Multi-Agent RL with Invisible Collaborators: Marginal Advantage Estimation for Indirect Cooperation

Jianglin Qiao[*1,2] **Zehong Cao**[†1] **Siyi Hu**[1,3] **Mingjun Fan**[1] **Mahardhika Pratama**[1] **Ryszard Kowalczyk**[1,4]

[1]School of CSIT, Adelaide University, Adelaide, SA, Australia
[2]Australian Centre for Robotics, The University of Sydney, Sydney, NSW, Australia
[3]School of EECMS, Curtin University, Bentley, WA, Australia
[4]Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Abstract

Cooperative Multi-Agent Reinforcement Learning (MARL) has primarily focused on direct coordination among agents. However, many real-world systems require indirect cooperation, where agents share a team objective but are structurally isolated in their observation or interaction pathways. In such settings, estimating individual contributions from sparse global rewards becomes challenging because reward-relevant teammates may be absent from an agent's local observation history. We propose Marginal Advantage Estimation (MAE), a representation-level contribution estimation method for cooperative MARL under structural isolation. MAE computes a full-vs-masked value difference inside the centralised critic by masking the agent-specific input block, and uses this signal as a stop-gradient auxiliary term alongside standard GAE. We provide conditional variance-reduction analysis under explicit critic-error and advantage-decomposition assumptions. Experiments across 18 tasks in Partitioned MPE and Basilisk show that MAE is most beneficial in structurally isolated, reward-coupled settings, with the most consistent improvements observed when integrated with MAPPO.

1 INTRODUCTION

Cooperative Multi-Agent Reinforcement Learning (MARL) studies how multiple agents learn coordinated behaviour under a shared team objective [Sunehag et al., 2017]. Contemporary approaches, including value decomposition methods (e.g., QMIX [Rashid et al., 2020b], WQMIX [Rashid et al., 2020a], QPLEX [Wang et al., 2020a]) and centralised-critic frameworks (e.g., MADDPG [Lowe et al., 2017], MAPPO [Yu et al., 2022], HAPPO [Zhong et al., 2024]), have demonstrated strong empirical performance across standard cooperative benchmarks. By leveraging structured value factorisation or centralised value estimation under the Centralised Training with Decentralised Execution (CTDE) paradigm, these methods enable agents to optimise a shared global objective through coordinated policies.

Despite this progress, existing MARL algorithms largely assume a direct cooperation paradigm [Lowe et al., 2017, Rashid et al., 2020b, Yu et al., 2022]. In this setting, agents operate in overlapping spatial or informational domains and can influence one another through observable interactions [Ma et al., 2021, Li et al., 2024, Zhou et al., 2025]. Under such conditions, correlations between individual behaviour and global return are discoverable from trajectory data, allowing advantage estimators or value factorisation mechanisms to meaningfully attribute individual contributions. However, many real-world systems violate this assumption [Zhu et al., 2024]. In distributed robotic teams [Drew, 2021], partitioned sensor networks [Mosalli and Aghdam, 2024] and satellite constellations [Zilberstein et al., 2025], agents share a global objective yet are permanently separated by physical, temporal, or organisational constraints. Although they share a common reward, specific subsets of agents are structurally prevented from directly interacting with or observing one another. We refer to this setting as indirect cooperation: agents are logically coupled through the team objective but structurally isolated in their interaction pathways.

This structural isolation introduces a fundamental challenge for marginal contribution estimation. A common strategy for individual contribution evaluation in cooperative MARL is to construct action-level counterfactuals, measuring how the team return changes when a focal agent deviates from the joint action. Representative approaches include counterfactual advantage estimators such as COMA [Foerster et al., 2018] and game-theoretic value allocation methods based on Shapley [Wang et al., 2020b] or Nucleolus prin-

[*]Corresponding Author
[†]Co-corresponding Author

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:5573–5600.

ciples [Li et al., 2025]. These estimators implicitly assume that modifying an agent's action or coalition participation can meaningfully alter coordination outcomes. Under indirect cooperation, however, this assumption breaks down. When interaction pathways are permanently absent, altering an agent's action may not affect the teammates with whom coordination is logically required. Consequently, action perturbations induce negligible changes in global value, producing weak or uninformative marginal signals. The issue is therefore not merely statistical bias due to partial observability, but a structural collapse of the action-level counterfactual signal under isolation. Current MARL methods primarily rely on these action-level counterfactuals, where an agent's contribution is estimated by varying its action while keeping other agents' actions fixed. This fundamentally introduces high variance and unstable credit assignment due to the noisy nature of action-level perturbations.

To address this limitation, we propose Marginal Advantage Estimation (MAE), a plug-in credit-assignment module for cooperative MARL under structural isolation. Rather than perturbing actions or enumerating coalitions, MAE computes a representation-level marginal signal inside the centralised critic. Specifically, it compares the factual critic value with a masked critic value in which the agent-specific input block is removed. This full-vs-masked value difference estimates the contribution carried by an agent's information under the current joint training sample. The resulting signal is added to standard GAE as a stop-gradient auxiliary term, so the policy-gradient advantage remains provided by GAE while MAE supplies additional credit information during centralised training. Under explicit critic-error assumptions, this construction can reduce estimator variance relative to independent simulation-based counterfactual estimation. We evaluate MAE across 18 tasks, including Partitioned MPE [Lowe et al., 2017] and the high-fidelity Basilisk satellite simulator [Kenneally et al., 2020, Harris et al., 2022, Herrmann et al., 2024, Stephenson and Schaub, 2024]. The results show that MAE is most beneficial in structurally isolated, reward-coupled settings, with the strongest and most consistent improvements observed when MAE is integrated with MAPPO. Our key contributions are:

- **Problem Formulation.** We formalise indirect cooperation as a structurally isolated but reward-coupled Dec-POMDP setting, where local histories may provide weak information about reward-relevant teammates.
- **Marginal Advantage Estimation.** We introduce MAE, a representation-level full-vs-masked critic signal that provides auxiliary marginal credit information without requiring additional rollouts or coalition enumeration.
- **Theoretical and Empirical Analysis.** We provide conditional variance-reduction analysis under explicit critic-error and advantage-decomposition assumptions, and empirically evaluate MAE across 18 MPE and Basilisk tasks.

2 RELATED WORK

Observability and Cooperative MARL. Partial observability is a fundamental challenge in multi-agent systems, formally modelled as Decentralised POMDPs (Dec-POMDPs) [Oliehoek et al., 2016, Gronauer and Diepold, 2022, Peralez et al., 2025]. While the Centralised Training with Decentralised Execution (CTDE) paradigm [Lowe et al., 2017, Rashid et al., 2020b, Yu et al., 2022] has enabled significant breakthroughs by leveraging global states during training (e.g., QMIX [Rashid et al., 2020b], MAPPO [Yu et al., 2022]), decentralised execution remains bottlenecked by local information limits. To mitigate this, standard approaches typically infer teammates' latent states by aggregating observation histories via RNNs [Hausknecht and Stone, 2015, Igl et al., 2018] or establishing explicit communication channels [Foerster et al., 2016]. Crucially, these solutions rely on a core assumption inherent to direct cooperation: that teammates' states are statistically inferable from local observation sequences or communication signals. Under indirect cooperation, this premise becomes less reliable. Structural isolation limits the information that an agent's local history contains about teammates in other operating regions, making history-based inference less informative when sparse rewards depend on those unobserved teammates. Unlike existing methods, our work directly targets this structurally isolated but reward-coupled setting, where local histories may provide weak signals for assigning individual contributions.

Counterfactuals in Cooperative MARL. Difference rewards evaluate individual contributions via counterfactual baselines [Wolpert and Tumer, 2001]. While methods like COMA [Foerster et al., 2018] marginalise over actions to reduce variance, game-theoretic approaches (e.g., Shapley-Q [Wang et al., 2020b], Nucleolus-Q [Li et al., 2025]) leverage coalition permutations for fair credit assignment. However, these action-level and coalition-based estimators implicitly assume structural coherence among teammates. Under indirect cooperation, physical isolation can make action perturbations weakly informative, because changing one agent's local action may not immediately affect distant teammates that are nevertheless coupled through the shared reward. Coalition-based estimators face a related difficulty: their marginal values can become expensive and noisy to estimate when reward-relevant agents are separated in observation and interaction pathways. Our proposed MAE addresses this by shifting the counterfactual construction from action or coalition perturbation to representation-level masking inside the centralised critic. Instead of simulating alternative actions, MAE masks the agent-specific input block and compares the factual critic value with the masked critic value under the same sampled training batch. This provides a plug-in marginal contribution signal that can be added to standard GAE as a stop-gradient auxiliary term, without requiring extra rollouts or coalition enumeration. By

evaluating the full and masked critic inputs in one batched critic computation, MAE keeps the training sample and critic parameters fixed, which can reduce estimator variance under the correlated-error assumptions discussed in Sec. 3.

3 METHOD

3.1 PROBLEM FORMULATION

We formalise the cooperative multi-agent environment as a Dec-POMDP operating under the specific setting of indirect cooperation. The problem is defined by the tuple $\mathcal{M} = \langle \mathcal{N}, \mathcal{S}, \mathcal{A}, T, R, \Omega, O, \gamma \rangle$. Specifically, $\mathcal{N} = \{1, \ldots, n\}$ denotes the set of n agents operating within a bounded physical domain $\mathcal{X} \subseteq \mathbb{R}^d$. The global state space $\mathcal{S}$ captures the complete system configuration; crucially, it embeds the physical location of each agent, meaning any state $s \in \mathcal{S}$ defines a specific coordinate configuration in $\mathcal{X}$. At each timestep t, the system receives a joint action $\mathbf{a}_t = \{a_{1,t}, \ldots, a_{n,t}\}$ drawn from the joint action space $\mathcal{A} = \times_{i=1}^{n} \mathcal{A}_i$, governing the system evolution via the state transition dynamics $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$. To induce cooperation, the team shares a single global reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, aiming to maximise the discounted return $J = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t]$ with a discount factor $\gamma \in [0, 1)$. Thus, the setting should not be interpreted as a collection of independent MDPs: agents remain coupled through the shared transition dynamics and team reward, while the structural separation appears in the observation and direct-interaction pathways. Finally, $\Omega = \times_{i=1}^{n} \Omega_i$ represents the joint observation space, where the observation function $O : \mathcal{S} \times \mathcal{N} \to \Omega$ determines the local information available to each agent.

In our indirect cooperation setting, the observation and direct-interaction pathways are constrained by the geometry of $\mathcal{X}$. Each agent i is assigned a fixed physically accessible or operating region $\mathcal{B}_i \subset \mathcal{X}$. We assume structural isolation between different agents' operating regions, i.e., $\mathcal{B}_i \cap \mathcal{B}_j = \emptyset, \quad \forall i \neq j$. The local observation of agent i is obtained from the region-restricted projection of the global state, $o_{i,t} = O_i(s_t) = \eta_i\big(\mathrm{proj}_{\mathcal{B}_i}(s_t)\big)$, where $\mathrm{proj}_{\mathcal{B}_i}$ extracts the components of the global state that lie within agent i's accessible region, and η_i represents local sensing limitations such as noise, occlusion, or partial field of view. When a time-varying sensing footprint is present, it is treated as a local subset of the operating region, e.g., $\mathcal{F}_{i,t} \subseteq \mathcal{B}_i$, rather than as the region partition itself. This formulation imposes two types of information limitation. First, structural inter-agent isolation prevents an agent's local history from directly observing or inferring the state of teammates located in other regions. Such teammates are therefore invisible from the perspective of the decentralised policy. Second, within its own region, each agent may still face ordinary local partial observability due to sensor noise and incomplete perception. Importantly, this isolation does not make the agents independent. The global transition T and team reward R remain defined over the joint state and joint action, so agents are still reward-coupled and must solve a cooperative task. This creates the key credit-assignment difficulty studied in this paper: sparse global rewards may depend on teammates whose states and behaviours are absent from an agent's local observation history. As a result, local or action-level credit signals can become weak or high-variance under structural isolation, motivating a centralised-training signal that estimates each agent's marginal contribution from the critic representation.

3.2 MARGINAL ADVANTAGE ESTIMATION

To address the credit-assignment difficulty caused by indirect cooperation, we propose MAE that follows the standard CTDE paradigm: during training, a centralised critic can access a joint critic input, while each decentralised policy still conditions only on its own local observation during execution. Importantly, MAE is not a standalone action-conditioned policy-gradient advantage estimator. Instead, it is a critic-derived representation-level marginal-contribution signal that augments standard GAE.

Let x_t^{cent} denote the centralised critic input at time t, which may contain the joint observation, global state information, or benchmark-specific task variables available during centralised training. For each agent i, we define a binary mask m^{-i} that zeroes out only the agent-specific input block corresponding to agent i and keeps the remaining critic input unchanged. The masked critic input is

$$x_t^{-i} = x_t^{\mathrm{cent}} \odot m^{-i}. \tag{1}$$

The MAE signal is then computed as the full-vs-masked critic value difference:

$$M_i(t) = V_\phi(x_t^{\mathrm{cent}}) - V_\phi(x_t^{-i}). \tag{2}$$

Here, masking should not be interpreted as physically removing agent i from the environment or changing the executed trajectory. Rather, it removes the information carried by agent i's input block from the centralised critic representation. Thus, $M_i(t)$ estimates the marginal value contribution associated with agent i's information under the current joint training sample.

The policy update combines this critic-derived signal with the standard GAE term:

$$\hat{A}_i^{\mathrm{total}}(t) = \hat{A}_i^{\mathrm{GAE}}(t) + \lambda \, \mathrm{sg}[M_i(t)], \tag{3}$$

where λ controls the strength of the MAE signal and $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator. Therefore, the action-conditioned policy-gradient signal remains provided by GAE, while MAE acts as an auxiliary credit signal derived from the centralised critic. This design allows MAE to

provide additional credit information under structural isolation without requiring extra rollouts, coalition enumeration, or inter-agent communication during execution.

In implementation, MAE requires two critic value queries on the same sampled training batch: one factual query using x_t^{cent} and one masked query using x_t^{-i}. These two queries can be evaluated efficiently by stacking the full and masked critic inputs along the batch dimension and passing them through the same centralised critic in one batched forward computation. This implementation does not require additional trajectory sampling or a separate replay buffer.

We next analyse why this full-vs-masked construction can provide a lower-variance marginal signal than simulation-based counterfactual estimation. Consider a simulation-based counterfactual baseline, rooted in difference rewards [Wolpert and Tumer, 2001], that estimates an agent's marginal contribution using two separate rollouts: one factual rollout and one counterfactual rollout where the agent is removed or intervened upon. Such independent rollouts introduce independent estimation errors from environmental stochasticity and sampling noise. In contrast, MAE evaluates the factual and masked critic values using the same training sample and the same critic parameters, which can induce positively correlated estimation errors. The following result formalises this conditional variance-reduction effect.

Theorem 1 (Variance Reduction under Correlated Critic Errors)**.** *Let a simulation-based Monte Carlo counterfactual estimator be written as*

$$\hat{\Delta}_i^{\text{MC}} = \Delta_i + \epsilon_{\text{fact}} - \epsilon_{\text{cf}},$$

where Δ_i is the target marginal quantity, and the factual and counterfactual errors are independent with variances $\text{Var}(\epsilon_{\text{fact}}) = \sigma_{\text{fact}}^2$ and $\text{Var}(\epsilon_{\text{cf}}) = \sigma_{\text{cf}}^2$. Let the MAE estimator be written as

$$\hat{\Delta}_i^{\text{MAE}} = \Delta_i + \delta_{\text{fact}} - \delta_{\text{mask}},$$

where the factual and masked critic errors have no larger marginal variances, $\text{Var}(\delta_{\text{fact}}) \le \sigma_{\text{fact}}^2$ and $\text{Var}(\delta_{\text{mask}}) \le \sigma_{\text{cf}}^2$, and have positive covariance $\text{Cov}(\delta_{\text{fact}}, \delta_{\text{mask}}) = c > 0$. Then

$$\text{Var}(\hat{\Delta}_i^{\text{MAE}}) < \text{Var}(\hat{\Delta}_i^{\text{MC}}). \tag{4}$$

Proof Sketch. For the independent simulation-based estimator, the two rollout errors are uncorrelated, giving

$$\text{Var}(\hat{\Delta}_i^{\text{MC}}) = \sigma_{\text{fact}}^2 + \sigma_{\text{cf}}^2.$$

For MAE, the factual and masked critic values are evaluated from the same sampled training batch using the same critic parameters. Under the stated positive-correlation assumption,

$$\begin{aligned}\text{Var}(\hat{\Delta}_i^{\text{MAE}}) &= \text{Var}(\delta_{\text{fact}}) + \text{Var}(\delta_{\text{mask}}) - 2\text{Cov}(\delta_{\text{fact}}, \delta_{\text{mask}}) \\ &\le \sigma_{\text{fact}}^2 + \sigma_{\text{cf}}^2 - 2c \\ &< \sigma_{\text{fact}}^2 + \sigma_{\text{cf}}^2 = \text{Var}(\hat{\Delta}_i^{\text{MC}}).\end{aligned}$$

Thus, when the shared critic induces positively correlated factual and masked estimation errors without increasing their marginal variances, the MAE estimator has lower variance than an independent simulation-based counterfactual estimator. The full derivation is provided in Appendix A.1. □

This result should be interpreted as a conditional variance-reduction statement under the stated critic-error assumptions. It does not imply that MAE is an unbiased action-conditioned advantage estimator, nor does it provide a universal convergence guarantee for arbitrary MARL dynamics. In practice, the reliability of the masked value difference depends on critic quality and on whether the agent-specific input block is sufficiently factorised for masking to remain meaningful.

3.3 MAE-GUIDED POLICY OPTIMISATION

We integrate MAE into centralised-critic policy-gradient optimisation as a lightweight auxiliary credit signal. The standard GAE term remains the action-conditioned policy-gradient advantage computed from the shared team reward, while MAE provides an additional critic-derived marginal signal under indirect cooperation. For each agent i, we define

$$\hat{A}_i^{\text{total}}(t) = \hat{A}_i^{\text{GAE}}(t) + \lambda \cdot \text{sg}[M_i(t)], \tag{5}$$

where $M_i(t)$ is the full-vs-masked critic signal defined in Eq. (2), $\lambda \ge 0$ controls its strength, and $\text{sg}[\cdot]$ denotes the stop-gradient operator. The stop-gradient prevents the actor update from back-propagating through the critic representation; hence MAE acts as a fixed auxiliary credit weight during policy optimisation rather than as a separate objective. During execution, each policy still uses only its local observation.

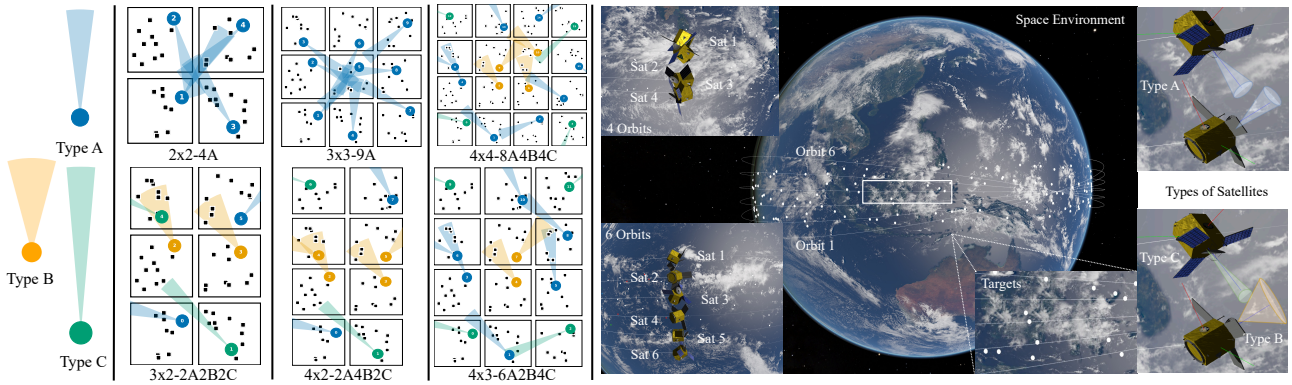
The intuition is that, under structural isolation, the global advantage may contain high-variance components caused by teammates that are not observable from agent i's local history. MAE uses the centralised critic during training to estimate the value contribution associated with agent i's own information, thereby providing an additional local credit signal.

Theorem 2 (Conditional Gradient-Variance Reduction)**.** *Suppose the global advantage for agent i can be decomposed as*

$$\hat{A}_i^{\text{GAE}}(t) = A_i^{\text{loc}}(t) + \xi_{-i}(t),$$

where $A_i^{\text{loc}}(t)$ is the local contribution and $\xi_{-i}(t)$ is a zero-mean latent component induced by structurally isolated teammates. Suppose the MAE signal satisfies

$$M_i(t) = A_i^{\text{loc}}(t) + \epsilon_i(t),$$



(a) Partitioned MPE Settings

(b) Basilisk Settings

Figure 1: Illustration of indirect cooperation scenarios with physical isolation and local partial observability: (a) Partitioned MPE: 12 tasks featuring 4-16 agents confined to 4-16 disjoint zones. (b) Basilisk: 4 Earth observation tasks involving 4 or 6 satellites coordinated across distinct orbital planes.

where $\epsilon_i(t)$ is a zero-mean approximation error with lower variance than $\xi_{-i}(t)$. Then the score-weighted MAE signal provides a lower-variance local credit direction than using the global advantage alone.

Proof Sketch. Let $z_i(t) = \nabla_{\theta_i} \log \pi_{\theta_i}(a_{i,t} \mid o_{i,t})$. Under the above decomposition, the global estimator contains the residual term $z_i(t)\xi_{-i}(t)$, while the MAE-based marginal signal contains $z_i(t)\epsilon_i(t)$. If $\epsilon_i(t)$ has lower variance than the teammate-induced latent component $\xi_{-i}(t)$, then the residual variance in the MAE-weighted direction is lower. Thus, under the stated assumptions, MAE can reduce the gradient noise associated with unobservable teammates. The full derivation is provided in Appendix A.2. □

Theorems 1 and 2 should be interpreted as conditional variance-reduction results under explicit critic-error and advantage-decomposition assumptions. They do not imply universal convergence or optimality guarantees. Instead, they justify MAE as a representation-level auxiliary credit signal for structurally isolated but reward-coupled cooperative tasks.

4 EXPERIMENTAL SETUP

We evaluate MAE across 18 tasks in two environments covering both direct and indirect cooperation settings (Fig. 1). First, based on the Multi-Agent Particle Environment (MPE) [Lowe et al., 2017], we construct 14 tasks: 2 standard boundary-free tasks as direct-cooperation baselines, and 12 Partitioned MPE tasks that enforce structural isolation by confining agents to disjoint spatial zones. Second, we use Basilisk [Kenneally et al., 2020, Harris et al., 2022, Stephenson and Schaub, 2024], a high-fidelity astrodynamics simulator for spacecraft dynamics, guidance, navigation, control, and mission-level tasking, to design 4 Earth-observation tasks. Basilisk has been used in studies of spacecraft autonomy and reinforcement-learning-based tasking, making it well-suited to evaluating MARL methods under realistic orbital and sensing constraints. In our tasks, heterogeneous satellites must coordinate across distinct orbital planes to image ground targets under time-varying visibility windows, limited field-of-view constraints, and sparse target-priority rewards. This setting is closely related to Earth-observation satellite scheduling, where spacecraft decisions are constrained by orbital access, pointing opportunities, and temporal coordination requirements [Herrmann and Schaub, 2022, Herrmann et al., 2024]. All tasks follow the CTDE paradigm. During execution, each decentralised policy observes only its local observation o_i and does not access joint observations or inter-agent communication. During training, the centralised critic receives the joint critic input required for value estimation and MAE computation. In Partitioned MPE, each local observation contains the agent velocity, relative target positions, and zone-boundary information for discrete movement control. In Basilisk, each local observation includes spacecraft attitude, orbital position, and instrument status for discrete attitude and sensor-activation commands. Both environments require coordinated target observation to obtain the shared team reward. Partitioned MPE provides a sparse global reward of $+1.0$ only when specified landmarks are simultaneously observed across different zones for 5 consecutive steps, whereas Basilisk assigns target-specific priority scores ($p_g \in [0.1, 1.0]$) for joint observations executed within a predefined 5-minute operational window.

To mitigate reward sparsity, we optionally use Potential-Based Reward Shaping (PBRS) [Ng et al., 1999] and Random Network Distillation (RND) [Burda et al., 2019]. These are algorithm-side auxiliary training signals used to improve learning under sparse rewards and exploration difficulty. They are not part of the environment defin-

ition and do not redefine the underlying task reward or the MAE marginal estimator. To provide a comprehensive benchmark, we augment three policy-gradient MARL algorithms, MAPPO [Yu et al., 2022], HATRPO [Kuba et al., 2022], and HAPPO [Zhong et al., 2024], with the MAE module. We compare them against value-based and game-theoretic credit-assignment methods, including COMA [Foerster et al., 2018], Shapley-Q [Wang et al., 2020b], HAD3QN [Zhong et al., 2024], HASAC [Liu et al., 2024], HAA2C [Zhong et al., 2024], and Nucleolus-Q [Li et al., 2025]. All models are implemented within the HARL [Zhong et al., 2024] framework. Practically, MAE is implemented through batched full-vs-masked critic evaluations and can be added to GAE as a plug-in auxiliary signal. We report the mean and standard deviation over five fixed random seeds. Training is executed on AMD EPYC 7T83 CPUs and NVIDIA 48 GB GPUs. Full hyperparameters and environment specifications are provided in Appendices A.3, B.1, and B.2.[1]

5 RESULTS

The 18-task suite is designed to evaluate MAE across four operational dimensions: (1) *scale*, varying the number of agents or satellites to assess scalability; (2) *heterogeneity*, comparing homogeneous teams with mixed-sensor or role-diverse configurations; (3) *spatial structure*, covering symmetric and asymmetric Partitioned MPE layouts with different region arrangements; and (4) *dynamics*, spanning grid-based particle motion in Partitioned MPE and high-fidelity orbital constraints in Basilisk. Learning curves and quantitative results are presented in Fig. 2 and Table 1, with detailed results provided in Appendices C.1 and C.2.

5.1 PERFORMANCE ANALYSIS

We evaluate the training dynamics and final performance of MAE-augmented algorithms against their standard baselines in Table 1. Overall, the results show that MAE is most beneficial in structurally isolated, reward-coupled settings, with the most consistent gains observed for MAE-MAPPO. The effects on HAPPO and HATRPO are more mixed, suggesting that the same critic-side MAE signal can interact differently with different actor-update procedures.

MPE with Direct Cooperation To examine whether MAE acts as a general-purpose performance booster or specifically addresses indirect cooperation, we conduct an ablation study by removing the physical partitions in the `2x2-4A` and `2x2-2A2C` environments. These tasks are reported as the MPE direct-cooperation group in Table 1. In these boundary-free settings, agents can directly observe or interact with one another, so standard advantage estimators already provide informative learning signals. Consistent with this interpretation, MAE yields only small and mixed changes: for example, MAE-MAPPO improves MAPPO by +3.8% in `Vanilla-2x2-4A`, but decreases by -2.1% in `Vanilla-2x2-2A2C`; MAE-HAPPO also decreases by -5.1% in `Vanilla-2x2-4A`. These results suggest that MAE is not simply a generic architectural enhancer. Its benefit becomes clearer when structural isolation is introduced, where local histories provide weaker information about reward-relevant teammates.

Partitioned MPE In the Partitioned MPE tasks, agents are confined to disjoint spatial regions while the shared reward depends on cross-region target observation. Under this setting, MAE-MAPPO provides the most consistent improvements across task scales and heterogeneity levels. In small tasks, MAE-MAPPO improves over MAPPO by +24.5% in `2x2-4A`, +18.0% in `2x2-2A2C`, +14.5% in `3x2-6A`, and +17.7% in `3x2-2A2B2C`. These gains indicate that the full-vs-masked critic signal is useful when sparse global rewards depend on teammates outside an agent's local observation pathway.

For medium and large tasks, MAE-MAPPO remains generally robust, with gains such as +13.4% in `4x2-2A4B2C`, +12.1% in `3x3-9A`, +11.0% in `3x3-4A1B4C`, +12.8% in `4x4-16A`, and +6.3% in `4x4-8A4B4C`. However, the improvements are not uniform across all base algorithms. MAE-HAPPO and MAE-HATRPO show positive gains in several settings, but also exhibit small degradations in some larger or more heterogeneous tasks. This suggests that MAE provides a useful auxiliary credit signal, but its final effect also depends on how the base actor-critic algorithm incorporates the signal during policy optimisation.

Basilisk We further evaluate MAE on four high-fidelity Basilisk satellite constellation tasks. These tasks introduce orbital dynamics, time-varying visibility windows, pointing constraints, and sparse target-priority rewards. The strongest and most consistent Basilisk results are again obtained by MAE-MAPPO. Compared with MAPPO, MAE-MAPPO improves by +6.8% in `4Ori-4A`, +9.6% in `4Ori-2B2C`, +13.0% in `6Ori-6A`, and +16.6% in `6Ori-2A2B2C`. These results support the usefulness of representation-level marginal credit signals in long-horizon satellite tasking scenarios where each policy observes only local spacecraft and target-visibility information.

The Basilisk results also highlight the scope of MAE. MAE-HAPPO and MAE-HATRPO do not consistently improve over their vanilla counterparts, especially in heterogeneous settings. This should not be attributed to a different MAE computation, since all MAE variants use the same full-vs-masked critic signal. Rather, the difference is likely due to the actor-update side: MAPPO uses a simpler PPO-style

[1]Code is available at `https://github.com/Hidenshadow/HARL_MPE_Satellite`.

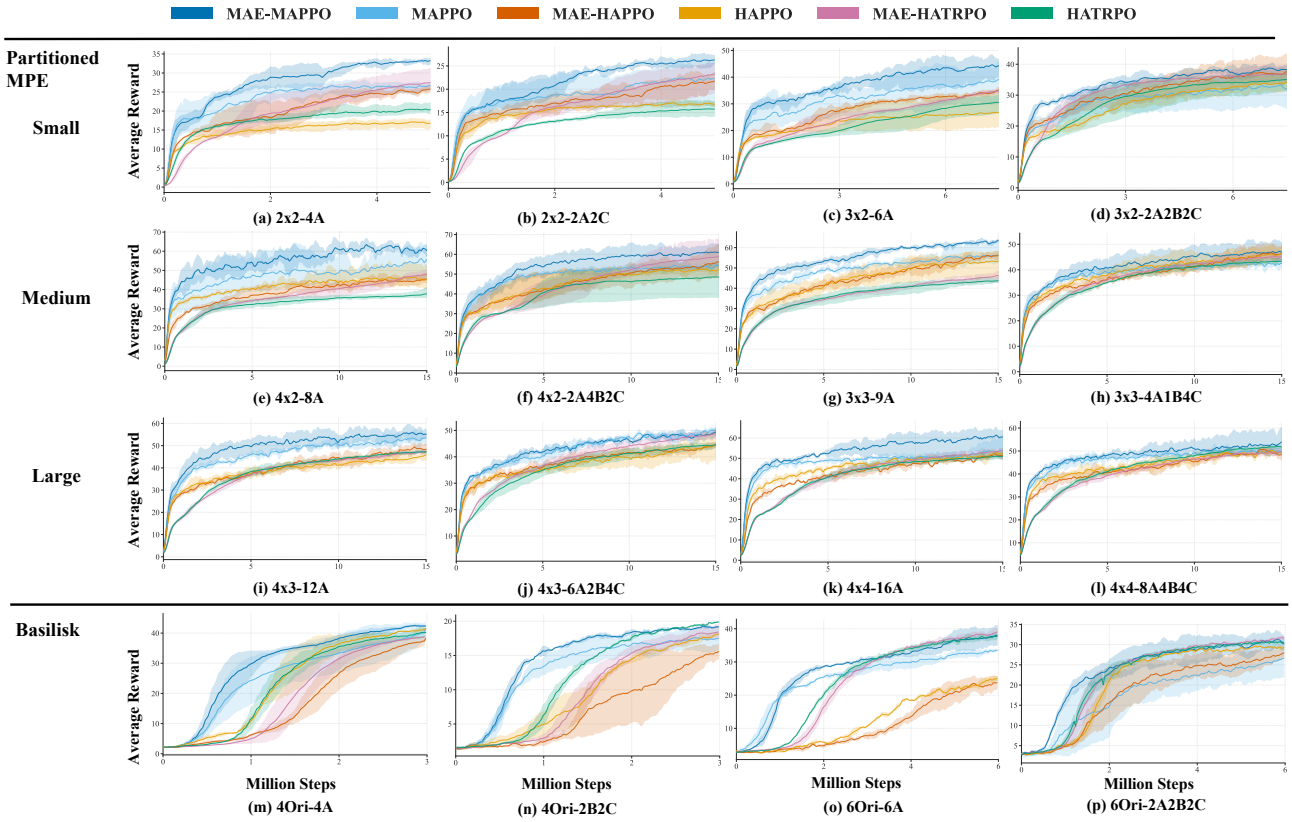


Figure 2: MARL performance of (1) 12 tasks in Partitioned MPE: Average episode rewards for 4-6 agents (a–d), 8-9 agents (e–h), and 12-16 agents (i–l); (2) 4 tasks in Basilisk: Average episode rewards for 4 satellites (m-n) and 6 satellites (o-p).

update with shared policy parameters, while HAPPO and HATRPO use heterogeneous-agent updates and, in the case of HATRPO, trust-region constraints. Therefore, the empirical claim should be read as strongest for MAE-MAPPO and more conditional for other base optimisers.

5.2 ANALYSIS OF COOPERATIVE BEHAVIOUR

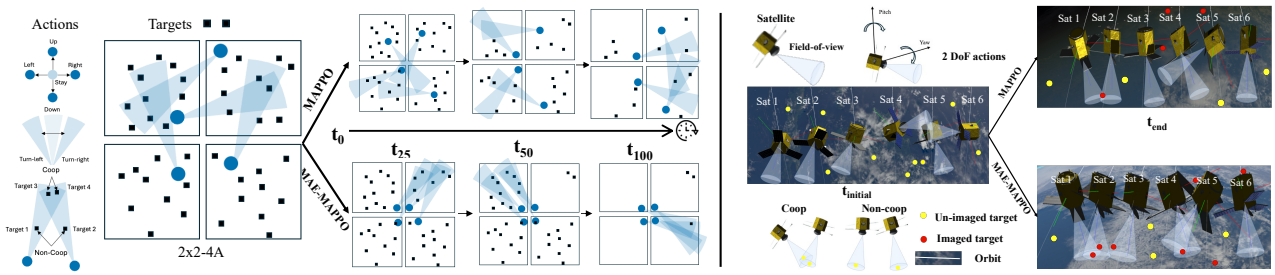
To explicitly contrast learning dynamics under indirect cooperation with standard direct cooperation assumptions, we examine the emergent behaviours of agents during training on two representative tasks: Partitioned MPE (`2x2-4A`) and Basilisk (`6Ori-6A`). As illustrated in Fig. 3, we compare MAPPO and MAE-MAPPO across different training stages to illustrate how structural isolation affects learned behaviours and how MAE-MAPPO can produce more coordinated patterns in these representative tasks.

Partitioned MPE: 2×2-4A Task As shown in Fig. 3(a), we compare agent trajectories across four representative timesteps. At $t = 0$, agents from both methods are randomly initialised. By $t = 25$, MAE-MAPPO agents begin to move toward complementary areas within their respective regions, whereas MAPPO agents show more overlapping and redundant trajectories. At $t = 50$, MAE-MAPPO achieves broader target coverage, suggesting that the MAE signal helps agents form more differentiated behaviours under structural isolation. In contrast, MAPPO agents remain more concentrated in a subset of regions or make less coordinated progress, reflecting the difficulty of assigning credit from sparse global rewards when teammates are not directly observable. By $t = 100$, MAE-MAPPO forms a more organised coverage pattern and completes more task goals within the same episode, while the baseline remains less coordinated.

Basilisk: 6Ori-6A Task For the satellite constellation task, Fig. 3(b) compares the pitch/yaw command distributions at early and late training stages. A key challenge in this environment is learning to orient sensors toward useful Earth-observation opportunities under orbital motion and limited visibility windows. In the early stage, MAPPO exhibits more scattered pointing commands across several satellites, whereas MAE-MAPPO learns more consistent Earth-facing orientations. As training progresses, MAE-MAPPO further spreads its pointing commands across satellites, which suggests improved coverage diversity and reduced redundant observations. By the late stage, MAE-MAPPO shows a more stable and structured distribution of pitch/yaw

Table 1: MARL performance comparisons on Partitioned MPE and Basilisk maps. We report the mean episode reward ± standard deviation over 5 seeds. Percentages in parentheses indicate the relative change of each MAE variant over its corresponding Vanilla baseline. Within each Vanilla–MAE pair, bold indicates the higher mean reward.

Environments	Tasks	MAPPO		HAPPO		HATRPO	
		Vanilla	MAE	Vanilla	MAE	Vanilla	MAE
Direct Cooperation MPE	Vanilla-2x2-4A	22.15 ± 0.66	**22.98 ± 0.19** (+3.8%)	**14.57 ± 0.41**	13.82 ± 0.49 (−5.1%)	**13.82 ± 2.04**	13.43 ± 0.98 (−2.8%)
	Vanilla-2x2-2A2C	**16.89 ± 0.12**	16.53 ± 0.23 (−2.1%)	12.32 ± 0.05	**13.42 ± 0.13** (+8.9%)	10.62 ± 0.64	**11.62 ± 0.28** (+9.4%)
Indirect Cooperation Partitioned MPE	2x2-4A	26.64 ± 0.74	**33.17 ± 0.83** (+24.5%)	16.85 ± 1.11	**25.61 ± 0.68** (+52.1%)	20.37 ± 1.52	**27.38 ± 2.82** (+34.4%)
	2x2-2A2C	22.23 ± 4.58	**26.23 ± 0.99** (+18.0%)	16.81 ± 0.36	**21.59 ± 1.63** (+28.4%)	15.73 ± 1.43	**23.14 ± 3.19** (+47.0%)
Small	3x2-6A	38.51 ± 7.36	**44.10 ± 4.19** (+14.5%)	26.60 ± 6.26	**34.46 ± 0.76** (+29.5%)	30.34 ± 3.56	**34.45 ± 0.58** (+13.6%)
	3x2-2A2B2C	32.51 ± 5.97	**38.26 ± 0.97** (+17.7%)	33.85 ± 3.33	**37.11 ± 4.80** (+9.6%)	34.87 ± 3.15	**36.77 ± 2.33** (+5.4%)
Medium	4x2-8A	55.02 ± 8.19	**60.67 ± 2.32** (+10.3%)	**45.33 ± 6.20**	43.80 ± 0.53 (−3.4%)	37.46 ± 2.36	**47.65 ± 2.38** (+27.2%)
	4x2-2A4B2C	53.82 ± 1.21	**61.01 ± 5.60** (+13.4%)	51.98 ± 1.90	**55.66 ± 7.82** (+7.1%)	48.49 ± 8.96	**58.55 ± 7.79** (+20.7%)
	3x3-9A	55.89 ± 1.68	**62.94 ± 0.69** (+12.1%)	53.18 ± 6.90	**55.89 ± 1.68** (+5.1%)	43.59 ± 1.27	**45.93 ± 2.56** (+5.4%)
	3x3-4A1B4C	42.29 ± 2.34	**46.95 ± 3.75** (+11.0%)	46.07 ± 2.86	**46.29 ± 2.64** (+0.5%)	43.29 ± 0.98	**45.04 ± 1.31** (+4.0%)
Large	4x3-12A	53.14 ± 2.72	**55.24 ± 2.17** (+4.0%)	45.29 ± 1.59	**48.46 ± 1.87** (+7.0%)	**47.13 ± 0.55**	47.07 ± 0.97 (−0.1%)
	4x3-6A2B4C	**49.88 ± 0.20**	48.88 ± 1.56 (−2.0%)	**44.18 ± 5.25**	43.60 ± 0.01 (−1.3%)	44.33 ± 1.10	**47.80 ± 1.41** (+7.8%)
	4x4-16A	53.68 ± 5.07	**60.57 ± 5.30** (+12.8%)	**52.14 ± 1.43**	51.39 ± 1.85 (−1.4%)	50.93 ± 1.70	**53.42 ± 1.90** (+4.9%)
	4x4-8A4B4C	50.00 ± 1.10	**53.15 ± 5.06** (+6.3%)	49.42 ± 0.22	**49.55 ± 0.43** (+0.3%)	**51.85 ± 0.61**	49.44 ± 1.41 (−4.6%)
Basilisk	4Ori-4A	39.46 ± 3.27	**42.16 ± 0.39** (+6.8%)	**40.85 ± 0.37**	36.76 ± 1.91 (−10.0%)	**39.81 ± 1.04**	38.93 ± 0.24 (−2.2%)
	4Ori-2B2C	17.43 ± 0.38	**19.10 ± 0.07** (+9.6%)	**17.72 ± 0.26**	15.02 ± 1.45 (−13.8%)	**19.74 ± 0.14**	18.23 ± 0.95 (−7.7%)
	6Ori-6A	33.38 ± 0.55	**37.73 ± 4.07** (+13.0%)	**24.59 ± 1.12**	23.31 ± 2.87 (−5.2%)	37.56 ± 1.18	**38.53 ± 1.46** (+2.6%)
	6Ori-2A2B2C	26.14 ± 6.43	**30.48 ± 2.44** (+16.6%)	**29.14 ± 0.11**	27.53 ± 1.56 (−5.5%)	30.58 ± 0.27	**31.50 ± 0.50** (+3.0%)



(a) Agents are learning to cooperate in Partitioned-MPE tasks. (b) Satellites are learning to cooperate in Basilisk tasks.

Figure 3: Visualisation of cooperative behaviours. (a) In the Partitioned MPE `2x2-4A` task, MAE-MAPPO agents show broader target coverage than MAPPO across representative timesteps. (b) In the Basilisk `6Ori-6A` task, pitch/yaw command heatmaps suggest that MAE-MAPPO learns more stable Earth-facing pointing behaviours than MAPPO.

commands, while MAPPO remains more variable. These behavioural differences are consistent with the quantitative results, indicating that MAE provides useful auxiliary credit information for coordinated satellite tasking under indirect cooperation.

5.3 EMPIRICAL DIAGNOSTICS

To examine whether the observed training behaviour is consistent with the conditional variance-reduction analysis in Sec. 3, we analyse the statistical characteristics of learning signals across the 16 indirect-cooperation tasks. These diagnostics are not intended as direct proofs of the theoretical results; rather, they provide empirical evidence about whether the learning dynamics follow the trends predicted by the stated assumptions. We consider two complementary quantities: Critic Value Loss, which reflects the stability of value estimation, and Critic Gradient Norm, which reflects the variability of policy-update signals during training.

First, for Theorem 1, we track the Critic Value Loss during the convergence phase as a diagnostic of value-estimation stability. The theorem suggests that full-vs-masked critic evaluations can yield a lower-variance marginal signal when factual and masked critic errors are positively correlated and their marginal variances are not increased. As detailed in Appendix C.3, MAE-based methods generally exhibit lower standard deviations than their corresponding vanilla baselines, with reductions of approximately 30%–60% in many tasks. This trend is consistent with the interpretation that using the same sampled training batch and shared critic parameters can stabilise the full-vs-masked value difference, especially in structurally isolated tasks where independent counterfactual rollouts would introduce additional sampling noise.

Second, for Theorem 2, we examine gradient-norm statistics during training. In indirect-cooperation settings, sparse global rewards may depend on teammates that are not observable from an agent's local history, which can increase

the residual noise in the advantage signal used for policy updates. As presented in Appendix C.4, vanilla algorithms often show larger gradient fluctuations in such settings. In comparison, MAE-based methods tend to produce lower or more stable gradient norms, particularly in large-scale heterogeneous scenarios such as `4x4-8A4B4C` and `6Ori-2A2B2C`. These diagnostics support the view that MAE can reduce residual update variability by providing an auxiliary marginal credit signal, while remaining subject to the critic-quality, representation-factorisation, and advantage-decomposition assumptions discussed in Sec. 3.

5.4 ADDITIONAL RESULTS

We further evaluate representative value-based and game-theoretic MARL algorithms, including COMA [Foerster et al., 2018], Shapley-Q [Wang et al., 2020b], Nucleolus-Q [Li et al., 2025], and HAD3QN [Zhong et al., 2024]. As detailed in Appendix C.5, these methods struggle in the indirect-cooperation tasks considered here. Their weaker performance is likely due to the difficulty of propagating sparse global rewards through bootstrapped value backups or action/coalition-level credit signals when agents are structurally isolated. Similar challenges appear in Basilisk, where long horizons, orbital motion, and time-varying visibility windows further increase credit-assignment difficulty. Applying MAE to value-based methods through MAE-Q provides only limited improvements, suggesting that the proposed full-vs-masked marginal signal is not a universal replacement for value backup or coalition-value computation. Instead, the results indicate that MAE is most effective when used as an auxiliary credit term in centralised-critic policy-gradient methods. In particular, MAPPO-style actor-critic training benefits more consistently from MAE because the critic-derived marginal signal can be added directly to GAE without replacing the base policy-gradient estimator. COMA also underperforms MAPPO on representative tasks such as `2x2-4A` and `2x2-2A2C`, with slower learning and higher variability; therefore, we report it in the appendix rather than the main comparison table.

Finally, to isolate the contribution of each component, we conduct an ablation study on the `2x2-4A` task (Appendix C.6). The results show that MAE is the main contributor to final performance improvement, while PBRS and RND play complementary roles in sparse-reward learning. Specifically, PBRS improves training stability by providing a smoother auxiliary reward signal, whereas RND accelerates early-stage exploration by encouraging visits to novel states. However, neither PBRS nor RND directly addresses the structural credit-assignment ambiguity caused by indirect cooperation. This separation supports our interpretation that MAE is the core credit-assignment mechanism, while PBRS and RND are auxiliary training aids that improve optimisation and exploration efficiency.

6 CONCLUSION

We introduced MAE for cooperative MARL under indirect cooperation. MAE uses full-vs-masked critic evaluations to provide a representation-level auxiliary credit signal when local histories contain limited information about reward-relevant teammates. By adding this signal to standard GAE as a stop-gradient term, MAE preserves the underlying centralised-critic policy-gradient update while providing additional marginal contribution information during training. Across Partitioned MPE and Basilisk tasks, the results show that MAE is most beneficial in structurally isolated, reward-coupled settings. The strongest and most consistent gains are observed when MAE is integrated with MAPPO, while results with HAPPO and HATRPO are more mixed, indicating that the effectiveness of the same critic-side marginal signal also depends on the actor-update procedure. Empirical diagnostics further suggest that MAE can improve value-estimation stability and reduce residual update variability under the stated assumptions. Overall, these findings support MAE as a simple and practical credit-shaping mechanism for structurally isolated cooperative tasks. Future work will study more distribution-preserving masking mechanisms, richer critic representations, and broader applications in large-scale robotic and satellite-constellation systems.

Acknowledgements

This work has been supported by the SmartSat CRC, whose activities are funded by the Australian Government's CRC Program. This study is also partially supported by Australian Research Council (ARC) Projects DE220100265, DP250103612, and LP240200636.

References

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations*, 2019.

Daniel S Drew. Multi-agent systems for search and rescue applications. *Current Robotics Reports*, 2021.

Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 2016.

Jakob Foerster, Gregory Farquhar, Tadas Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *AAAI*, 2018.

Sven Gronauer and Klaus Diepold. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review*, 2022.

Andrew Harris, Trace Valade, Thibaud Teil, and Hanspeter Schaub. Generation of spacecraft operations procedures using deep reinforcement learning. *Journal of Spacecraft and Rockets*, 59(2):611–626, 2022.

Matthew J Hausknecht and Peter Stone. Deep recurrent q-learning for partially observable mdps. In *AAAI fall symposia*, 2015.

Adam Herrmann, Mark A Stephenson, and Hanspeter Schaub. Single-agent reinforcement learning for scalable earth-observing satellite constellation operations. *Journal of Spacecraft and Rockets*, 2024.

Adam P Herrmann and Hanspeter Schaub. Monte carlo tree search methods for the earth-observing satellite scheduling problem. *Journal of Aerospace Information Systems*, 19(1):70–82, 2022.

Maximilian Igl, Luisa Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. Deep variational reinforcement learning for pomdps. In *International conference on machine learning*, 2018.

Patrick W Kenneally, Scott Piggott, and Hanspeter Schaub. Basilisk: A flexible, scalable and modular astrodynamics simulation framework. *Journal of aerospace information systems*, 2020.

Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. Trust region policy optimisation in multi-agent reinforcement learning. In *International Conference on Learning Representations*, 2022.

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*, 2025.

Z Li, Y Yang, and H Cheng. Efficient multi-agent cooperation: Scalable reinforcement learning with heterogeneous graph networks and limited communication. *Knowledge-Based Systems*, 2024.

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

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

Xiaoteng Ma, Yiqin Yang, Chenghao Li, Yiwen Lu, Qianchuan Zhao, and Jun Yang. Modeling the interaction between agents in cooperative multi-agent reinforcement learning. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pages 853–861, 2021.

Hesam Mosalli and Amir G Aghdam. A distributed reinforcement learning strategy to maximize coverage in a hybrid heterogeneous sensor network. In *2024 IEEE 63rd Conference on Decision and Control (CDC)*, pages 2475–2480. IEEE, 2024.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pages 278–287. Citeseer, 1999.

Frans A Oliehoek, Christopher Amato, et al. *A concise introduction to decentralized POMDPs*. Springer, 2016.

Johan Peralez, Aurelien Delage, Jacopo Castellini, Rafael F Cunha, and Jilles S Dibangoye. Optimally solving simultaneous-move dec-pomdps: The sequential central planning approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Tabish Rashid, Gregory Farquhar, Bei Peng, and Shimon Whiteson. Weighted qmix: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 33:10199–10210, 2020a.

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*, 2020b.

Mark A Stephenson and Hanspeter Schaub. Bsk-rl: Modular, high-fidelity reinforcement learning environments for spacecraft tasking. In *75th International Astronautical Congress*, 2024.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*, 2017.

Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. Qplex: Duplex dueling multi-agent q-learning. *arXiv preprint arXiv:2008.01062*, 2020a.

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*, 2020b.

David H Wolpert and Kagan Tumer. Optimal payoff functions for members of collectives. *Advances in Complex Systems*, 4(02n03):265–279, 2001.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 2022.

Yifan Zhong, Jakub Grudzien Kuba, Xidong Feng, Siyi Hu, Jiaming Ji, and Yaodong Yang. Heterogeneous-agent reinforcement learning. *Journal of Machine Learning Research*, 2024.

Guangchong Zhou, Zeren Zhang, and Guoliang Fan. Air: Unifying individual and collective exploration in cooperative multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Shaohao Zhu, Jiacheng Zhou, Anjun Chen, Mingming Bai, Jiming Chen, and Jinming Xu. Maexp: A generic platform for rl-based multi-agent exploration. In *2024 IEEE International Conference on Robotics and Automation*, 2024.

Itai Zilberstein, Ananya Rao, Matthew Salis, and Steve Chien. Decentralized, decomposition-based observation scheduling for a large-scale satellite constellation. *Journal of Artificial Intelligence Research*, 82:169–208, 2025.

A THEORETICAL PROOFS

In this appendix, we provide the derivations for the variance-reduction results discussed in the main text. These results should be interpreted under the stated assumptions. They do not establish universal convergence or optimality guarantees for arbitrary multi-agent reinforcement learning dynamics.

A.1 PROOF OF THEOREM 1

Let the target marginal quantity be denoted by Δ_i. A simulation-based Monte Carlo counterfactual estimator can be written as

$$\hat{\Delta}_i^{\mathrm{MC}} = \Delta_i + \epsilon_F - \epsilon_{CF}, \tag{6}$$

where ϵ_F and ϵ_{CF} are the factual and counterfactual rollout errors, respectively. Since the factual and counterfactual terms are estimated from independent rollouts, we assume

$$\mathrm{Cov}(\epsilon_F, \epsilon_{CF}) = 0. \tag{7}$$

Let

$$\mathrm{Var}(\epsilon_F) = \sigma_F^2, \qquad \mathrm{Var}(\epsilon_{CF}) = \sigma_{CF}^2. \tag{8}$$

Then the variance of the simulation-based estimator is

$$\mathrm{Var}(\hat{\Delta}_i^{\mathrm{MC}}) = \sigma_F^2 + \sigma_{CF}^2. \tag{9}$$

For MAE, the factual and masked values are evaluated on the same sampled training batch using the same centralised critic parameters. We write the MAE estimator as

$$\hat{\Delta}_i^{\mathrm{MAE}} = \Delta_i + \delta_F - \delta_M, \tag{10}$$

where δ_F and δ_M are the factual and masked critic estimation errors. Assume that the shared critic induces positive error correlation:

$$\mathrm{Cov}(\delta_F, \delta_M) = c > 0. \tag{11}$$

Also assume that the marginal error variances are no larger than those of the independent simulation baseline:

$$\mathrm{Var}(\delta_F) \le \sigma_F^2, \qquad \mathrm{Var}(\delta_M) \le \sigma_{CF}^2. \tag{12}$$

Then

$$\begin{aligned} \mathrm{Var}(\hat{\Delta}_i^{\mathrm{MAE}}) &= \mathrm{Var}(\delta_F - \delta_M) \\ &= \mathrm{Var}(\delta_F) + \mathrm{Var}(\delta_M) - 2\mathrm{Cov}(\delta_F, \delta_M) \\ &\le \sigma_F^2 + \sigma_{CF}^2 - 2c \\ &< \sigma_F^2 + \sigma_{CF}^2 = \mathrm{Var}(\hat{\Delta}_i^{\mathrm{MC}}). \end{aligned} \tag{13}$$

Therefore, under positively correlated critic errors and no larger marginal error variance, MAE has lower variance than the independent simulation-based counterfactual estimator. □

A.2 PROOF OF THEOREM 2

Let

$$z_i(t) = \nabla_{\theta_i} \log \pi_{\theta_i}(a_{i,t} \mid o_{i,t}) \tag{14}$$

denote the policy score function of agent i. Suppose the global advantage used by the centralised-critic policy-gradient update can be approximately decomposed as

$$\hat{A}_i^{\mathrm{GAE}}(t) = A_i^{\mathrm{loc}}(t) + \xi_{-i}(t), \tag{15}$$

where $A_i^{\mathrm{loc}}(t)$ is the local contribution of agent i, and $\xi_{-i}(t)$ is a zero-mean latent component induced by structurally isolated teammates.

Suppose the MAE signal approximates the local contribution:

$$M_i(t) = A_i^{\mathrm{loc}}(t) + \epsilon_i(t), \tag{16}$$

where $\epsilon_i(t)$ is a zero-mean approximation error. The score-weighted global signal is

$$z_i(t)\hat{A}_i^{\mathrm{GAE}}(t) = z_i(t)A_i^{\mathrm{loc}}(t) + z_i(t)\xi_{-i}(t), \tag{17}$$

whereas the score-weighted MAE signal is

$$z_i(t)M_i(t) = z_i(t)A_i^{\mathrm{loc}}(t) + z_i(t)\epsilon_i(t). \tag{18}$$

Assume that $\xi_{-i}(t)$ and $\epsilon_i(t)$ are uncorrelated with the score function conditional on the local history, and that

$$\mathrm{Var}(\epsilon_i(t)) < \mathrm{Var}(\xi_{-i}(t)). \tag{19}$$

Then the residual variance associated with the MAE-weighted direction is lower than that associated with the latent teammate component. Equivalently,

$$\mathrm{Var}[z_i(t)M_i(t)] < \mathrm{Var}\left[z_i(t)\hat{A}_i^{\mathrm{GAE}}(t)\right]. \tag{20}$$

Therefore, under the stated decomposition and approximation assumptions, MAE can reduce the gradient noise associated with unobservable teammates. □

A.3 DETAILED IMPLEMENTATION

Our implementation is built upon the HARL library [Zhong et al., 2024]. MAE is implemented as a modular critic-side computation and can be added to MAPPO, HAPPO, and HATRPO without changing the rollout collection procedure.

Network Architecture. All actor-critic methods use multi-layer perceptron networks. The actor contains two hidden layers with ReLU activations and outputs a categorical distribution for discrete actions. The centralized critic receives the joint critic input and outputs a scalar value estimate. For MAE, the masked input x_t^{-i} is constructed by zeroing the agent-specific input block corresponding to agent i while keeping the remaining critic input unchanged. The factual and masked critic values are evaluated through batched full-vs-masked forward computation.

Auxiliary Training Signals. Potential-Based Reward Shaping (PBRS) [Ng et al., 1999] and Random Network Distillation (RND) [Burda et al., 2019] are used only as optional algorithm-side auxiliary signals for sparse-reward settings. They are not part of the MAE marginal estimator and do not redefine the environment reward. Unless otherwise stated, MAE refers specifically to the full-vs-masked critic value difference in Eq. (2).

Policy-Update Signal. The main policy-update advantage is

$$\hat{A}_i^{\mathrm{total}}(t) = \hat{A}_i^{\mathrm{GAE}}(t) + \lambda \cdot \mathrm{sg}[M_i(t)]\,. \tag{21}$$

When auxiliary PBRS or RND signals are enabled, they are treated as separate training aids and are reported in the corresponding ablations.

Hyperparameters. Table 2 lists the common hyperparameters used in our experiments. Unless otherwise stated, we use Adam with learning rate 5×10^{-4}, discount factor $\gamma = 0.99$, GAE parameter $\zeta = 0.95$, PPO clipping range 0.2, and MAE weight $\lambda = 0.15$.

B DETAILED EXPERIMENTAL SETTINGS

B.1 PARTITIONED MPE ENVIRONMENT

To evaluate our method under indirect cooperation, we design a customised version of the Multi-Agent Particle Environment (MPE) Lowe et al. [2017], termed Partitioned MPE. Unlike the standard MPE setup, Partitioned MPE introduces structured

Table 2: Hyperparameter settings used in the experiments.

Hyperparameter	**Value**
Optimizer	Adam
Learning Rate (Actor/Critic)	5×10^{-4}
Discount Factor (γ)	0.99
GAE Parameter (ζ)	0.95
Batch Size	1000 (Small/Med), 3200 (Large/BSK)
Hidden Layer Size	[128, 128]
Activation Function	ReLU
PPO Clip Range	0.2
PPO Epochs	5
MAE Weight (λ)	**0.15** (`coop_bonus_weight`)
PBRS Weight (α)	0.25 (`potential_weight`)
RND Weight (β)	0.10 (`int_rnd_weight`)

spatial partitions, heterogeneous observability, and sparse global rewards to simulate realistic cooperation under limited sensing and physical isolation. Each agent is confined to a fixed non-overlapping spatial region and can only act within this bounded exploration space. Its local perception is further limited by its field of view (FoV) and sensing radius. Global rewards are only granted when agents from distinct regions successfully cooperate to monitor distributed targets jointly for a sustained period. The environment features diverse map configurations, including both symmetric and asymmetric designs, with varying agent counts, region layouts, and sensing capabilities. These settings simulate practical multi-agent constraints such as irregular sensor coverage, role specialisation, and decentralised information.

B.1.1 Observation and operational constraints

Each agent i is restricted to a fixed, non-overlapping *exploration space* $\mathcal{B}_i \subset \mathcal{X}$, corresponding to its designated grid region. The agent's movement and actions are confined to $\mathcal{B}_i$ for the entire episode. Different exploration spaces are disjoint:

$$\mathcal{B}_i \cap \mathcal{B}_j = \emptyset, \quad i \neq j.$$

Perception is constrained by two parameters:

- Sensing radius r_i: the maximum distance an agent can perceive from its current location.
- Viewing angle θ_i: the aperture of an agent's field-of-view cone, aligned with its heading direction.

The exploration space $\mathcal{B}_i$ defines where agent i can move and act, while r_i and θ_i define its time-varying local sensing footprint inside that region. The agent's local observation is computed by applying a region- and FoV-restricted projection to the global state, filtering out entities outside its accessible region or outside its current sensing footprint. In heterogeneous settings, agents differ in r_i and θ_i, leading to asymmetric observability across the team.

Formally, the local observation set of agent i at state s_t is defined as:

$$\mathcal{O}_i(s_t) = \left\{ e \in s_t : p_e \in \mathcal{B}_i,\ \mathrm{dist}(p_{i,t}, p_e) \leq r_i,\ \alpha_{i,t}(e) \leq \frac{1}{2}\theta_i \right\},$$

where $p_{i,t}$ denotes the position of agent i at time t, p_e denotes the position of an observable entity e such as a landmark, and $\alpha_{i,t}(e)$ denotes the angle between agent i's current heading direction and the vector from $p_{i,t}$ to p_e. This definition separates the fixed operating region $\mathcal{B}_i$ from the time-varying FoV constraint.

To support dynamic perceptual control, we extend the original MPE action space by introducing *field-of-view (FoV) rotation actions*. In addition to standard planar movements, each agent is equipped with discrete control actions to adjust its viewing direction. At each timestep, an agent can rotate its FoV by $\pm 15^\circ$, allowing it to scan its local region more effectively under limited observability. Formally, the action space of each agent is extended as:

$$A_i = A_i^{\mathrm{move}} \times A_i^{\mathrm{rotate}},$$

where:

- $A_i^{\text{move}} = \{\text{up}, \text{down}, \text{left}, \text{right}, \text{stay}\}$ denotes the planar movement actions;
- $A_i^{\text{rotate}} = \{-15°, 0°, +15°\}$ specifies discrete FoV angle adjustments relative to the current viewing direction.

B.1.2 Reward structure

The cooperative objective is to monitor spatially distributed landmarks that represent observation targets. A landmark is deemed *completed* only when it is simultaneously observed by at least two agents from distinct exploration regions for five consecutive steps. Upon success, a global sparse reward is issued, and the landmark is removed from the environment for the remainder of the episode. No local or intermediate rewards are given. This design enforces long-horizon, decentralised coordination under operational constraints, where agents must navigate partial observability, sparse feedback, and limited mobility to achieve cooperative success.

Let $L = \{l_1, l_2, \ldots, l_m\}$ denote the set of landmarks. Define the set of agents observing landmark l_k at time t as:

$$\mathcal{V}_k(t) = \{i \in \mathcal{N} : l_k \in \mathcal{O}_i(s_t)\}.$$

A landmark l_k is considered completed at step t if, for each step in the previous five-step window, there exist at least two agents from distinct exploration regions that simultaneously observe it:

$$C_k(t) = 1 \iff \forall \tau \in \{t-4, t-3, t-2, t-1, t\},\ \exists i, j \in \mathcal{V}_k(\tau) \text{ such that } i \neq j \text{ and } \mathcal{B}_i \neq \mathcal{B}_j.$$

The global reward function is:

$$r_t = \begin{cases} 1, & \text{if there exists } l_k \in L \text{ such that } C_k(t) = 1, \\ 0, & \text{otherwise.} \end{cases}$$

Table 3: Partitioned MPE configurations across different sizes, with symmetric and asymmetric setups.

Size	Map	Targets	Grid Size	Agent Type	FOV Angle	Radius	Episode Length
Small	2x2-4A	40	2×2	All	15°	1.25	100
	2x2-2A2C	40	2×2	Agents 0,3	15°	1.25	
				Agents 1,2	10°	1.5	
	3x2-6A	60	2×3	All	15°	1.25	125
	3x2-2A2B2C	60	2×3	Agents 0,5	15°	1.25	
				Agents 1,4	10°	1.5	
				Agents 2,3	30°	1.0	
Medium	4x2-8A	80	2×4	All	15°	1.25	150
	4x2-2A4B2C	80	2×4	Agents 0,7	15°	1.25	
				Agents 1,6	10°	1.5	
				Agents 2–5	30°	1.0	
	3x3-9A	90	3×3	All	15°	1.25	150
	3x3-4A1B4C	90	3×3	Agents 1,3,5,7	15°	1.25	
				Agents 0,2,6,8	10°	1.5	
				Agent 4	30°	1.0	
Large	4x3-12A	96	3×4	All	15°	1.25	150
	4x3-6A2B4C	96	3×4	Agents 1,3,5,6,8,10	15°	1.25	
				Agents 0,2,9,11	10°	1.5	
				Agents 4,7	30°	1.0	
	4x4-16A	112	4×4	All	15°	1.25	150
	4x4-8A4B4C	112	4×4	Agents 1,2,4,7,8,11,13,14	15°	1.25	
				Agents 0,3,12,15	10°	1.5	
				Agents 5,6,9,10	30°	1.0	

B.1.3 Task configurations

To systematically evaluate the cooperative performance, we design a suite of 12 multi-agent tasks with increasing scale and heterogeneity. Each task is named following a standardised format: [Grid-size]-[Agent Type]. The configuration parameters, including agent types, FOV angles, detection radii, and episode lengths, are summarised in Table 3. We categorise tasks into three sizes: small, medium, and large, based on the number of agents and spatial partitions. Across all configurations, indirect cooperation is induced by the fixed spatial partition: agents operate in disjoint regions and cannot directly observe teammates outside their local information pathway, while the shared reward depends on cross-region target observation. The asymmetric variants further introduce heterogeneous sensing parameters, allowing us to evaluate whether credit assignment remains stable under sensor diversity and role imbalance.

Small tasks (4–6 agents) The small-scale tasks include the `2x2-4A` and `3x2-6A` maps, each with symmetric (homogeneous) and asymmetric (heterogeneous) observability versions. In the homogeneous settings, all agents share identical sensor parameters (FOV = 15°, radius = 1.25). In contrast, heterogeneous variants, `2x2-2A2C` and `3x2-2A2B2C`, assign diverse roles with different angular coverages and detection ranges. This design enables us to examine the impact of partial role heterogeneity in low-agent settings.

Medium tasks (8–9 agents) The medium-scale tasks expand both spatial layout and agent count. For instance, `4x2-8A` and `3x3-9A` represent symmetric baselines, while `4x2-2A4B2C` and `3x3-4A1B4C` introduce complex asymmetries. These maps contain agents with FOV angles ranging from 10° to 30°, mimicking realistic sensor diversity. The increased density of agents and targets adds coordination complexity and makes learning under physical isolation and local partial observability more challenging.

Large tasks (12–16 agents) The largest-scale tasks, `4x3-12A` and `4x4-16A`, feature dense agent populations and extended grid layouts. Their heterogeneous variants (`4x3-6A2B4C`, `4x4-8A4B4C`) incorporate diverse perception parameters such as varying field-of-view angles and sensing radii. These tasks are designed to evaluate the scalability and robustness of each algorithm under strict observability constraints and increased agent diversity.

B.2 BASILISK ENVIRONMENT

B.2.1 Simulator overview

Basilisk is a high-fidelity astrodynamics simulation framework for modelling spacecraft dynamics, guidance, navigation, control, and mission-level tasking [Kenneally et al., 2020, Harris et al., 2022, Stephenson and Schaub, 2024]. In this paper, we use Basilisk to construct cooperative Earth-observation tasks for satellite constellations. Each satellite is modelled as an autonomous agent that moves according to orbital dynamics and controls its sensor pointing direction to image ground targets. Compared with abstract particle environments, Basilisk introduces realistic spatio-temporal constraints: satellites move along orbital trajectories, target visibility changes over time, sensing is limited by field-of-view and elevation constraints, and successful observations require correct timing and pointing.

The Basilisk environment is well suited to our indirect-cooperation setting. Satellite tasking and Earth-observation scheduling are naturally long-horizon, time-window-constrained decision problems, where spacecraft must coordinate observations under orbital visibility, pointing, and resource constraints [Herrmann and Schaub, 2022, Herrmann et al., 2024]. Satellites may contribute to a shared Earth-observation objective without being able to directly observe or communicate the full state of other satellites during execution. Each decentralised policy only receives local spacecraft information and locally visible target information, while the team reward depends on coordinated imaging outcomes across the constellation. Thus, the isolation is not induced by fixed ground partitions as in Partitioned MPE, but by distinct orbital planes, time-varying visibility windows, limited sensor footprints, and local spacecraft observations. The centralised critic used during training can access the joint critic input, whereas decentralised policies are executed using only local observations.

B.2.2 Observation and operational constraints

Each satellite agent follows a deterministic orbit defined by Classical Orbital Elements (COEs) and is equipped with constrained pointing control. The orbital parameters used in different configurations are summarised in Table 4, where all satellites share the same semi-major axis ($a = 6798.3$ km) and near-zero eccentricity, but differ in inclination to produce

Table 4: Basilisk orbital configurations for the four-satellite and six-satellite constellations.

Config.	Satellite	a (km)	e	i	Ω	ω	ν_0
4-sat	Sat1	6798.3	0.000691	-2°	0	0	0
4-sat	Sat2	6798.3	0.000691	2°	0	0	0
4-sat	Sat3	6798.3	0.000691	-6°	0	0	0
4-sat	Sat4	6798.3	0.000691	6°	0	0	0
6-sat	Sat1	6798.3	0.000691	2°	0	0	0
6-sat	Sat2	6798.3	0.000691	-2°	0	0	0
6-sat	Sat3	6798.3	0.000691	6°	0	0	0
6-sat	Sat4	6798.3	0.000691	-6°	0	0	0
6-sat	Sat5	6798.3	0.000691	10°	0	0	0
6-sat	Sat6	6798.3	0.000691	-10°	0	0	0

varied ground coverage across latitude bands.

- Orbital motion: Satellites move along pre-defined Keplerian orbits with fixed COEs. Inclination angles range from $\pm 2^\circ$ to $\pm 10^\circ$, generating diverse visibility footprints over time.
- Pointing actions: Each satellite has 2-DoF pointing control over pitch and yaw. For each axis, the discrete pointing adjustment is selected from $\{-5^\circ, 0^\circ, +5^\circ\}$, yielding nine discrete pointing commands.
- Field-of-view (FoV): Each sensor defines a fixed conical viewing area, with aperture angle determined by the agent type. A target is considered visible only if it lies within the projected FoV footprint and satisfies the minimum elevation threshold.

Each decentralised policy receives local observations composed of:

- Satellite state: orbital position, velocity, pitch-yaw pointing, and angular velocity.
- Visible targets: ground targets that are currently within the satellite's FoV projection and above the elevation threshold.

Observation constraints. A ground target g_k is deemed *visible* to satellite agent i at time t if it lies within the projected sensor footprint and satisfies the elevation constraint:

$$\mathrm{vis}_{i,k}(t) = \begin{cases} 1, & \text{if } g_k \in \mathrm{FoV}_i(t) \text{ and } \mathrm{elev}(i, g_k, t) \geq \delta, \\ 0, & \text{otherwise.} \end{cases}$$

Here, $\mathrm{FoV}_i(t)$ denotes the time-varying ground footprint of satellite i under its current orbital position and pointing direction, and δ is the minimum elevation threshold. The detailed geometric visibility computation is handled by the Basilisk simulator.

B.2.3 Reward structure

The cooperative task is to image geographically distributed ground targets using the satellite constellation. Each target is assigned a fixed priority weight and is completed only if it is successfully imaged by at least two distinct satellites within a 5-minute temporal window. A valid imaging attempt requires the target to be visible to the satellite under the current FoV and elevation constraints and the satellite to execute the corresponding imaging action. The native Basilisk task reward is sparse and issued only upon successful joint imaging of previously uncompleted targets; environment-side dense reward shaping is not used.

Let $G = \{g_1, \ldots, g_K\}$ denote the set of ground targets, and let $w_k \in [0.1, 1.0]$ denote the priority weight of target g_k. Define

$$\mathrm{img}_{i,k}(t) = 1$$

if satellite i successfully images target g_k at time t, and $\mathrm{img}_{i,k}(t) = 0$ otherwise. For a temporal window of length Δt, define the set of satellites that have imaged target g_k within the current window as

$$\mathcal{S}_k(t) = \left\{ i \in \mathcal{N} : \exists t' \in [t - \Delta t, t] \text{ such that } \mathrm{img}_{i,k}(t') = 1 \right\}.$$

A target is completed at time t if it has not been completed before and at least two distinct satellites have imaged it within the temporal window:

$$C_k(t) = 1 \quad \Longleftrightarrow \quad \text{done}_k(t) = 0 \text{ and } |\mathcal{S}_k(t)| \geq 2.$$

The global reward is then

$$r_t = \sum_{k=1}^{K} w_k C_k(t).$$

Once a target is completed, it is removed from further reward consideration, so no repeated reward is issued for the same target. In our experiments, we set $\Delta t = 5$ minutes. This design requires satellites to coordinate their pointing and imaging schedules across distinct orbital planes despite only observing local spacecraft and target-visibility information during execution.

B.2.4 Task configurations

Basilisk considers four simulated Earth observation task scenarios of increasing complexity, systematically varying the number of satellites, orbital planes, and sensing heterogeneity. Each scenario is named following a standardised format: *[Orbit Count]Ori-[Agent Type]*, where the suffix reflects whether agents are homogeneous (e.g., "4A") or heterogeneous (e.g., "2B2C"). Table 5 summarises the key parameters. All tasks share consistent constraints on observation radius (500 km) and minimum target elevation angle (35°), while differing in the number of satellites, field-of-view (FOV) angles, and agent roles.

Table 5: Basilisk task configurations with varying numbers of orbits, satellites, targets, and sensing capabilities.

Map	Targets	Orbital Number	Agent Type	FOV Angle	Obs. Sat Radius (km)	Obs. Target Elev.	Episode Length
`4Ori-4A`	100	4	All	15°	500	35°	315
`4Ori-2B2C`	100	4	Agents 0,2	10°	500	35°	315
			Agents 1,3	20°	500	35°	
`6Ori-6A`	150	6	All	15°	500	35°	315
`6Ori-2A2B2C`	150	6	Agents 0,3	15°	500	35°	315
			Agents 1,4	10°	500	35°	
			Agents 2,5	20°	500	35°	

4Ori-4A A homogeneous configuration with four satellites on four distinct orbits. All agents have identical capabilities, each equipped with a 15° field of view (FOV), and operate under the same system dynamics.

4Ori-2B2C A heterogeneous setting where satellites are divided into two groups. Agents 0 and 2 use narrow 10° FOV sensors, while Agents 1 and 3 use wider 20° FOV sensors, introducing sensing asymmetry.

6Ori-6A A larger homogeneous configuration with six satellites across six orbits. All agents use 15° FOV sensors, increasing the scale of coordination while maintaining uniform sensing.

6Ori-2A2B2C A heterogeneous configuration with six satellites across six orbits. Agents 0 and 3 use 15° FOV sensors, Agents 1 and 4 use narrower 10° FOV sensors, and Agents 2 and 5 use wider 20° FOV sensors. This setup introduces three sensing roles while keeping two satellites in each role.

Each episode in Basilisk spans 315 simulation steps, which approximately corresponds to a full orbital period of a typical low Earth orbit (LEO) satellite. This duration ensures that every satellite completes a full revolution around the Earth, thereby experiencing one complete cycle of target visibility and occlusion. Structuring the episode around a full orbit allows for effective evaluation of long-horizon cooperation strategies. It also guarantees that every satellite has at least one opportunity to observe most global targets, making performance across different configurations temporally comparable.

B.3 TRAINING CONFIGURATION AND HYPERPARAMETERS

All MARL algorithms are trained under a consistent set of hyperparameters to ensure fairness and reproducibility. We use PPO with GAE enabled. Both actor and critic are trained for 5 epochs per update using one mini-batch. The actor and critic

share an MLP backbone with hidden sizes of 128. Parameter sharing and fixed agent ordering are used across all runs. This setup ensures that performance differences are primarily attributable to algorithmic design rather than tuning discrepancies. Table 6 details the full configuration.

Table 6: Unified hyperparameter configuration used in MARL.

Category	Hyperparameter Setting
Seed & Device	Seed = 1, CUDA = True, Threads = 4
Training Steps	Env Steps = 15M, Episode Length = 315
Parallelism	Rollout Threads = 30, Eval Threads = 10
Network	MLP [128, 128], Activation = ReLU
Optimisation	LR = 5e-4, Critic LR = 5e-4, Adam Eps = 1e-5
Initialisation	Orthogonal, Gain = 0.01
Normalisation	Feature Norm = True, ValueNorm = True
Recurrent Policy	Disabled
PPO Epochs	Actor = 5, Critic = 5
GAE	$\gamma = 0.99$, $\lambda = 0.95$
Loss	Clip = 0.2, Entropy Coef = 0.01, Huber $\delta = 10.0$
Gradient Clipping	Max Grad Norm = 10.0
Action Aggregation	Method = prod
Misc	Share Param = True, Fixed Order = True

B.4 WALL-CLOCK TRAINING TIME (REPORTING NOTES)

Table 7 reports wall-clock training time (in hours) for each algorithm across the four environment groups and their map setups. Columns are grouped by environment: *Partitioned MPE (Small)* with `2x2`, `2x3`; *Partitioned MPE (Medium)* with `3x3`, `2x4`; *Partitioned MPE (Large)* with `4x4`, `3x4`; and *Basilisk* with `4ori`, `6ori`. Rows list the base methods (MAPPO, HAPPO, HATRPO) and their MAE-augmented counterparts (MAE-MAPPO, MAE-HAPPO, MAE-HATRPO).

All runs use the same training budget (identical number of environment steps/updates), batch sizes, and rollout lengths, under the hardware/software stack. Wall-clock time is measured from the start of optimisation to the end of the budget, excluding evaluation-only episodes and checkpoint I/O. The table is intended to provide a direct, like-for-like view of training time across algorithms and map setups, complementing the performance results.

C ADDITIONAL EXPERIMENTAL RESULTS

C.1 LEARNING PERFORMANCE IN PARTITIONED MPE ENVIRONMENT

We present the complete quantitative results across 12 Partitioned MPE tasks and 2 unconstrained Vanilla MPE ablation tasks. We report the mean episode returns (mean $\pm$ std) of all algorithms under both homogeneous and heterogeneous settings, covering agent scales from 4 to 16. Each task is defined on a structured grid-based cooperative environment with fixed spatial partitions, sparse global rewards, and varying degrees of sensing heterogeneity.

Across many tasks, MAE-MAPPO shows the most consistent improvements over its vanilla counterpart, particularly in small- and medium-scale Partitioned MPE settings. MAE-HAPPO and MAE-HATRPO also improve performance in several tasks, but their behaviour is more mixed, especially in larger or more heterogeneous environments. This trend is consistent with the main results: the same critic-side MAE signal can interact differently with different actor-update procedures. MAPPO uses a simpler PPO-style update with shared policy parameters, while HAPPO and HATRPO rely on heterogeneous-agent updates and, in the case of HATRPO, trust-region constraints.

Overall, these results support a scoped interpretation of MAE. The method is most reliable when used as an auxiliary credit signal in MAPPO-style centralised-critic policy-gradient training. Its benefits are most apparent in structurally isolated, reward-coupled tasks, where sparse global rewards depend on teammates outside an agent's local observation pathway.

Table 7: Wall-clock training time (hours) across environment groups and map setups.

	Partitioned MPE (Small)		Partitioned MPE (Medium)		Partitioned MPE (Large)		Basilisk	
Algorithm	2x2	2x3	3x3	2x4	4x4	3x4	4ori	6ori
MAPPO	1.5	2.0	9.2	8.7	25.5	26.2	24.0	44.0
MAE-MAPPO	2.6	3.3	15.3	14.1	39.1	40.2	30.0	72.0
HAPPO	1.5	2.0	9.2	8.7	25.5	26.2	24.0	44.0
MAE-HAPPO	2.5	3.3	15.3	14.2	39.1	40.2	30.0	72.0
HATRPO	2.0	2.7	14.3	14.9	29.3	48.2	26.0	54.0
MAE-HATRPO	4.2	4.9	22.7	28.8	43.1	54.8	35.0	80.0

C.1.1 Vanilla MPE: Unconstrained Direct Cooperation

To explicitly isolate the performance gains attributable to resolving structural isolation, we first present the results of two ablation tasks without physical boundaries: `Vanilla-2x2-4A` and `Vanilla-2x2-2A2C`.

Table 8: Performance comparison of Vanilla MPE tasks (ablation without physical boundaries): `Vanilla-2x2-4A` and `Vanilla-2x2-2A2C`.

Task	Algorithm	1M	3M	5M	Improvement
`Vanilla-2x2-4A`	HAPPO	10.65 ± 1.15	13.92 ± 0.34	14.57 ± 0.41	–
	HATRPO	7.24 ± 0.77	11.53 ± 1.43	13.82 ± 2.04	–
	MAPPO	15.46 ± 0.87	20.05 ± 0.46	22.15 ± 0.66	–
	MAE-HAPPO	10.25 ± 0.24	12.68 ± 0.68	13.82 ± 0.49	-5.1%
	MAE-HATRPO	6.48 ± 0.23	11.21 ± 0.31	13.43 ± 0.98	-2.8%
	MAE-MAPPO	15.41 ± 1.17	19.08 ± 0.40	$\mathbf{22.98 \pm 0.19}$	**+3.8%**
`Vanilla-2x2-2A2C`	HAPPO	8.65 ± 0.21	10.86 ± 0.51	12.32 ± 0.05	–
	HATRPO	6.37 ± 0.13	9.48 ± 0.36	10.62 ± 0.64	–
	MAPPO	13.60 ± 0.21	15.61 ± 0.74	16.89 ± 0.12	–
	MAE-HAPPO	8.67 ± 0.43	11.35 ± 0.29	$\mathbf{13.42 \pm 0.13}$	+8.9%
	MAE-HATRPO	5.71 ± 0.32	9.70 ± 0.34	$\mathbf{11.62 \pm 0.28}$	**+9.4%**
	MAE-MAPPO	13.60 ± 0.46	16.30 ± 0.19	16.53 ± 0.23	-2.1%

Vanilla-2x2-4A and Vanilla-2x2-2A2C In these environments, physical partitions are removed, allowing agents to move freely and observe one another directly. The interaction graph is fully observable, converting the problem back to a standard direct cooperation paradigm. As shown in Table 8, under these conditions, the baseline algorithms (MAPPO, HAPPO, HATRPO) are inherently capable of establishing effective coordination using conventional advantage estimators. The integration of MAE yields only marginal gains (e.g., +3.8% for MAE-MAPPO) or even minor performance degradations (e.g., -5.1% for MAE-HAPPO and -2.1% for MAE-MAPPO in the heterogeneous setting). These ablation results confirm that MAE is not a general reward-inflation mechanism. Instead, its counterfactual marginalisation is specifically engineered to address the ambiguous credit assignment dilemma, which only emerges when structural isolation severs direct interaction links.

C.1.2 Small-size tasks: 4 agents

2x2-4A All four agents are functionally identical and evenly distributed across the environment. This setting provides a compact and cooperative scenario where agents share the same sensing and action capabilities, yet must still cooperate under decentralised execution and global rewards. MAE-HAPPO achieves the highest relative gain of 52.1%, indicating that shaping decentralised updates using local contribution signals is highly effective even in low-dimensional setups. MAE-HATRPO also exhibits a strong boost of +34.4%, showing that trust-region updates benefit from clearer agent-wise credit attribution. MAE-MAPPO achieves the best absolute performance, reaching a return of 33.17 at the end of training (24.5% improvement over MAPPO). Notably, it converges faster and with lower variance, highlighting MAE's capacity to regularise centralised critic training via targeted observation masking.

Table 9: Performance comparison of small-size tasks under structural isolation: 2x2-4A and 2x2-2A2C.

Task	Algorithm	1M	3M	5M	Improvement
2x2-4A	HAPPO	13.52 ± 0.84	16.14 ± 1.14	16.85 ± 1.11	–
	HATRPO	15.83 ± 0.38	19.04 ± 1.36	20.37 ± 1.52	–
	MAPPO	19.48 ± 3.87	25.93 ± 1.65	26.64 ± 0.74	–
	MAE-HAPPO	15.96 ± 2.34	22.46 ± 2.06	25.61 ± 0.68	**+52.1%**
	MAE-HATRPO	12.93 ± 0.14	22.07 ± 3.23	27.38 ± 2.82	+34.4%
	MAE-MAPPO	23.14 ± 0.68	29.14 ± 2.53	**33.17 ± 0.83**	+24.5%
2x2-2A2C	HAPPO	14.10 ± 0.56	16.30 ± 0.19	16.81 ± 0.36	–
	HATRPO	10.78 ± 0.35	14.42 ± 0.62	15.73 ± 1.43	–
	MAPPO	17.17 ± 2.74	18.99 ± 3.09	22.23 ± 4.58	–
	MAE-HAPPO	14.65 ± 1.11	18.16 ± 2.04	21.59 ± 1.63	+28.4%
	MAE-HATRPO	9.78 ± 0.18	18.80 ± 2.18	23.14 ± 3.19	**+47.0%**
	MAE-MAPPO	17.43 ± 2.14	24.03 ± 0.79	**26.23 ± 0.99**	+18.0%

2x2-2A2C It introduces functional heterogeneity: two agents of type A and two agents of type C. This setup increases cooperation complexity, as the reward-triggering conditions require complementary behaviours from different agent roles. Under this setting, MAE-HATRPO delivers the most significant gain of 47.0%, leveraging MAE to disambiguate the effect of functionally distinct agents in the global return. This results in more informed trust-region updates that account for the asymmetry in agent capabilities. MAE-HAPPO achieves a substantial 28.4% improvement, reinforcing the benefit of decentralised shaping when role differentiation is present. Although MAE-MAPPO exhibits the highest final return of 26.23 (an 18.0% improvement), its consistent improvements across all variants confirm that the framework effectively addresses the structural attribution challenge under indirect cooperation with functionally heterogeneous agents.

C.1.3 Small-size tasks: 6 agents

Table 10: Performance comparison of small-size tasks: 3x2-6A and 3x2-2A2B2C.

Task	Algorithm	2.5M	5M	7.5M	Improvement
3x2-6A	HAPPO	21.43 ± 2.28	25.15 ± 5.67	26.60 ± 6.26	–
	HATRPO	18.39 ± 1.00	25.01 ± 4.16	30.34 ± 3.56	–
	MAPPO	30.02 ± 5.72	35.35 ± 7.11	38.51 ± 7.36	–
	MAE-HAPPO	24.31 ± 2.14	31.41 ± 0.69	34.46 ± 0.76	**+29.5%**
	MAE-HATRPO	21.27 ± 3.59	28.32 ± 0.15	34.45 ± 0.58	+13.6%
	MAE-MAPPO	33.91 ± 0.71	41.24 ± 5.06	**44.10 ± 4.19**	+14.5%
3x2-2A2B2C	HAPPO	23.73 ± 0.44	31.10 ± 2.76	33.85 ± 3.33	–
	HATRPO	26.54 ± 5.62	33.13 ± 4.15	34.87 ± 3.15	–
	MAPPO	28.37 ± 4.03	30.38 ± 4.41	32.51 ± 5.97	–
	MAE-HAPPO	28.95 ± 2.76	34.36 ± 3.75	37.11 ± 4.80	+9.6%
	MAE-HATRPO	31.30 ± 1.92	35.40 ± 1.58	36.77 ± 2.33	+5.4%
	MAE-MAPPO	32.92 ± 0.71	36.68 ± 1.31	**38.26 ± 0.97**	**+17.7%**

3x2-6A It involves six homogeneous agents uniformly distributed across three zones. Under this setup, all MAE-augmented algorithms exhibit stable and consistent improvements. MAE-HAPPO delivers the highest relative gain of 29.5%, demonstrating that local contribution-based shaping is especially effective in stabilising decentralised updates. The early learning acceleration (from 2.5M to 5M steps) highlights MAE's role in guiding exploration when global feedback is sparse. MAE-HATRPO also benefits from the structured shaping signal (+13.6%), addressing the conservative nature of trust-region updates when lacking precise credit information. MAE-MAPPO achieves the best overall return of 44.10 (+14.5%), suggesting that MAE can improve training efficiency even in homogeneous structurally isolated teams.

3x2-2A2B2C Here, agents are functionally heterogeneous (A, B, C). This setup increases coordination complexity, as policies must align behaviours across agents with different functional capabilities. MAE-MAPPO attains the highest final

return of 38.26 (+17.7%). By estimating individual marginal contributions via masked observations, MAE enables the central critic to better assign value to functionally distinct agent behaviours, mitigating credit diffusion. MAE-HAPPO and MAE-HATRPO also deliver meaningful gains (9.6% and +5.4%, respectively), suggesting that structured advantage signals improve the alignment of functionally distinct agent policies toward global objectives.

C.1.4 Medium-size tasks: 8 agents

Table 11: Performance comparison of medium-size tasks: `4x2-8A` and `4x2-2A4B2C`.

Task	Algorithm	5M	10M	15M	Improvement
`4x2-8A`	HAPPO	40.22 ± 0.62	44.17 ± 3.68	45.33 ± 6.20	–
	HATRPO	32.10 ± 1.92	35.56 ± 0.99	37.46 ± 2.36	–
	MAPPO	46.50 ± 8.00	48.51 ± 6.97	55.02 ± 8.19	–
	MAE-HAPPO	37.38 ± 3.01	41.08 ± 1.75	43.80 ± 0.53	-3.4%
	MAE-HATRPO	33.82 ± 0.56	40.29 ± 3.11	47.65 ± 2.38	**+27.2%**
	MAE-MAPPO	53.42 ± 5.11	60.42 ± 4.33	$\mathbf{60.67 \pm 2.32}$	+10.3%
`4x2-2A4B2C`	HAPPO	40.33 ± 8.30	47.72 ± 4.45	51.98 ± 1.90	–
	HATRPO	39.43 ± 5.30	46.26 ± 7.45	48.49 ± 8.96	–
	MAPPO	49.85 ± 0.92	53.18 ± 1.10	53.82 ± 1.21	–
	MAE-HAPPO	41.41 ± 6.62	50.09 ± 5.43	55.66 ± 7.82	+7.1%
	MAE-HATRPO	38.45 ± 5.53	51.72 ± 1.91	58.55 ± 7.79	**+20.7%**
	MAE-MAPPO	53.93 ± 4.54	59.17 ± 5.46	$\mathbf{61.01 \pm 5.60}$	+13.4%

4x2-8A Here, eight agents of the same type are uniformly distributed. MAE-MAPPO achieves the highest final return of 60.67 (+10.3%). This indicates that structured advantage estimation effectively enhances centralised critic-based updates in larger homogeneous teams. MAE-HATRPO demonstrates the most significant relative improvement, boosting performance from 37.46 to 47.65 (+27.2%), providing valuable stability to trust-region updates. However, MAE-HAPPO exhibits a slight decline (-3.4%), potentially reflecting sensitivity to over-shaping effects when team size increases in fully decentralised updates.

4x2-2A4B2C This task features agents of three distinct functional types. MAE-MAPPO maintains its strong performance, achieving the highest return of 61.01 (+13.4%). MAE-HATRPO benefits notably as well, improving from 48.49 to 58.55 (+20.7%), confirming that marginal contribution shaping helps stabilise learning even when agent roles differ significantly. MAE-HAPPO achieves a moderate improvement (+7.1%). These results collectively reinforce MAE's robustness in scaling to mid-sized heterogeneous teams.

Table 12: Performance comparison of medium-size tasks: `3x3-9A` and `3x3-4A1B4C`.

Task	Algorithm	5M	10M	15M	Improvement
`3x3-9A`	HAPPO	39.92 ± 3.08	49.38 ± 5.42	53.18 ± 6.90	–
	HATRPO	34.70 ± 2.41	40.67 ± 2.72	43.59 ± 1.27	–
	MAPPO	48.34 ± 0.64	52.36 ± 1.70	56.13 ± 5.69	–
	MAE-HAPPO	40.97 ± 1.71	48.97 ± 2.09	55.89 ± 1.68	+5.1%
	MAE-HATRPO	34.03 ± 0.85	40.72 ± 1.02	45.93 ± 2.56	+5.4%
	MAE-MAPPO	52.66 ± 1.05	59.76 ± 0.42	$\mathbf{62.94 \pm 0.69}$	**+12.1%**
`3x3-4A1B4C`	HAPPO	37.53 ± 1.96	42.64 ± 3.65	46.07 ± 2.86	–
	HATRPO	34.38 ± 0.21	41.06 ± 1.33	43.29 ± 0.98	–
	MAPPO	38.85 ± 2.14	41.00 ± 1.29	42.29 ± 2.34	–
	MAE-HAPPO	35.66 ± 1.29	41.03 ± 2.20	46.29 ± 2.64	+0.5%
	MAE-HATRPO	35.51 ± 0.96	42.17 ± 0.35	45.04 ± 1.31	+4.0%
	MAE-MAPPO	40.63 ± 2.61	45.24 ± 3.07	$\mathbf{46.95 \pm 3.75}$	**+11.0%**

C.1.5 Medium-size tasks: 9 agents

3x3-9A It features nine homogeneous agents in a 3×3 grid. MAE-MAPPO achieves the best final return of 62.94 (+12.1%), confirming the effectiveness of MAE in scaling structured advantage estimation to moderately sized teams. MAE-HAPPO and MAE-HATRPO also exhibit stable gains of 5.1% and 5.4%, respectively. Notably, MAE-HAPPO shows improved stability during middle-stage learning, while MAE-HATRPO maintains consistent updates across training.

3x3-4A1B4C Here, the agent population consists of three functional types (A, B, C). Despite the increased heterogeneity, MAE-MAPPO achieves the highest final return of 46.95 (+11.0%). This reflects the capability of MAE to provide structured advantage signals that generalise to agents with dissimilar roles. MAE-HATRPO also yields a meaningful improvement of 4.0%, suggesting that marginal contribution estimation enhances robustness under trust-region constraints even when policy roles diverge.

Table 13: Performance comparison of large-size tasks: `4x3-12A` and `4x3-6A2B4C`.

Task	Algorithm	5M	10M	15M	Improvement
`4x3-12A`	HAPPO	37.33 ± 1.90	41.39 ± 1.48	45.29 ± 1.59	–
	HATRPO	37.61 ± 1.38	43.75 ± 0.67	47.13 ± 0.55	–
	MAPPO	45.30 ± 0.59	50.52 ± 1.39	53.14 ± 2.72	–
	MAE-HAPPO	36.56 ± 0.46	44.44 ± 2.30	48.46 ± 1.87	**+7.0%**
	MAE-HATRPO	36.32 ± 1.04	43.14 ± 0.31	47.07 ± 0.97	-0.1%
	MAE-MAPPO	49.72 ± 2.93	52.43 ± 2.91	$\mathbf{55.24 \pm 2.17}$	+4.0%
`4x3-6A2B4C`	HAPPO	34.93 ± 0.66	39.80 ± 2.95	44.18 ± 5.25	–
	HATRPO	34.17 ± 2.36	41.05 ± 2.07	44.33 ± 1.10	–
	MAPPO	39.89 ± 1.97	47.39 ± 0.65	49.88 ± 0.20	–
	MAE-HAPPO	35.37 ± 1.34	40.52 ± 0.16	43.60 ± 0.01	-1.3%
	MAE-HATRPO	35.16 ± 0.12	43.01 ± 0.49	$\mathbf{47.80 \pm 1.41}$	**+7.8%**
	MAE-MAPPO	41.71 ± 0.36	46.44 ± 0.87	48.88 ± 1.56	-2.0%

C.1.6 Large-size tasks: 12 agents

4x3-12A It involves 12 homogeneous agents. MAE-MAPPO yields the best final return of 55.24 (+4.0%). This suggests that the MAE framework effectively scales structured advantage estimation to mid-sized teams, maintaining centralised critic consistency. MAE-HAPPO also demonstrates a significant performance gain of 7.0%, indicating that local advantage shaping from MAE can complement decentralised optimisation schemes when agent roles are symmetric.

4x3-6A2B4C This task introduces considerable heterogeneity (6A, 2B, 4C). MAE-HATRPO achieves the best outcome, reaching a final return of 47.80 (+7.8%). This suggests that MAE can enhance stability in trust-region updates even in heterogeneous populations.

C.1.7 Large-size tasks: 16 agents

4x4-16A This represents the largest homogeneous task. MAE-MAPPO demonstrates strong scalability, achieving a final return of 60.57 (+12.8%). This suggests that the centralised critic in MAPPO can effectively leverage the structured shaping signals from MAE, even in dense-agent settings. MAE-HATRPO also provides a moderate gain of 4.9%.

4x4-8A4B4C This introduces significant agent diversity across 16 entities. MAE-MAPPO continues to perform reliably, reaching a final return of 53.15 (+6.3%). This demonstrates the robustness of MAE when integrated with centralised critics, as it can still provide effective structured advantage estimates despite functional heterogeneity.

Table 14: Performance comparison of large-size tasks: `4x4-16A` and `4x4-8A4B4C`.

Task	Algorithm	5M	10M	15M	Improvement
4x4-16A	HAPPO	44.53 ± 0.47	49.13 ± 0.76	52.14 ± 1.43	–
	HATRPO	40.05 ± 1.29	48.20 ± 1.40	50.93 ± 1.70	–
	MAPPO	48.82 ± 0.44	49.93 ± 1.18	53.68 ± 5.07	–
	MAE-HAPPO	41.07 ± 0.96	47.13 ± 1.50	51.39 ± 1.85	-1.4%
	MAE-HATRPO	39.97 ± 1.88	49.04 ± 1.04	53.42 ± 1.90	+4.9%
	MAE-MAPPO	52.14 ± 0.84	57.62 ± 6.32	$\mathbf{60.57 \pm 5.30}$	**+12.8%**
4x4-8A4B4C	HAPPO	41.90 ± 1.25	46.26 ± 2.29	49.42 ± 0.22	–
	HATRPO	40.48 ± 0.87	47.66 ± 0.28	51.85 ± 0.61	–
	MAPPO	46.72 ± 1.03	48.67 ± 0.67	50.00 ± 1.10	–
	MAE-HAPPO	39.37 ± 0.37	44.78 ± 1.03	49.55 ± 0.43	+0.3%
	MAE-HATRPO	38.67 ± 1.05	46.27 ± 0.51	49.44 ± 1.41	-4.6%
	MAE-MAPPO	47.79 ± 0.68	51.18 ± 2.67	$\mathbf{53.15 \pm 5.06}$	+6.3%

Table 15: Performance comparison of 4 satellite tasks: `4Ori-4A` and `4Ori-2B2C`.

Task	Algorithm	1M	2M	3M	Improvement
4Ori-4A	HAPPO	9.13 ± 0.85	34.93 ± 2.68	40.85 ± 0.37	–
	HATRPO	8.67 ± 1.75	34.09 ± 1.80	39.81 ± 1.04	–
	MAPPO	21.81 ± 6.46	32.57 ± 5.61	39.46 ± 3.27	–
	MAE-HAPPO	5.45 ± 0.35	23.75 ± 4.39	36.76 ± 1.91	-10.0%
	MAE-HATRPO	4.94 ± 1.17	31.70 ± 2.48	38.93 ± 0.24	-2.2%
	MAE-MAPPO	26.92 ± 5.84	37.33 ± 1.29	$\mathbf{42.16 \pm 0.39}$	**+6.8%**
4Ori-2B2C	HAPPO	4.37 ± 1.55	13.51 ± 0.23	17.72 ± 0.26	–
	HATRPO	4.94 ± 1.12	16.83 ± 0.15	19.74 ± 0.14	–
	MAPPO	5.02 ± 1.15	15.95 ± 0.17	17.43 ± 0.38	–
	MAE-HAPPO	2.15 ± 0.35	9.12 ± 4.38	15.02 ± 1.45	-13.8%
	MAE-HATRPO	2.68 ± 0.36	14.30 ± 1.11	18.23 ± 0.95	-7.65%
	MAE-MAPPO	14.26 ± 0.18	18.09 ± 0.27	$\mathbf{19.10 \pm 0.07}$	**+9.6%**

C.2 LEARNING PERFORMANCE IN BASILISK ENVIRONMENT

C.2.1 Target tasking with 4 satellites

4Ori-4A This task involves 4 identical satellites. MAE-MAPPO achieves the best performance, with a final return of 42.16 (+6.8%). This highlights the ability of centralised critics to exploit structured shaping signals. In contrast, both MAE-HAPPO and MAE-HATRPO underperform compared to their baselines, suggesting that decentralised updates may require stronger signals in small-team settings to fully benefit from individual contribution estimates.

4Ori-2B2C This introduces heterogeneity by allocating two satellites of type B and two of type C. MAE-MAPPO again demonstrates robustness, achieving a final return of 19.10 (+9.6%). This confirms that MAE remains effective when paired with centralised critics. However, MAE-HAPPO and MAE-HATRPO show degradation.

C.2.2 Target tasking with 6 satellites

6Ori-6A This models a medium-scale homogeneous satellite cooperation scenario. MAE-MAPPO achieves a strong return of 37.73 (+13.0%), demonstrating that structured shaping from MAE can enhance centralised critics. MAE-HATRPO achieves the highest absolute return of 38.53 (+2.6%).

6Ori-2A2B2C This introduces agent heterogeneity. MAE-MAPPO achieves a significant 16.6% gain, reaching 30.48, which affirms the adaptability of MAE when used with centralised training. MAE-HATRPO shows a consistent gain (+3.0%), while MAE-HAPPO lags behind.

Table 16: Performance comparison of 6 satellite tasks: 6Ori-6A and 6Ori-2A2B2C.

Task	Algorithm	2M	4M	6M	Improvement
6Ori-6A	HAPPO	5.70 ± 0.16	17.82 ± 0.13	24.59 ± 1.12	–
	HATRPO	17.53 ± 0.54	33.66 ± 0.90	37.56 ± 1.18	–
	MAPPO	24.76 ± 0.20	29.73 ± 2.46	33.38 ± 0.55	–
	MAE-HAPPO	4.72 ± 0.45	12.65 ± 1.76	23.31 ± 2.87	-5.2%
	MAE-HATRPO	13.34 ± 2.26	33.87 ± 0.51	$\mathbf{38.53 \pm 1.46}$	+2.6%
	MAE-MAPPO	27.92 ± 0.59	32.48 ± 1.15	37.73 ± 4.07	**+13.0%**
6Ori-2A2B2C	HAPPO	18.03 ± 2.12	28.47 ± 0.28	29.14 ± 0.11	–
	HATRPO	21.86 ± 3.17	29.08 ± 0.57	30.58 ± 0.27	–
	MAPPO	13.97 ± 10.48	22.26 ± 5.39	26.14 ± 6.43	–
	MAE-HAPPO	14.12 ± 6.30	24.39 ± 3.92	27.53 ± 1.56	-5.5%
	MAE-HATRPO	21.15 ± 3.02	29.39 ± 0.34	31.50 ± 0.50	+3.0%
	MAE-MAPPO	23.30 ± 2.66	28.61 ± 4.79	$\mathbf{30.48 \pm 2.44}$	**+16.6%**

Table 17: Critic Gradient Norm Diagnostics. Statistical comparison of gradient norms (Mean $\pm$ Std) across the 16 indirect-cooperation scenarios. Maps are grouped by environment scale (Small, Medium, Large, Basilisk). Within each Vanilla–MAE pair, bold indicates the lower mean gradient norm. Standard deviations are reported without bolding. These statistics are used as empirical diagnostics of update variability and should not be interpreted as direct proof of variance reduction.

Size	Map Configuration	MAPPO		HAPPO		HATRPO	
		Vanilla	MAE	Vanilla	MAE	Vanilla	MAE
Small	2x2-4A	0.1810 ± 0.0777	$\mathbf{0.1486 \pm 0.0660}$	0.3106 ± 0.0950	$\mathbf{0.1963 \pm 0.0746}$	1.0308 ± 0.2332	$\mathbf{0.1340 \pm 0.0684}$
	2x2-2A2C	0.2812 ± 0.1034	$\mathbf{0.2133 \pm 0.0817}$	0.2911 ± 0.0921	$\mathbf{0.2869 \pm 0.1052}$	0.2246 ± 0.0804	$\mathbf{0.1746 \pm 0.0975}$
	3x2-6A	0.1905 ± 0.0799	$\mathbf{0.1656 \pm 0.0828}$	0.2583 ± 0.1237	$\mathbf{0.1549 \pm 0.0653}$	0.1631 ± 0.0801	$\mathbf{0.1165 \pm 0.0577}$
	3x2-2A2B2C	0.2446 ± 0.0989	$\mathbf{0.1763 \pm 0.0671}$	0.2141 ± 0.0766	$\mathbf{0.2040 \pm 0.0776}$	$\mathbf{0.1346 \pm 0.0435}$	0.1445 ± 0.0532
Medium	4x2-8A	0.2696 ± 0.1402	$\mathbf{0.2400 \pm 0.1296}$	0.3044 ± 0.0958	$\mathbf{0.1576 \pm 0.0573}$	0.1940 ± 0.0680	$\mathbf{0.1415 \pm 0.0622}$
	4x2-2A4B2C	0.1414 ± 0.0526	$\mathbf{0.1170 \pm 0.0486}$	0.1458 ± 0.0559	$\mathbf{0.1165 \pm 0.0469}$	0.1255 ± 0.0545	$\mathbf{0.1154 \pm 0.0407}$
	3x3-9A	0.1584 ± 0.0761	$\mathbf{0.1130 \pm 0.0465}$	$\mathbf{0.1282 \pm 0.0555}$	0.1507 ± 0.0631	$\mathbf{0.1520 \pm 0.0615}$	0.1552 ± 0.0735
	3x3-4A1B4C	0.1909 ± 0.0652	$\mathbf{0.1577 \pm 0.0637}$	$\mathbf{0.1651 \pm 0.0743}$	0.1662 ± 0.0726	$\mathbf{0.1445 \pm 0.0498}$	0.1488 ± 0.0627
Large	4x3-12A	0.1799 ± 0.0867	$\mathbf{0.1781 \pm 0.0940}$	0.1778 ± 0.0776	$\mathbf{0.1591 \pm 0.0744}$	$\mathbf{0.1416 \pm 0.0704}$	0.1466 ± 0.0603
	4x3-6A2B4C	$\mathbf{0.1834 \pm 0.0971}$	0.1945 ± 0.0961	$\mathbf{0.1946 \pm 0.0841}$	0.2052 ± 0.0876	$\mathbf{0.1394 \pm 0.0552}$	0.1548 ± 0.0661
	4x4-16A	0.1802 ± 0.0843	$\mathbf{0.1393 \pm 0.0700}$	0.1595 ± 0.0722	$\mathbf{0.1562 \pm 0.0815}$	0.1548 ± 0.0772	$\mathbf{0.1390 \pm 0.0654}$
	4x4-8A4B4C	0.2264 ± 0.1081	$\mathbf{0.1910 \pm 0.1054}$	0.1906 ± 0.0854	$\mathbf{0.1893 \pm 0.1027}$	$\mathbf{0.1398 \pm 0.0652}$	0.1493 ± 0.0740
Basilisk	4Ori-4A	0.2530 ± 0.0825	$\mathbf{0.2109 \pm 0.0523}$	$\mathbf{0.2244 \pm 0.0629}$	0.3190 ± 0.1411	$\mathbf{0.2201 \pm 0.0634}$	0.2259 ± 0.0737
	4Ori-2B2C	0.6429 ± 0.1732	$\mathbf{0.6194 \pm 0.1984}$	$\mathbf{0.4973 \pm 0.1806}$	0.7273 ± 0.3565	$\mathbf{0.3636 \pm 0.0975}$	0.4610 ± 0.1372
	6Ori-6A	0.3008 ± 0.1014	$\mathbf{0.2587 \pm 0.0638}$	$\mathbf{0.3200 \pm 0.1192}$	0.3739 ± 0.1702	0.2103 ± 0.0563	$\mathbf{0.2090 \pm 0.0573}$
	6Ori-2A2B2C	0.3559 ± 0.1024	$\mathbf{0.2772 \pm 0.0806}$	$\mathbf{0.2295 \pm 0.0652}$	0.2623 ± 0.0944	0.2328 ± 0.0708	$\mathbf{0.2238 \pm 0.0697}$

C.3 DIAGNOSTICS FOR THEOREM 1: VALUE-ESTIMATION STABILITY

Theorem 1 provides a conditional variance-reduction result: under positively correlated critic errors and no larger marginal error variance, the full-vs-masked MAE estimator has lower variance than an independent simulation-based counterfactual estimator. To examine whether the empirical training behaviour is consistent with this analysis, we record the statistical properties of the Critic Value Loss during the convergence phase. The results are summarised in Table 18.

- **Value-estimation stability:** As shown in the "Std Dev" columns, MAE-based methods often exhibit lower standard deviations than their corresponding vanilla baselines. In many tasks, the reduction is approximately 30%–60%, suggesting that the full-vs-masked critic computation can provide a more stable value-estimation signal under structural isolation.

- **Behaviour across task scales:** This trend appears not only in small-scale tasks such as 2x2-4A, but also in several larger or heterogeneous settings such as 4x4-8A4B4C. These diagnostics are consistent with the conditional variance-reduction interpretation in Theorem 1, while remaining subject to the critic-error assumptions stated in the theorem.

C.4 DIAGNOSTICS FOR THEOREM 2: GRADIENT STABILITY

Theorem 2 suggests that, when the global advantage contains a latent component induced by structurally isolated teammates, an auxiliary MAE signal that better approximates the local contribution can reduce residual gradient noise under the stated assumptions. To examine this effect empirically, we analyse Critic Gradient Norm statistics, as reported in Table 17.

- **Gradient-stability trend:** In several indirect-cooperation tasks, vanilla baselines exhibit larger gradient fluctuations. MAE-based variants tend to produce lower or more stable gradient norms in these cases. For example, in the `2x2-4A` task, HATRPO shows a high gradient norm, while MAE-HATRPO reduces it to a much lower range.
- **Behaviour in larger tasks:** Similar trends are observed in larger or more complex settings. For instance, in the `6Ori-6A` Basilisk task, MAE-MAPPO reduces the gradient-norm standard deviation from 0.1014 to 0.0638 compared with MAPPO. These results are consistent with the view that MAE can reduce residual update variability under structural isolation, but they should be interpreted as empirical diagnostics rather than direct proof of the theorem.

Table 18: Critic Value Loss Analysis. Statistical comparison of Critic Value Loss (Mean ± Std) across all 16 experimental scenarios. Maps are grouped by environment scale. Lower standard deviations in MAE variants are consistent with the conditional variance-reduction analysis.

Size	Map Configuration	MAPPO		HAPPO		HATRPO	
		Vanilla	MAE	Vanilla	MAE	Vanilla	MAE
Small	2x2-4A	0.0250 ± 0.0084	**0.0152 ± 0.0044**	0.0724 ± 0.0150	**0.0281 ± 0.0082**	0.1026 ± 0.0362	**0.0153 ± 0.0050**
	2x2-2A2C	0.0473 ± 0.0133	**0.0292 ± 0.0070**	0.0646 ± 0.0093	**0.0558 ± 0.0136**	0.0595 ± 0.0135	**0.0251 ± 0.0130**
	3x2-6A	0.0320 ± 0.0138	**0.0238 ± 0.0075**	0.0574 ± 0.0283	**0.0213 ± 0.0051**	0.0259 ± 0.0134	**0.0145 ± 0.0055**
	3x2-2A2B2C	0.0483 ± 0.0181	**0.0313 ± 0.0071**	0.0408 ± 0.0109	**0.0324 ± 0.0102**	**0.0207 ± 0.0041**	0.0220 ± 0.0039
Medium	4x2-8A	0.0554 ± 0.0295	**0.0547 ± 0.0223**	0.0634 ± 0.0171	**0.0264 ± 0.0036**	0.0404 ± 0.0098	**0.0200 ± 0.0035**
	4x2-2A4B2C	0.0269 ± 0.0079	**0.0195 ± 0.0038**	0.0235 ± 0.0080	**0.0183 ± 0.0082**	0.0232 ± 0.0103	**0.0169 ± 0.0030**
	3x3-9A	0.0256 ± 0.0054	**0.0205 ± 0.0037**	**0.0205 ± 0.0071**	0.0250 ± 0.0046	0.0300 ± 0.0088	**0.0271 ± 0.0096**
	3x3-4A1B4C	0.0478 ± 0.0085	**0.0386 ± 0.0145**	0.0337 ± 0.0114	**0.0304 ± 0.0038**	0.0283 ± 0.0052	**0.0267 ± 0.0040**
Large	4x3-12A	0.0303 ± 0.0037	**0.0288 ± 0.0065**	0.0337 ± 0.0057	**0.0287 ± 0.0051**	**0.0187 ± 0.0024**	0.0203 ± 0.0023
	4x3-6A2B4C	**0.0343 ± 0.0042**	0.0371 ± 0.0052	0.0426 ± 0.0089	**0.0378 ± 0.0048**	0.0264 ± 0.0026	**0.0238 ± 0.0032**
	4x4-16A	0.0392 ± 0.0055	**0.0293 ± 0.0072**	0.0292 ± 0.0027	**0.0273 ± 0.0034**	0.0202 ± 0.0025	**0.0178 ± 0.0022**
	4x4-8A4B4C	0.0481 ± 0.0057	**0.0418 ± 0.0079**	0.0426 ± 0.0059	**0.0393 ± 0.0055**	**0.0240 ± 0.0025**	0.0253 ± 0.0024
Basilisk	4Ori-4A	0.0301 ± 0.0072	**0.0233 ± 0.0018**	**0.0261 ± 0.0057**	0.0479 ± 0.0213	**0.0238 ± 0.0036**	0.0254 ± 0.0053
	4Ori-2B2C	0.0639 ± 0.0102	**0.0571 ± 0.0043**	**0.0930 ± 0.0220**	0.1879 ± 0.1229	**0.0608 ± 0.0080**	0.0810 ± 0.0225
	6Ori-6A	0.0413 ± 0.0048	**0.0351 ± 0.0024**	**0.0644 ± 0.0069**	0.0839 ± **0.0249**	0.0271 ± 0.0022	**0.0271 ± 0.0021**
	6Ori-2A2B2C	0.0473 ± 0.0057	**0.0393 ± 0.0065**	**0.0336 ± 0.0025**	0.0444 ± 0.0099	0.0334 ± 0.0022	**0.0323 ± 0.0026**

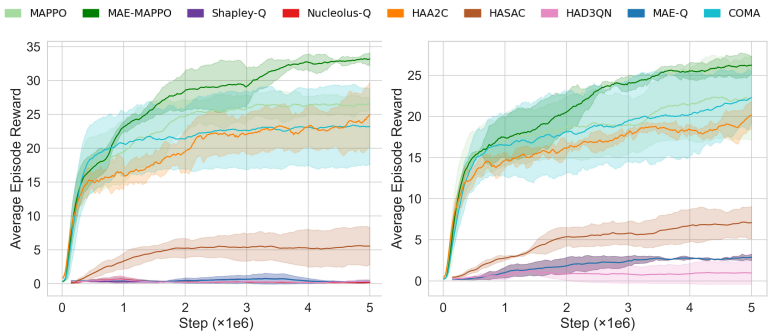
C.5 COMPARISON WITH VALUE-BASED AND GAME-THEORETIC MARL BASELINES

To further examine MAE under indirect cooperation, we compare MAE-augmented methods with representative value-based and game-theoretic MARL baselines. These include Shapley-Q [Wang et al., 2020b] and Nucleolus-Q [Li et al., 2025], which estimate cooperative payoff allocations from game-theoretic principles, and HAD3QN and HASAC [Zhong et al., 2024], which represent value-decomposition and soft actor-critic approaches for heterogeneous teams. We also include MAE-Q, a value-based variant that incorporates the MAE signal into Q-learning. We evaluate these methods on two representative Partitioned MPE tasks: the homogeneous `2x2-4A` task and the heterogeneous `2x2-2A2C` task. The results are summarised in Table 19 and visualised in Fig. 4.

Limitations of Value-based and Game-Theoretic Methods. Shapley-Q and Nucleolus-Q estimate marginal contributions using permutation-based or excess-based cooperative-game metrics. In the indirect-cooperation tasks considered here, these methods obtain much lower returns than policy-gradient baselines. As shown in Table 19, Shapley-Q and Nucleolus-Q remain below 3.0 final return on both tasks, whereas MAPPO reaches 26.64 on `2x2-4A` and 22.23 on `2x2-2A2C`. This suggests that coalition- or action-level credit signals are difficult to estimate reliably when agents are structurally isolated and rewards are sparse. MAE-Q also provides only limited improvement over value-based baselines, indicating that incorporating a masked marginal signal into Q-learning does not by itself resolve the bootstrapping difficulty under severe partial observability.

Table 19: Performance comparison of value-based and policy-gradient MARL algorithms on Partitioned MPE tasks `2x2-4A` and `2x2-2A2C`. The results indicate that MAE-MAPPO performs strongly in these indirect-cooperation settings, while value-based and game-theoretic baselines show limited performance under sparse global rewards and structural isolation.

Task	Type	Algorithm	1M	3M	5M
`2x2-4A`	Value	Nucleolus-Q	0.64 ± 0.42	0.26 ± 0.06	0.15 ± 0.00
	Value	Shapley-Q	0.29 ± 0.08	0.67 ± 0.56	0.32 ± 0.18
	Value	HAD3QN	0.57 ± 0.52	0.30 ± 0.12	0.33 ± 0.25
	Value	MAE-Q	0.49 ± 0.39	0.64 ± 0.59	0.32 ± 0.18
	Value	HASAC	3.34 ± 1.16	5.33 ± 1.99	5.55 ± 2.75
	Policy	HAA2C	16.05 ± 1.73	22.06 ± 2.63	24.92 ± 4.41
	Policy	MAPPO	19.48 ± 3.87	25.93 ± 1.65	26.64 ± 0.74
	Policy	**MAE-MAPPO**	$\mathbf{23.14 \pm 0.68}$	$\mathbf{29.14 \pm 2.53}$	$\mathbf{33.17 \pm 0.83}$
`2x2-2A2C`	Value	Nucleolus-Q	0.88 ± 0.42	2.24 ± 0.62	2.67 ± 0.43
	Value	Shapley-Q	0.95 ± 0.34	2.37 ± 0.75	2.57 ± 0.32
	Value	HAD3QN	0.66 ± 0.49	0.84 ± 1.13	0.97 ± 1.37
	Value	MAE-Q	1.00 ± 0.63	2.45 ± 0.57	2.87 ± 0.40
	Value	HASAC	2.77 ± 0.24	5.81 ± 1.44	7.08 ± 1.96
	Policy	HAA2C	14.66 ± 0.80	18.02 ± 0.37	20.13 ± 1.53
	Policy	MAPPO	17.17 ± 2.74	18.99 ± 3.09	22.23 ± 4.58
	Policy	**MAE-MAPPO**	$\mathbf{17.43 \pm 2.14}$	$\mathbf{24.03 \pm 0.79}$	$\mathbf{26.23 \pm 0.99}$



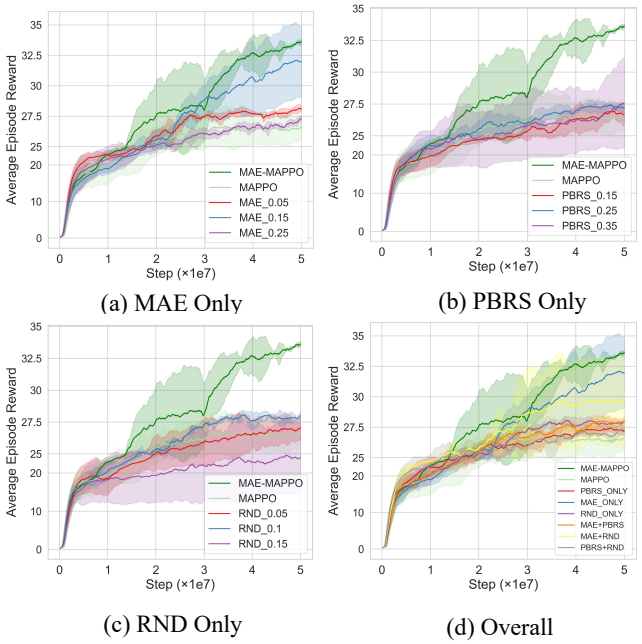
(a) 2x2-4A (b) 2x2-2A2C

Figure 4: Learning curves comparison. We contrast the performance of Value-based methods (e.g., Shapley-Q, Nucleolus-Q) against Policy-based methods (MAPPO, MAE-MAPPO) on Partitioned MPE tasks. The results show that value-based and game-theoretic methods struggle in these sparse-reward, structurally isolated settings, while MAE-MAPPO achieves stronger performance.

Policy-Gradient Methods. Policy-gradient methods perform more strongly in these two tasks. MAPPO, which uses CTDE and GAE, achieves substantially higher returns than the value-based baselines. MAE-MAPPO further improves performance, reaching 33.17 on `2x2-4A` and 26.23 on `2x2-2A2C`. These results support the intended scope of MAE: the full-vs-masked critic signal is most effective when used as an auxiliary credit term in centralised-critic policy-gradient optimisation, rather than as a direct replacement for value-based backup or coalition-value computation. HAA2C and HASAC perform better than pure Q-learning methods but still trail MAE-MAPPO, suggesting that parameter sharing or entropy regularisation alone is insufficient to match the benefit of an explicit marginal credit signal in these structurally isolated tasks.

C.6 ABLATION STUDY

To assess the individual contributions of MAE and the auxiliary training mechanisms, PBRS and RND, we conduct an ablation study on the representative Partitioned MPE task `2x2-4A`. We analyse the isolated and combined effects of each component within the MAE-MAPPO framework.



(a) MAE Only (b) PBRS Only

(c) RND Only (d) Overall

Figure 5: Component analysis of MAE-MAPPO on the `2x2-4A` task. (a) Effect of the MAE weight λ with PBRS and RND disabled. (b) Effect of the PBRS weight α without MAE or RND. (c) Effect of the RND weight β without MAE or PBRS. (d) Comparison of individual and combined modules. The results suggest that MAE is the main contributor to final performance, while PBRS and RND mainly improve early-stage learning.

Impact of MAE. Fig. 5(a) isolates the contribution of the MAE signal by varying the weighting coefficient $\lambda \in \{0.05, 0.15, 0.25\}$, with PBRS and RND disabled. Compared with the MAPPO baseline, all tested MAE variants improve sample efficiency and final return on this task. This supports the interpretation that the full-vs-masked critic signal provides useful auxiliary credit information under structural isolation. Among the tested values, $\lambda = 0.15$ provides the best balance between stability and final performance. Increasing the weight to $\lambda = 0.25$ accelerates early learning but also increases variability, likely because critic approximation errors are amplified in the policy-update signal.

Impact of PBRS. Fig. 5(b) evaluates Potential-Based Reward Shaping (PBRS) across weights $\alpha \in \{0.15, 0.25, 0.35\}$, with MAE and RND disabled. PBRS improves learning over standard MAPPO in this sparse-reward task, with $\alpha = 0.25$ performing best among the tested values. However, its effect is mainly concentrated in early-stage learning, as PBRS provides an auxiliary shaping signal but does not directly address the structural credit-assignment difficulty caused by indirect cooperation.

Impact of RND. Fig. 5(c) examines Random Network Distillation (RND) with weights $\beta \in \{0.05, 0.10, 0.15\}$. RND improves early reward acquisition by encouraging exploration under partial observability, with $\beta = 0.10$ giving the best trade-off among the tested settings. However, RND alone yields a lower final return than MAE-based variants, suggesting that exploration bonuses help agents discover rewarding states but do not fully replace the need for a marginal credit signal.

Combined Modules. Fig. 5(d) compares the full MAE-MAPPO configuration against single-module and pairwise combinations. The results show three trends. First, configurations without MAE, such as `PBRS_ONLY`, `RND_ONLY`, and `PBRS+RND`, do not reach the same final return as MAE-based configurations. Second, PBRS and RND accelerate learning when combined with MAE, indicating that they are useful auxiliary training aids in sparse-reward settings. Third, the full combination, with $\lambda = 0.15$, $\alpha = 0.25$, and $\beta = 0.10$, achieves the best overall learning curve among the tested configurations. These results support the interpretation that MAE provides the core credit-assignment signal, while PBRS and RND help improve optimisation and exploration efficiency.