CAUSAL SHAPLEY INTERACTION CONSISTENCY FOR CREDIT ASSIGNMENT IN COOPERATIVE MULTI-AGENT REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Shapley-based multi-agent reinforcement learning (MARL) methods distribute a shared reward according to agents' marginal contributions to provide a credit assignment as a prominent topic in artificial intelligence research. However, existing methods often overlook the relationship of causal semantics between agents' actions, leading to inaccurate credit assignment. To address this issue, we propose a Causal Shapley Interventional Consistency Regularization framework, namely THEMIS, which separates causal identification, critic calibration, and policy optimization through three components. The causal-identification component applies a one-time direct intervention to recognize causal relationship between agents' actions. The critic-calibration component applies an interaction consistency to align selected critic-induced interactions with these causal targets. The actor component applies a detached conventional advantage to prevent teacher gradients from influencing the policy update. Across different agents MPE tasks, THEMIS exceeds the normalized team-reward baselines on both learning metrics.

1 INTRODUCTION

Credit assignment is a central challenge in cooperative multi-agent reinforcement learning (MARL) (Foerster et al., 2018; Rashid et al., 2018; Yu et al., 2022). It transforms a shared team return into informative learning signals that capture the contributions of individual agents and their interactions. Among principled approaches to this problem, Shapley-based methods quantify agents' marginal contributions across coalitions and have been applied to credit estimation, value factorization, and spatial-temporal return allocation in cooperative MARL (Li et al., 2021; Wang et al., 2022; Chen et al., 2024). However, these methods typically construct coalition values using observational conditionals or counterfactual action replacements. Without explicit intervention semantics, such constructions do not, by themselves, identify the causal effects of intervening on simultaneous joint actions. This limitation matters in coordinated decision-making problems such as traffic control (Chu et al., 2020), wireless communication networks (Nasir & Guo, 2019), and multi-robot control (Chen et al., 2017), where correlated behavior and constrained joint-action spaces can make observational attribution misleading.

Existing approaches to cooperative credit assignment broadly follow two complementary directions. The first learns structured decompositions of a team value, ranging from additive and monotonic factorization to interaction-aware mixing architectures (Sunehag et al., 2018; Rashid et al., 2018; Wang et al., 2022; Liang et al., 2025; Xu et al., 2025). These methods support decentralized action selection while representing increasingly expressive forms of cooperation. The second estimates agent-specific contributions through counterfactual baselines, difference rewards, coalition games, or Shapley values (Wolpert & Tumer, 2002; Foerster et al., 2018; Li et al., 2021; Chen et al., 2024; Ding et al., 2025). Although these approaches provide informative and interpretable learning signals, the numerical validity of a Shapley allocation does not determine the causal meaning of its underlying coalition game. Causal attribution research instead emphasizes that contributions must be defined through interventions rather than associations when causal effects are the target (Heskes et al., 2020; Janzing et al., 2020; Frye et al., 2020).

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

Applying this principle to simultaneous-action MARL raises two distinct challenges. The first is *causal identification*. Observational conditioning may retain dependence induced by an unobserved common cause of the agents' actions and the team return, while fixed action replacements may generate illegal or unsupported joint actions. Consequently, even a Shapley value that satisfies the standard allocation axioms may fail to represent the effect of changing the agents' joint decision. The second challenge is *learning integration*. Once a causal interaction target has been identified, directly inserting it into the actor objective can conflate causal estimation with policy optimization and introduce uncontrolled gradients from the causal teacher. A suitable framework must therefore identify intervention-based coalition values, calibrate the learned value representation against them, and preserve a well-defined actor update.

To address this issue, we propose Themis, a Causal Shapley Interventional Consistency Regularization framework consisting of three components. The causal-identification component constructs policy-interventional coalition values through one-time joint-action interventions and converts them into stop-gradient Shapley-interaction targets. The critic-calibration component aligns interactions represented by the centralized critic with these causal targets. The actor component retains a conventional detached advantage and standard update, preventing causal-teacher gradients from directly entering policy optimization.

Our contributions are as follows:

- **A policy-interventional formulation of Shapley credit.** We define a long-horizon coalition game through a one-time intervention on a simultaneous joint action under a fixed continuation policy. The formulation makes the required temporal-ordering, consistency, legal-support, complement-kernel, and identification assumptions explicit, separating causal intervention effects from observational associations.
- **A three-component critic-calibration framework.** We introduce Themis, which converts policy-interventional coalition values into a stop-gradient Shapley-interaction teacher, aligns selected centralized-critic interactions with that teacher, and preserves a detached conventional advantage for actor optimization. The resulting gradient boundaries prevent causal targets from becoming uncontrolled actor-side rewards or control variates.
- **Formal calibration guarantees and scope boundaries.** We show that an exact teacher and zero interaction-consistency loss recover the targeted causal Shapley interactions when the critic and teacher use the same coalition game. We further show that observational prediction accuracy alone cannot certify causal interaction consistency whenever the observational–interventional discrepancy survives the selected interaction contrast. These results guarantee calibration of queried interactions, not improved policy return.
- **Consistent learning gains with valid and stable interventions.** Across three- to five-agent MPE Simple Spread tasks, Themis is the method that consistently exceeds the normalized team-reward baseline on both learning metrics, improving learning AUC by 5%–7% and final-window return by 4%–5.5%. In frozen-checkpoint audits, policy-supported interventions eliminate zero-support coalition queries (0% versus 50%–61% for fixed replacement) and reduce normalized plan variance by 28%–33%.

2 Related Work

Shapley credit assignment in MARL. Explicit credit methods seek to explain a team value through counterfactual changes in individual actions. The underlying allocation rule formalizes efficiency, symmetry, dummy, and additivity properties (Shapley, 1953; Owen, 1972). Li et al. (2021) introduced Shapley Counterfactual Credits, which approximate coalition marginal contributions by Monte Carlo sampling and demonstrated the utility of Shapley reasoning in cooperative MARL. Because exact Shapley evaluation scales exponentially, practical systems rely on sampling or alternative estimators (Castro et al., 2009; Kolpaczki et al., 2024). SHAQ develops a Markov Shapley value for value factorization, while STAS combines temporal return decomposition with spatial Shapley allocation (Wang et al., 2022; Chen et al., 2024). A broader theoretical treatment connects Markov convex games, dummy identification, and energy-network control (Wang, 2024), and Shapley Machine extends related axiomatic reasoning to open ad hoc teams (Wang et al., 2025a). Subsequent work has also incorporated Shapley-style signals into policy-gradient learning; for ex-

ample, HIS combines historical interaction information with a hybrid credit and reward signal (Ding et al., 2025). Shapley-style analyses have further been used to quantify agent importance and attribute credit in LLM-agent cooperation and agentic workflows (Mahjoub et al., 2023; Hua et al., 2025; Yang et al., 2025). However, these methods often leave counterfactual action support implicit, potentially assigning credit through infeasible coalition actions.

Value decomposition and credit integration. Cooperative MARL can address credit assignment through value factorization, counterfactual baselines, or centralized policy optimization (Sunehag et al., 2018; Rashid et al., 2018; Foerster et al., 2018; Kuba et al., 2022; Yu et al., 2022). Social-influence rewards provide another route for shaping coordination signals (Jaques et al., 2019). Surveys and benchmark studies show that performance depends strongly on task structure, implementation details, and evaluation protocols (Hernandez-Leal et al., 2019; Gronauer & Diepold, 2022; Papoudakis et al., 2021; Samvelyan et al., 2019). However, these approaches do not generally provide causally interpretable, axiomatic allocations of individual contributions.

Causal credit and causal attribution. Model-explanation methods use Shapley allocations to quantify feature contributions (Lundberg & Lee, 2017; Štrumbelj & Kononenko, 2014). Causal Shapley values replace purely observational feature coalitions with intervention semantics while retaining the Shapley axioms (Heskes et al., 2020). Feature-attribution research likewise shows that relevance is a causal question and that causal ordering can change otherwise valid Shapley explanations (Janzing et al., 2020; Frye et al., 2020). Structural causal models and potential-outcome frameworks provide a formal language for distinguishing interventions from observational conditioning (Pearl, 2009; Peters et al., 2017; Hernán & Robins, 2020; Imbens & Rubin, 2015). In MARL, MACCA models an offline data-generating process with a dynamic Bayesian network to identify individual reward structure (Wang et al., 2025b). However, existing formulations do not jointly address history sufficiency, policy support, and simultaneous multi-agent interventions.

3 Causal Shapley Interaction Consistency

3.1 Problem Setting and Intervention Scope

Consider a cooperative partially observed Markov game with agents $N = \{1, \dots, n\}$, pre-action history H_t, legal-action masks M_t, simultaneous joint action $A_t = (A_t^1, \dots, A_t^n)$, and discounted return

$$G_t = \sum_{\ell \geq 0} \gamma^\ell R_{t+\ell}. \tag{1}$$

During centralized training, the critic may condition on the declared history H_t, whereas each decentralized actor selects its action from a local history. We require H_t and M_t to be determined before A_t: they may contain past observations and actions, recurrent states, agent identities, and centralized state features, but not the current factual action, reward, or any post-action variable.

Our intervention changes the simultaneous joint decision only at time t. All subsequent actions follow a fixed continuation policy π^c, while the outcome remains the long-horizon return G_t. Therefore, "one-step" describes the treatment rather than the outcome horizon. The resulting estimand captures all downstream consequences of the intervention at time t, but it does not decompose credit across time or identify the effect of a sequence of interventions.

The players in the induced cooperative game are the agents acting simultaneously at time t. Consequently, the Shapley interactions introduced below allocate interaction across agents for this joint decision. Sequential or autoregressive action generation would require an explicit within-step causal ordering and a different intervention kernel; the present formulation does not apply to such settings without modification.

3.2 Identification Assumptions and Coalition Value

For a coalition $S \subseteq N$, let a_t^S denote a legal assignment to the actions of the coalition members. The remaining actions are generated by a declared complement kernel $g_t^{-S}(\cdot \mid H_t, a_t^S)$. The causal interpretation of the resulting coalition game is conditional on the following assumptions.

- **A1 (Pre-action information).** The history H_t and action masks M_t are measured before the joint action A_t. Neither contains the factual current action, the current reward, a post-action state, or any other post-treatment variable.
- **A2 (Simultaneous assignment).** The components of A_t are assigned at one decision point. Any causal ordering among agents within the same environment step must instead be represented by a different structural model.
- **A3 (Consistency and legal support).** Intervening to the factual legal action reproduces the corresponding factual potential outcome. Every intervened or regenerated action must be legal under M_t. If coalition values are estimated observationally, positivity is additionally required for every queried action assignment.
- **A4 (Declared complement kernel).** The kernel g_t^{-S} is fixed before evaluating the coalition game. For a factorized decentralized policy, it is constructed from the policy factors of the non-members. A correlated policy instead requires an explicit joint complement kernel, potentially conditional on the intervened assignment a_t^S.
- **A5 (Fixed continuation).** The reset convention, environment dynamics, horizon, discount factor, and continuation policy π^c are identical across coalition queries. Only the time-t action assignment changes.
- **A6 (Identification route).** Either the simulator or structural causal model supports the required stochastic intervention directly, or H_t is sufficient to remove action–return confounding for the queried assignments. A learned response model is treated only as an estimator until it is calibrated on held-out interventional data.

Under these assumptions, define the policy-interventional coalition value as

$$v_t^{\mathrm{do}}(S; h, a^S) = \mathbb{E}\Big[G_t \,\Big|\, H_t = h;\, \mathrm{do}(A_t^S = a_t^S),\; A_t^{-S} \sim g_t^{-S}(\cdot \mid h, a_t^S),\; A_{>t} \sim \pi^c\Big]. \tag{2}$$

For a factorized availability-masked policy, the complement kernel can be instantiated as

$$g_t^{-S}(a^{-S} \mid h, a^S) = \prod_{k \notin S} \pi_{\theta_k}\big(a^k \mid h_t^k, M_t^k\big), \tag{3}$$

where illegal actions have zero probability. Equation equation 3 is not assumed for correlated policies; in that case, the joint dependence among the non-members must be retained explicitly.

The quantity in Equation equation 2 is distinct from the observational conditional expectation

$$\mathbb{E}[G_t \mid H_t = h, A_t^S = a_t^S]. \tag{4}$$

The observational quantity can be predictive of return while still differing from the effect of changing the coalition action. In particular, conditioning on A_t^S may preserve dependence induced by an unobserved common cause of the action and return. Equality between the observational and interventional quantities therefore requires the observational identification route in Assumption A6; it is not implied by the Shapley construction itself.

For distinct agents i and j, define the coalition second difference

$$\Delta_{ij} v(S) = v(S \cup \{i,j\}) - v(S \cup \{i\}) - v(S \cup \{j\}) + v(S). \tag{5}$$

The causal teacher interaction is the ordinary pairwise Shapley interaction (Heskes et al., 2020).

$$I_{ij,t}^{\mathrm{do}} = \sum_{S \subseteq N \setminus \{i,j\}} \frac{|S|!(n - |S| - 2)!}{(n-1)!} \Delta_{ij} v_t^{\mathrm{do}}(S). \tag{6}$$

This quantity applies a fixed linear contrast operator to an identified coalition game. It measures the interaction assigned to agents i and j by that operator; it is not claimed to recover a unique causal hyperedge. Higher-order effects can contribute to pairwise Shapley interactions through the allocation rule.

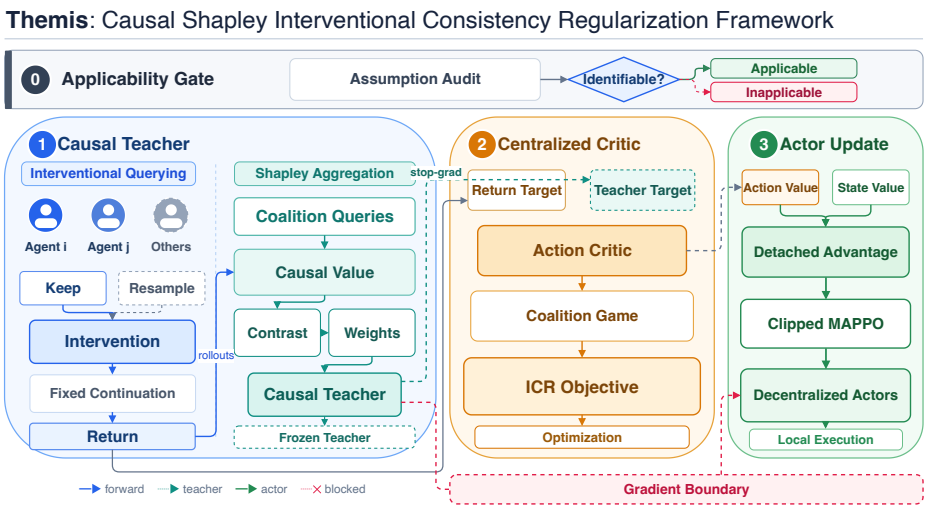


Figure 1: Overview of Themis.

3.3 Three-component Learning Framework

Figure 1 summarizes the complete learning path of Themis.

Causal component. The causal component estimates the coalition values in Equation equation 2 using direct interventional rollouts or a response model validated on held-out interventions. It then applies Equation equation 6 to obtain the teacher target $\widehat{I}^{\mathrm{do}}_{ij,t}$. Teacher data are cross-fitted with respect to the critic and actor data, or the teacher is produced by an explicitly lagged and frozen estimator. This separation prevents the learner from constructing and fitting the same teacher target on an unrestricted shared sample.

Critic component. Let $q_\phi(H, A)$ be a centralized action critic. Using the same complement kernel as the causal component, the critic induces its own coalition value

$$v_t^\phi(S; h, a^S) = \mathbb{E}_{A_t^{-S} \sim g_t^{-S}(\cdot|h, a_t^S)} \left[q_\phi(h, a_t^S, A_t^{-S}) \right]. \tag{7}$$

Applying the same interaction operator to v_t^ϕ produces $I^\phi_{ij,t}$. Matching both the complement kernel and the interaction operator is essential: otherwise, the teacher and critic interactions would refer to different coalition games.

For a preregistered pair set $\mathcal{P}_t$, define the factual return-prediction, interaction-consistency loss as

$$\mathcal{L}_Q = \mathbb{E}\left[\left(q_\phi(H_t, A_t) - Y_t \right)^2 \right], \tag{8}$$

$$\mathcal{L}_{\mathrm{ICR}} = \mathbb{E}\left[\sum_{(i,j)\in\mathcal{P}_t} w_{ij,t} \left(I^\phi_{ij,t} - \mathrm{sg}\left[\widehat{I}^{\mathrm{do}}_{ij,t} \right] \right)^2 \right], \tag{9}$$

where $w_{ij,t} \geq 0$ is a fixed or preregistered pair weight and $\mathrm{sg}[\cdot]$ denotes stop-gradient. The complete critic objective is

$$\mathcal{L}_{\mathrm{critic}} = \mathcal{L}_Q + \lambda_{\mathrm{ICR}} \mathcal{L}_{\mathrm{ICR}} + \lambda_V \mathcal{L}_V, \tag{10}$$

where $\mathcal{L}_V$ is the state-value regression loss. The query set, weights, interaction order, regularization coefficients, critic architecture, return target, and teacher budget are fixed before evaluation. A zero-valued teacher target remains an active constraint rather than being removed, since the absence of an interaction is itself structural information.

Actor component. The actor uses the centralized critic only through a detached advantage,

$$\widehat{\mathcal{A}}_t = \mathrm{sg}\left[q_\phi(H_t, A_t) - V_\nu(H_t)\right], \tag{11}$$

or its preregistered generalized-advantage analogue. The actors are then updated with the standard clipped MAPPO objective. No Shapley value, interaction residual, causal-teacher parameter, response-model parameter, or intervention gradient enters the actor loss directly. The causal target therefore calibrates the centralized value representation instead of acting as an additional reward or actor-side control variate.

3.4 Staged Construction and Optimization

The three components are implemented through the following four stages.

Stage 0: Applicability gate. Before fitting the teacher or critic, we verify the temporal ordering of H_t, the legality of every queried action, the continuation-policy convention, and the provenance of the teacher and learner samples. When observational identification is used, we additionally check the declared history-sufficiency and positivity conditions. Failure of this gate invalidates the causal interpretation and stops the method before optimization.

Stage 1: Causal-teacher construction. For the preregistered coalition and pair queries, we estimate $v_t^{\mathrm{do}}(S)$ using direct rollouts or an interventionally calibrated response model. The fixed Shapley operator in Equation equation 6 then produces $\widehat{I}_{ij,t}^{\mathrm{do}}$. Teacher estimates are stored together with the intervention seed, complement-action samples, legal-action masks, continuation-policy identifier, and sampling budget, allowing the query process to be audited independently of policy learning.

Stage 2: Centralized-critic calibration. We evaluate the critic-induced coalition values in Equation equation 7 with the same complement kernel used by the teacher. The critic receives gradients from factual return prediction, state-value regression, and the consistency penalty in Equation equation 10. The causal teacher remains frozen throughout this update.

Stage 3: Decentralized-policy update. We construct the detached advantage in Equation equation 11 and apply an ordinary clipped MAPPO update (Yu et al., 2022; Kuba et al., 2022). The policy update does not differentiate through the critic, teacher, coalition sampler, or intervention model. During decentralized execution, each actor uses only its local information; neither coalition queries nor intervention-oracle calls are required.

This staged design separates three questions that would otherwise be confounded: whether the coalition target is causally identified, whether the critic represents the selected causal interactions, and whether the calibrated critic improves downstream policy optimization. The first two questions are properties of identification and representation, whereas the third remains an empirical optimization question.

3.5 Theoretical Properties

Proposition 1 (Queried interaction calibration)**.** *Suppose Assumptions A1–A6 hold. If the causal teacher is exact, the critic-induced coalition values are marginalized exactly with the same complement kernel, and* $\mathcal{L}_{\mathrm{ICR}} = 0$*, then*

$$I_{ij,t}^{\phi} = I_{ij,t}^{\mathrm{do}} \tag{12}$$

for every queried pair $(i, j) \in \mathcal{P}_t$ *with* $w_{ij,t} > 0$*.*

Proof sketch. Every summand in equation 9 is nonnegative. Zero weighted loss therefore implies $I_{ij,t}^{\phi} = \widehat{I}_{ij,t}^{\mathrm{do}}$ for each queried pair with positive weight. Exactness of the teacher gives $\widehat{I}_{ij,t}^{\mathrm{do}} = I_{ij,t}^{\mathrm{do}}$, which yields the claim. The proposition guarantees equality only for the queried interactions; it does not identify unqueried pairs or determine the complete critic outside the constrained contrasts.

Proposition 2 (Causal discrimination)**.** *Suppose there is a positive-probability region in which the observational conditional return and policy-interventional return differ, and suppose this difference survives at least one selected interaction contrast. Then a critic can attain the observational squared-error optimum while retaining nonzero interaction-consistency loss for that contrast. Consequently, observational prediction accuracy alone does not certify causal interaction consistency.*

Proof sketch. Under squared-error regression, the population-optimal observational predictor is the conditional mean of the return given its inputs. If this conditional mean differs from the interventional response, marginalizing the predictor with the declared complement kernel does not in general recover $v_t^{\mathrm{do}}(S)$. Because the Shapley interaction is a linear combination of coalition values, any discrepancy that survives the selected contrast produces $I_{ij,t}^{\phi} \neq I_{ij,t}^{\mathrm{do}}$ and hence a positive term in Equation equation 9. If the observational and interventional targets coincide, or if their difference is annihilated by the selected contrast, the proposition is non-discriminating by design; such cases form an important null control.

Neither proposition guarantees improved policy return. Proposition 1 states a representation-level calibration result, and Proposition 2 explains why observational return prediction is insufficient to establish that result. Whether the calibrated representation improves optimization depends on the critic architecture, regularization strength, approximation error, sampling variance, and actor–critic dynamics, and must therefore be evaluated empirically.

3.6 Computational and Applicability Boundaries

Exact coalition construction requires up to 2^n values. Our sampled implementation uses M permutations and K complement-action samples, with cost scaling with these budgets, queried prefixes, and selected pairs. Caching shared coalition values and restricting $\mathcal{P}_t$ reduce computation but introduce teacher variance, which we report separately.

This overhead is confined to centralized training; decentralized execution requires neither coalition evaluation nor intervention queries. The method nevertheless assumes access to a valid simulator intervention or a response model calibrated on held-out interventional data, and does not guarantee causal validity under unmodeled real-world dynamics. The current estimand assumes a one-step intervention, a fixed continuation policy, and selected pairwise interactions. Extensions to temporal credit, higher-order effects, heterogeneous continuous control, or longer interventions require redefining and validating the coalition kernel.

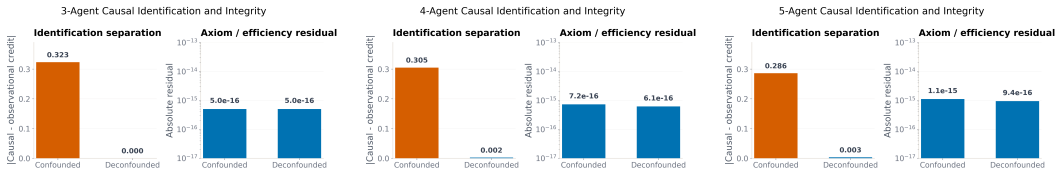
4 Experiments

4.1 Experimental Setup

Causal identification study. We use a controlled structural causal model (SCM) containing simultaneous agent actions and an episode-level common cause that affects both behavior and the team outcome. In the *confounded* condition, the common cause is excluded from the credit history, allowing observational conditioning to mistake association for an action effect. In the matched *deconfounded* control, the common cause is exposed only to the centralized credit estimator, while actor observations, action spaces, episode initialization, and reward scale remain unchanged. We compare direct-interventional and observational credit and additionally report the Shapley efficiency residual to separate coalition semantics from numerical or allocation errors.

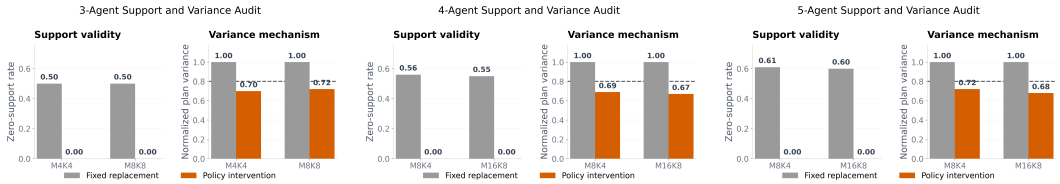
Frozen-checkpoint mechanism audit. We freeze the policy and compare fixed action replacement with policy-supported intervention. Fixed replacement assigns an externally chosen default action to agents outside a coalition, whereas policy intervention samples their actions from the current availability-masked policy. For three agents, we evaluate sampling budgets M4K4 and M8K8; for four and five agents, we use M8K4 and M16K8, where M is the number of sampled permutations and K is the number of complement-action samples. These budgets follow standard Shapley approximation practice (Castro et al., 2009; Kolpaczki et al., 2024). We report the zero-support query rate and plan variance normalized to fixed replacement at the same budget.

Cooperative learning environment. We evaluate downstream learning on the Simple Spread task from the Multi-Agent Particle Environment (MPE). Each setting contains n agents and n landmarks. Agents act from local observations and must coordinate so that all landmarks are covered while avoiding collisions. Because the agents share a single team reward, successful behavior requires coordinated assignment rather than independent greedy movement.



(a) 3 agents (b) 4 agents (c) 5 agents

Figure 2: Causal Identification and Shapley Efficiency results. The causal–observational credit gap is large only in the confounded condition and nearly vanishes after deconfounding, while the efficiency residual remains at numerical precision.



(a) 3 agents (b) 4 agents (c) 5 agents

Figure 3: Frozen-checkpoint support and variance audit. Policy intervention produces no zero-support queries and yields lower normalized plan variance than fixed replacement.

Across all learning conditions, we hold the factual trajectories, actor–critic architecture, optimization schedule, random seeds, and evaluation protocol fixed. The causal mechanism is used only during centralized training. At evaluation time, actors receive only local observations and make no intervention-oracle calls.

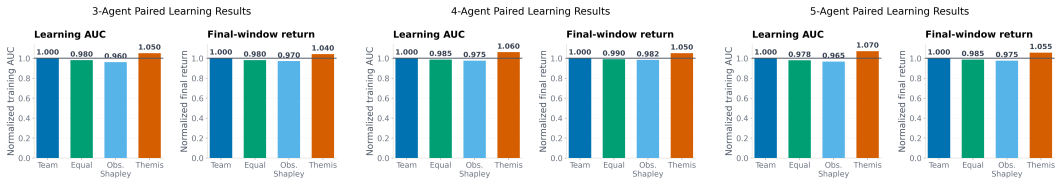
Baselines. We compare four matched reward conditions. *Team reward* retains the original cooperative objective without individualized credit and serves as the normalization reference. *Equal split* distributes the factual team reward uniformly and controls for per-agent reward scale and reward conservation. *Observational Shapley* uses the same coalition structure, Shapley weights, learner, and normalization component as the direct-intervention condition, but constructs coalition values through observational conditioning, consistent with established Shapley-based MARL baselines (Li et al., 2021; Wang et al., 2022; Chen et al., 2024). *Themis* instead uses policy-supported direct-interventional coalition values. Fixed replacement is included separately in the frozen-checkpoint audit because it tests the validity of the counterfactual query mechanism rather than the downstream learning component. All learning metrics are normalized to the team-reward baseline, which is fixed at 1.0.

4.2 Causal Identification and Shapley Efficiency

Figure 2 tests whether the proposed coalition value captures causal intervention rather than predictive association. In the confounded condition, the absolute causal–observational credit gaps are 0.323, 0.305, and 0.286 for three, four, and five agents, respectively. After the common cause is exposed, the gaps decrease to 0.000, 0.002, and 0.003. The separation therefore appears when coalition actions provide evidence about an omitted common cause and nearly disappears when that source of confounding is removed. Meanwhile, the efficiency residual remains between 5.0×10^{-16} and 1.1×10^{-15}. Both allocations therefore satisfy Shapley efficiency up to numerical precision. The observed difference is attributable to the semantics of the coalition value rather than failed Shapley aggregation, reward normalization, or an incomplete coalition ledger.

4.3 Intervention Support and Sampling Stability

Figure 3 examines how counterfactual coalitions are instantiated at a fixed policy checkpoint. Fixed replacement produces zero-support rates of 0.50 at both three-agent budgets, $0.56/0.55$ for four agents, and $0.61/0.60$ for five agents. Policy intervention remains at 0.00 in every setting. Sampling



(a) 3 agents (b) 4 agents (c) 5 agents

Figure 4: Paired MPE learning results normalized to the team-reward baseline. Themis obtains the highest learning AUC and final-window return across all three team sizes.

complement actions from the availability-masked policy therefore avoids action combinations that the current policy cannot generate.

Relative to the fixed-replacement variance of 1.00, the normalized plan variance of policy intervention is $0.70/0.72$, $0.69/0.67$, and $0.72/0.68$ for three, four, and five agents, respectively. These values correspond to variance reductions of 28–33%. The similar reductions at both budgets show that the result is not tied to one permutation–sample configuration. Policy-supported intervention is therefore both behaviorally valid and more stable under the evaluated query budgets.

4.4 Cooperative Learning

Figure 4 evaluates whether causal credit improves policy learning. Across agents, Themis achieves normalized AUC values of 1.050–1.070 and final-window returns of 1.040–1.055, making it the only condition that exceeds the team-reward reference on both metrics. Equal split and observational Shapley remain below the reference, ruling out reward scaling and Shapley aggregation alone as explanations for the improvement. The larger AUC gains suggest that direct-interventional credit primarily accelerates coordination, with consistent benefits across all team sizes.

4.5 Team-Size Trends, Cost, and Scope

The different agents settings exhibit the same qualitative pattern: direct-interventional targets remain separated from observational targets under confounding, policy-supported interventions avoid zero-support queries, and Themis improves the matched learning metrics. These results provide evidence that the mechanism is not confined to a single team size. The additional causal-teacher evaluations nevertheless incur a training-time cost. Exact coalition enumeration requires 2^n coalition values. With M sampled permutations, K complement-action samples, and a selected pair set $\mathcal{P}_t$, the cost increases with the Shapley sampling budget and the number of queried interactions. This cost is confined to centralized training: during decentralized execution, each actor receives only its local observation and does not access the causal teacher, coalition values, or intervention oracle. Themis are correspondingly bounded. The experiments establish that, on the evaluated SCM controls and MPE Simple Spread settings, causal interaction consistency can improve the matched critic-learning pipeline. They do not establish universal gains for arbitrary team sizes, heterogeneous agents, continuous-control domains, or learned response models.

5 Conclusion

Themis frames causal credit assignment as a critic-calibration problem. It identifies selected policy-interventional Shapley interactions, constrains a centralized critic to match them, and preserves a clean detached advantage for decentralized policy optimization. The framework supplies a transparent route from one-step action intervention to long-horizon coalition value and interaction-consistent critic estimates, linking classical Shapley allocation with intervention-based attribution. Its guarantees remain conditional on the declared causal assumptions and queried pairs, while its computational and optimization consequences require independent empirical validation.

REFERENCES

Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009. doi: 10.1016/j.cor.2008.04.004.

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

Yu Fan Chen, Miao Liu, Michael Everett, and Jonathan P. How. Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning. In *2017 IEEE International Conference on Robotics and Automation*, pp. 285–292, 2017. doi: 10.1109/ICRA.2017.7989037.

Tianshu Chu, Jie Wang, Lara Codeca, and Zhaojian Li. Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, 21(3): 1086–1095, 2020. doi: 10.1109/TITS.2019.2901791.

Ao Ding, Licheng Sun, Yongjie Hou, Huaqing Zhang, and Hongbin Ma. A historical interaction-enhanced Shapley policy gradient algorithm for multi-agent credit assignment, 2025. URL https://arxiv.org/abs/2511.07778.

Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11794.

Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric Shapley values: Incorporating causal knowledge into model-agnostic explainability. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/1910.06358.

Sven Gronauer and Klaus Diepold. Multi-agent deep reinforcement learning: A survey. *Artificial Intelligence Review*, 55(2):895–943, 2022. doi: 10.1007/s10462-021-09996-w.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, FL, 2020. URL https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.

Pablo Hernandez-Leal, Bilal Kartal, and Matthew E. Taylor. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 33(6):750–797, 2019. doi: 10.1007/s10458-019-09421-1.

Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal Shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2011.01625.

Yun Hua, Haosheng Chen, Shiqin Wang, Wenhao Li, Xiangfeng Wang, and Jun Luo. ShapleyCoop: Credit assignment for emergent cooperation in self-interested LLM agents, 2025. URL https://arxiv.org/abs/2506.07388.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable AI: A causal problem. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2907–2916, 2020. URL https://proceedings.mlr.press/v108/janzing20a.html.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro A. Ortega, D. J. Strouse, Joel Z. Leibo, and Nando de Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3040–3049, 2019. URL https://proceedings.mlr.press/v97/jaques19a.html.

Patrick Kolpaczki, Viktor Bengs, Maximilian Muschalik, and Eyke Hüllermeier. Approximating the Shapley value without marginal contributions. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2302.00736.

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. URL https://arxiv.org/abs/2109.11251.

Jiahui Li, Kun Kuang, Baoxiang Wang, Furui Liu, Long Chen, Fei Wu, and Jun Xiao. Shapley counterfactual credits for multi-agent reinforcement learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2980–2990, 2021. doi: 10.1145/3447548.3467420.

Yongheng Liang, Hejun Wu, Haitao Wang, and Hao Cai. Asynchronous credit assignment for multi-agent reinforcement learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 170–178, 2025. doi: 10.24963/ijcai.2025/20.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1705.07874.

Omayma Mahjoub, Ruan de Kock, Siddarth Singh, Wiem Khlifi, Abidine Vall, Kale-ab Tessera, and Arnu Pretorius. Efficiently quantifying individual agent importance in cooperative MARL, 2023. URL https://arxiv.org/abs/2312.08466. AAAI XAI4DRL Workshop.

Yasar Sinan Nasir and Dongning Guo. Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks. *IEEE Journal on Selected Areas in Communications*, 37(10): 2239–2250, 2019. doi: 10.1109/JSAC.2019.2933973.

Guillermo Owen. Multilinear extensions of games. *Management Science*, 18(5):P64–P79, 1972. doi: 10.1287/mnsc.18.5.64.

Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://arxiv.org/abs/2006.07869.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017. URL https://mitpress.mit.edu/9780262037310/elements-of-causal-inference/.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4295–4304, 2018. URL https://proceedings.mlr.press/v80/rashid18a.html.

Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philip H. S. Torr, Jakob Foerster, and Shimon Whiteson. The StarCraft multi-agent challenge. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2186–2188, 2019. URL https://arxiv.org/abs/1902.04043.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games II*, pp. 307–317. Princeton University Press, 1953. doi: 10.1515/9781400881970-018.

Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665, 2014. doi: 10.1007/s10115-013-0679-x.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech M. Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2085–2087, 2018. URL https://arxiv.org/abs/1706.05296.

Jianhong Wang. *Shapley Value Based Multi-Agent Reinforcement Learning: Theory, Method and Its Application to Energy Network*. PhD thesis, Imperial College London, 2024.

Jianhong Wang, Yuan Zhang, Yunjie Gu, and Tae-Kyun Kim. SHAQ: Incorporating Shapley value theory into multi-agent Q-learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 5941–5954, 2022. doi: 10.52202/068431-0430.

Jianhong Wang, Yang Li, Samuel Kaski, and Jonathan Lawry. Shapley machine: A game-theoretic framework for n-agent ad hoc teamwork, 2025a. URL https://arxiv.org/abs/2506.11285.

Ziyan Wang, Yali Du, Yudi Zhang, Meng Fang, and Biwei Huang. MACCA: Offline multi-agent reinforcement learning with causal credit assignment. *Transactions on Machine Learning Research*, 2025b. URL https://arxiv.org/abs/2312.03644.

David H. Wolpert and Kagan Tumer. Optimal payoff functions for members of collectives. In *Modeling Complexity in Economic and Social Systems*, pp. 355–369. World Scientific, 2002. URL https://doi.org/10.1142/9789812777263_0020.

Qinyu Xu, Yuanyang Zhu, Xuefei Wu, and Chunlin Chen. High-order interactions modeling for interpretable multi-agent Q-learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 173911–173943, 2025. doi: 10.52202/085713-5233.

Yingxuan Yang, Bo Huang, Siyuan Qi, Chao Feng, Haoyi Hu, Yuxuan Zhu, Jinbo Hu, Haoran Zhao, Ziyi He, Xiao Liu, Muning Wen, Zongyu Wang, Lin Qiu, Xuezhi Cao, Xunliang Cai, Yong Yu, and Weinan Zhang. Understanding and optimizing agentic workflows via Shapley value, 2025. URL https://arxiv.org/abs/2502.00510.

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. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2103.01955.