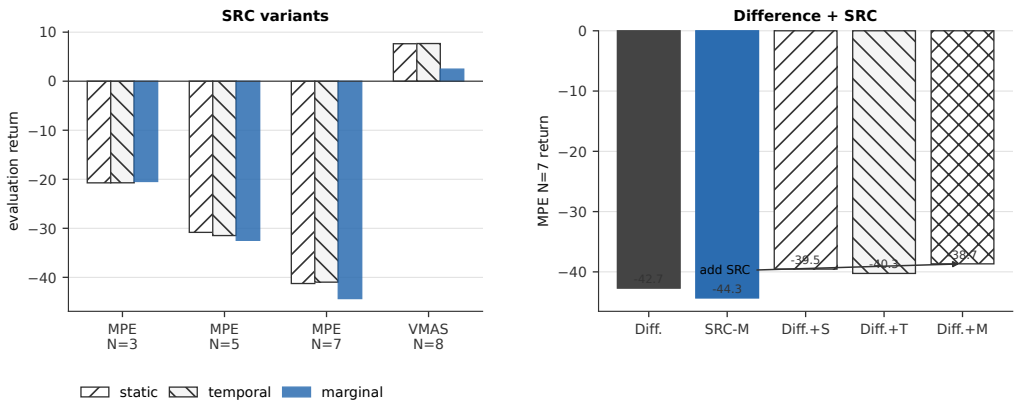

Figure 4: Ablation summary. Left: SRC variants show that static, temporal, and marginal responsibility estimators behave differently across MPE and VMAS. Right: the archived MPE $N = 7$ Difference+SRC sweep illustrates the composability hypothesis; Table 3 reports the stricter focused MPE and VMAS reruns, which are positive but original-metric-inconclusive.

B Reproducibility Details

The experimental code is organized around a fixed learning backbone and interchangeable credit signals. The core SRC implementation is in `experiments/src/ot_credit/`, including the Sinkhorn solver, credit-signal construction, and environment wrappers. Training entry points are in `experiments/scripts/`: `train_ippo_src.py` for MPE, `train_ippo_vmas_src.py` for VMAS, and analogous scripts for LBF, RWARE, and Pursuit. Sweep scripts such as `run_credit_sweep.py` and `run_vmas_credit_sweep.py` enumerate matched seeds and credit modes; each supports `--resume` to skip completed `metrics.csv` files and `--dry-run` to inspect commands before launching training.

Software setup. The repository separates lightweight tests from training dependencies. The expected setup is to create a Python virtual environment, install `requirements-core.txt` for unit tests, and install `requirements-training.txt` for Stable-Baselines3 (Raffin et al., 2021), PettingZoo/SuperSuit (Terry et al., 2021), VMAS (Bettini et al., 2022), LBF, and RWARE experiments. LBF and RWARE follow the cooperative benchmark suite introduced by Papoudakis et al. (2021). We used the repository tests under `experiments/tests/` to check the Sinkhorn solver, credit signals, wrappers, and result utilities before running sweeps. The code release also includes a separate BenchMARL/TorchRL entry point for this framework check. The reported BenchMARL result uses native MAPPO on VMAS navigation with an SRC reward adapter and final unshaped evaluation, while smoke runs and calibration probes remain excluded from the paper evidence.

Main MPE sweeps. The primary MPE runs use `run_credit_sweep.py` with PPO, `simple_spread`, `total_timesteps=200000`, `max_cycles=25`, `local_ratio=0.5`, `epsilon=0.1`, `lambda_credit=0.5`, and `alpha_temporal=0.2`. The credit modes are `shared`, `global_ot`, `src_marginal`, and `difference` for the main comparison, with `src_static` and `src_temporal` added for variant ablations. The main team sizes are $N = 3, 5, 7$, and the main comparisons use matched seeds. A representative command is:

```
python experiments/scripts/run_credit_sweep.py \
  --modes shared global_ot src_marginal difference \
  --seeds 0 1 ... 19 \
  --num-agents 3 5 7 \
  --total-timesteps 200000 \
  --lambda-credit 0.5 \
  --epsilon 0.1 \
  --alpha-temporal 0.2 \
```

```
--resume
```

The PPO configuration is the Stable-Baselines3 `MlpPolicy` with `gamma=0.99`, `learning_rate=3e-4`, `n_steps=1024`, and `batch_size=256`. Unless otherwise specified, we use the SB3 default MLP policy architecture for each algorithm. Auxiliary A2C and DQN checks use the same MPE wrapper and sweep script with `--algo a2c` or `--algo dqn`. The CC-PPO check uses a lightweight centralized critic with the global MPE state and an agent-id input, while the actor still observes each agent's local observation. A2C uses `gamma=0.99`, `learning_rate=3e-4`, and `n_steps=1024`. DQN uses a replay buffer of 50,000, `learning_starts=1000`, `batch_size=128`, `train_freq=4`, `gradient_steps=1`, `target_update_interval=1000`, and final exploration 0.05.

VMAS and stress-test sweeps. VMAS navigation uses `run_vmas_credit_sweep.py` with `scenario=navigation`, `num_agents=4,6,8`, `total_timesteps=200000`, `max_steps=200`, `epsilon=0.1`, `alpha_temporal=0.2`, and `lambda_credit=0.1`. The VMAS scale-up comparison uses 20 matched seeds for `shared`, `global_ot`, `src_marginal`, `src_static`, and `src_temporal`. LBF, RWARE, and Pursuit stress tests use the corresponding `run_*_credit_sweep.py` scripts with the same resume logic; these runs are reported as scope tests rather than as primary claims.

BenchMARL/TorchRL diagnostic. The BenchMARL/TorchRL check uses `experiments/scripts/run_benchmarl_vmas_src_pilot.py` in a separate BenchMARL/TorchRL virtual environment to avoid dependency conflicts with the Stable-Baselines3 stack. The reported setting is VMAS `navigation`, MAPPO, $N = 8$, `frames_per_batch=5000`, `n_iters=20`, 100k total frames, `eval_episodes=50`, and matched seeds 0–19. The comparison is native MAPPO versus `potential_src_static` with `lambda_credit=0.01`, centered credit normalization, and `potential_gamma=0.99`. We evaluate final unshaped return and success, and parse logged mean returns only as a learning-curve proxy.

Evaluation and aggregation. Training writes one directory per environment, credit mode, and seed under `experiments/runs/ppo_credit/`. Each completed run stores a trained model and `metrics.csv`. We report original task-level metrics except for explicitly labeled GRF and SMACv2 diagnostics. MPE statistics use `experiments/scripts/analyze_fair_mpe.py`; revision-batch VMAS/GRF/SMACv2 summaries use `experiments/scripts/analyze_revision_sync_results.py`; and paper CSVs are gathered with `experiments/scripts/curate_paper_results.py`. For interrupted sweeps, rerunning the same command with `--resume` skips completed runs.

C Statistical Evidence for the Main MPE Claim

The main statistical claim concerns MPE simple spread: under matched random seeds, SRC improves over shared rewards and global OT shaping. Table 7 reports paired benefits for original evaluation return. Table 8 repeats the analysis for final coverage cost, with signs reversed so that positive values favor SRC. Both metrics support the same conclusion: SRC reliably improves over shared rewards and global OT, while difference rewards remain competitive and often stronger. We report paired t-tests with Holm correction, Wilcoxon signed-rank tests with Holm correction, Cohen's d_z, 95% bootstrap confidence intervals, and seed-wise win counts.

Table 7: MPE original evaluation return paired statistics with 20 matched seeds. Positive benefit means SRC has higher return than the baseline. The primary SRC-vs-shared and SRC-vs-global-OT comparisons are significant after Holm correction for every team size.

Baseline	N	Benefit	95% CI	d_z	Wins	p_t Holm	p_w Holm
Shared	3	5.07	$[3.96, 6.21]$	1.94	20/20	3.53×10^{-7}	1.72×10^{-5}
Shared	5	14.22	$[11.79, 16.97]$	2.35	20/20	2.14×10^{-8}	1.72×10^{-5}
Shared	7	20.03	$[15.96, 24.33]$	2.04	20/20	1.75×10^{-7}	1.72×10^{-5}
Global OT	3	2.71	$[2.08, 3.39]$	1.77	20/20	1.20×10^{-6}	1.72×10^{-5}
Global OT	5	11.04	$[7.91, 14.50]$	1.42	19/20	2.08×10^{-5}	1.91×10^{-5}
Global OT	7	16.50	$[10.24, 22.91]$	1.12	18/20	3.28×10^{-4}	5.34×10^{-4}
Difference	3	-0.92	$[-1.54, -0.30]$	-0.63	4/20	3.23×10^{-2}	3.62×10^{-2}
Difference	5	-1.46	$[-2.50, -0.46]$	-0.62	5/20	3.23×10^{-2}	3.62×10^{-2}
Difference	7	-1.64	$[-4.01, 0.87]$	-0.29	7/20	2.14×10^{-1}	1.23×10^{-1}

Table 8: MPE final coverage cost paired statistics with 20 matched seeds. Coverage cost is lower-is-better, so the benefit is sign-reversed; positive values favor SRC. The coverage analysis independently supports the same spatial-responsibility conclusion as return.

Baseline	N	Benefit	95% CI	d_z	Wins	p_t Holm	p_w Holm
Shared	3	0.31	$[0.22, 0.40]$	1.43	20/20	2.39×10^{-5}	1.72×10^{-5}
Shared	5	0.59	$[0.48, 0.72]$	2.15	20/20	8.64×10^{-8}	1.72×10^{-5}
Shared	7	0.63	$[0.50, 0.76]$	2.06	20/20	1.59×10^{-7}	1.72×10^{-5}
Global OT	3	0.16	$[0.12, 0.20]$	1.75	19/20	1.62×10^{-6}	2.29×10^{-5}
Global OT	5	0.45	$[0.30, 0.62]$	1.20	19/20	1.72×10^{-4}	2.29×10^{-5}
Global OT	7	0.48	$[0.29, 0.67]$	1.06	18/20	5.46×10^{-4}	6.71×10^{-4}
Difference	3	-0.04	$[-0.06, -0.02]$	-0.71	4/20	1.12×10^{-2}	1.12×10^{-2}
Difference	5	-0.05	$[-0.08, -0.02]$	-0.74	4/20	1.12×10^{-2}	1.10×10^{-2}
Difference	7	-0.05	$[-0.12, 0.02]$	-0.28	6/20	2.21×10^{-1}	1.43×10^{-1}

D Experimental Setup and Auxiliary Results

The main text contains the primary MPE/VMAS results and compressed ablation analysis. Appendix A collects large supporting tables and schematic figures moved to the appendix for page-limit compliance. Tables 9, 10, and 12 report auxiliary backbone checks, including a lightweight centralized-critic PPO run, and Table 13 reports stress-test environments that are less central to the main claim but important for defining the boundary of the method.

Table 9: MPE $N = 7$ original evaluation return under auxiliary backbones. Higher is better. A2C and DQN use SB3; CC-PPO is a lightweight centralized-critic PPO implementation used only as a plug-in check. SRC-M improves over shared rewards and global OT under A2C, improves over global OT under DQN, and remains competitive with Difference. CC-PPO is mixed: SRC-M improves over shared rewards and Difference but not over global OT.

Backbone	Shared	Global OT	SRC-M	Difference
PPO (20 seeds)	-64.37	-60.83	-44.33	$\mathbf{-42.70}$
A2C (20 seeds)	-47.65	-49.97	-43.67	$\mathbf{-43.44}$
DQN (20 seeds)	-35.54	-36.36	-34.92	$\mathbf{-34.75}$
CC-PPO (20 seeds)	-57.37	$\mathbf{-50.67}$	-52.52	-59.53

Table 10: Lightweight centralized-critic PPO credit-signal check on MPE $N = 7$. Shared, Global OT, Difference, and SRC-M use 20 matched seeds; SRC-S and SRC-T are 10-seed variant probes. Values are original evaluation return, mean $\pm$ 95% CI; higher is better. The result supports compatibility with a centralized critic but does not show dominance over global OT.

Credit signal	Return	Benefit vs Shared	Wins vs Shared
Shared	-57.37 ± 5.61	–	–
Global OT	$\mathbf{-50.67 \pm 3.49}$	6.70	12/20
Difference	-59.53 ± 5.01	-2.16	9/20
SRC-M	-52.52 ± 2.08	4.85	12/20
SRC-S	-51.96 ± 2.99	6.36	7/10
SRC-T	-53.34 ± 7.99	4.98	7/10

Table 11: Native BenchMARL/TorchRL MAPPO diagnostic on VMAS navigation with $N = 8$, 100k frames, and 20 matched seeds. Values are final unshaped evaluation metrics, mean $\pm$ 95% CI. Potential-shaped SRC is portable to BenchMARL but does not significantly improve native MAPPO in this setting.

Credit signal	Return	Success	Paired return p
Native MAPPO	0.2803 ± 0.0038	0.973 ± 0.006	–
Potential SRC	0.2785 ± 0.0034	0.976 ± 0.006	0.388

Table 12: Matched-seed paired checks for auxiliary backbones. Positive benefit means SRC-M has higher return than the baseline under the same backbone and seed. These checks support the claim that SRC-M can be attached beyond the main PPO setting, but they remain auxiliary because PPO is the main controlled backbone and the results are learner-dependent.

Backbone	Baseline	Benefit	p	Wins
A2C	Shared	3.98	6.52×10^{-4}	15/20
A2C	Global OT	6.31	4.11×10^{-3}	16/20
A2C	Difference	-0.23	7.70×10^{-1}	12/20
DQN	Shared	0.61	3.42×10^{-1}	10/20
DQN	Global OT	1.44	2.10×10^{-3}	15/20
DQN	Difference	-0.18	6.93×10^{-1}	8/20
CC-PPO	Shared	4.85	1.12×10^{-1}	12/20
CC-PPO	Global OT	-1.85	3.54×10^{-1}	8/20
CC-PPO	Difference	7.01	1.08×10^{-2}	14/20

Table 13: Representative stress-test original evaluation returns with 10 seeds. Higher is better. LBF and RWARE are sparse enough that most unshaped evaluations remain near zero; Pursuit uses negative distance-style returns. These matched comparisons do not support a universal SRC claim and instead identify settings where interaction rules dominate geometric responsibility.

Stress test	Shared	Global OT	SRC-M	Interaction-aware / task utility
LBF 8x8	$\mathbf{0.20}$	0.07	0.00	Diff. 0.00
LBF 10x10	0.00	0.00	0.00	Diff. 0.00
RWARE tiny	$\mathbf{0.02}$	0.00	0.00	IA-SRC 0.00
Pursuit	$\mathbf{-15.46}$	-17.74	-17.81	IA-SRC -17.63

Table 14: Interaction-aware diagnostic utilities used in stress tests. These utilities are hand-designed probes for distance-only failures, not additional proposed algorithms. LBF, RWARE, and Pursuit use leave-one-out utility differences; SMACv2 uses responsibility-weighted combat readouts.

Benchmark	Diagnostic utility	Interaction state added beyond distance
LBF	IA-SRC	Logistic score for whether adjacent loading agents have enough total level for each food item; optionally conditions on the LOAD action.
RWARE	IA-SRC	Logistic score for whether requested shelves are adjacent to a robot or already carried by one.
Pursuit	IA-SRC	Logistic score for whether enough pursuers are adjacent to each evader to satisfy the capture condition.
SMACv2	Combat utilities	Responsibility-weighted enemy proximity augmented with low-health priority, attack-range bonus, and observed damage or kill events.

Table 15: SMACv2 diagnostic on `10gen_terran` with 5 allied units versus 6 enemies, PPO, 20k environment steps, and 10 matched seeds. Return values are mean $\pm$ 95% CI; success is the mean success rate. Geometry-only SRC hurts performance; combat-aware utilities recover part of the gap but do not produce a statistically reliable improvement over shared reward.

Credit signal	Original return	Original success	p vs Shared
Shared	1.03 ± 0.32	0.64	–
SRC-M	0.13 ± 0.14	0.16	5.39×10^{-4}
Combat SRC	0.97 ± 0.27	0.72	0.745
Damage-delta SRC	$\mathbf{1.07 \pm 0.41}$	**0.80**	0.873
Geo+health SRC	0.92 ± 0.24	0.74	0.435
Geo+range SRC	0.81 ± 0.38	0.72	0.363

Table 16: GRF diagnostic on `academy_3_vs_1_with_keeper` with PPO, 500k environment steps, and 20 matched seeds. SRC-M strongly improves the shaped spatial signal and learning-speed metrics, but the original scoring metrics remain noisy. Higher is better except time-to-threshold.

Metric	Shared	SRC-M	Paired p
Final shaped return	0.0003	0.393	2.00×10^{-5}
Shaped-return AUC	–	+0.424 benefit	1.39×10^{-8}
Steps to shaped return 0.1	–	508k-step benefit	1.32×10^{-12}
Shaped success	0.64	0.90	0.041
Original success	0.66	0.82	0.159
Original return	0.0003	−0.099	0.164

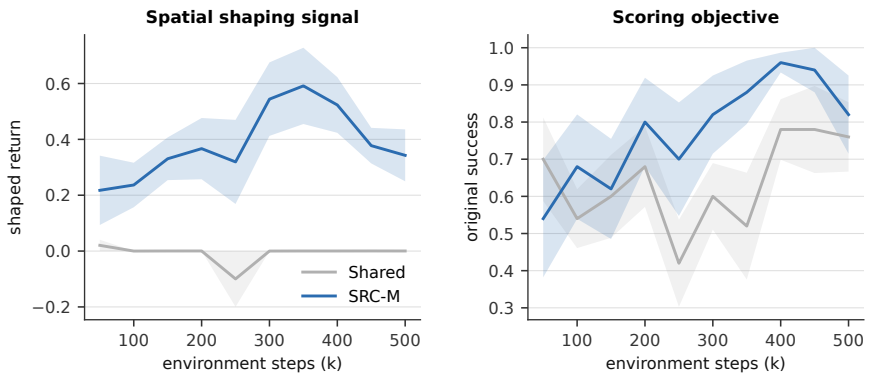

Figure 5: GRF diagnostic learning curves on `academy_3_vs_1_with_keeper`. Lines show the earlier 10-seed diagnostic mean and standard-error bands; Table 16 reports the updated 20-seed aggregate statistics. SRC-M rapidly improves the shaped spatial signal, while the original scoring success remains noisy and less clearly separated.