A Causality-Inspired Spatial-Temporal Return Decomposition Approach for Multi-Agent Reinforcement Learning

Extended Abstract

Yudi Zhang
Eindhoven University of Technology
Eindhoven, the Netherlands
y.zhang5@tue.nl

Yali Du
King's College London
London, United Kingdom
yali.du@kcl.ac.uk

Biwei Huang
University of California San Diego
California, United States
bih007@ucsd.edu

Meng Fang
University of Liverpool
Liverpool, United Kingdom
Meng.Fang@liverpool.ac.uk

Mykola Pechenizkiy
Eindhoven University of Technology
Eindhoven, the Netherlands
m.pechenizkiy@tue.nl

ABSTRACT

Cooperative multi-agent reinforcement learning (MARL) has achieved strong performance, but it remains limited in explaining how agents' decisions contribute to outcomes. This limitation is especially acute under delayed, episodic rewards, where credit must be assigned across both time and agents. We propose **CA**usally-inspired **S**patial-**T**emporal return decomposition (**CAST**) for episodic cooperative MARL. CAST provides an interpretable decomposition while relaxing common assumptions on multi-agent reward structure. Temporally, the episodic return is expressed as a sum of per-timestep team rewards. Spatially, team rewards are modeled as general nonlinear mixtures of individual rewards rather than simple additive forms, enabling more flexible and accurate credit assignment. We show that team rewards, individual rewards, and the underlying causal relations are identifiable under our framework, yielding structural constraints that improve interpretability. Experiments on MPE and variants demonstrate state-of-the-art performance and qualitative visualizations that reveal meaningful causal structure.

KEYWORDS

Multi-agent, Spatial-Temporal Credit Assignment, Causality

ACM Reference Format:
Yudi Zhang, Yali Du, Biwei Huang, Meng Fang, and Mykola Pechenizkiy. 2026. A Causality-Inspired Spatial-Temporal Return Decomposition Approach for Multi-Agent Reinforcement Learning: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/LWGT2184

1 INTRODUCTION

Cooperative multi-agent reinforcement learning (MARL) enables agents to learn coordinated policies toward a shared team objective [10, 20, 22, 23], with successes in games [13, 17, 19] and robotics [8, 15]. A central difficulty is *credit assignment*: each agent affects others through coupled dynamics, yet learning is driven by a single team reward [3, 6, 11, 16, 18]. This becomes particularly challenging when rewards are sparse and delayed until the end of an episode. Recent spatial-temporal approaches redistribute credit across agents and time to mitigate this issue [2, 14].

Despite progress, many methods provide limited interpretability about *why* particular credits are assigned. While some works leverage structural information for learning [4, 7], most spatial-temporal decompositions rely on a strict additive assumption that the team reward equals the sum of individual rewards [2, 14]. This assumption can yield identifiability, but it may be overly restrictive: in many tasks, team outcomes arise from unknown nonlinear interactions among agents. For example, in a team game, a medic's support can be essential for attackers' success, yet its contribution is not naturally captured by a simple linear decomposition.

We propose **CA**usality-inspired **S**patial-**T**emporal return decomposition (**CAST**) to provide interpretable credit assignment under sparse episodic rewards. CAST decomposes return temporally by attributing long-term outcomes to per-timestep team rewards [1, 12, 21], and decomposes team rewards spatially by modeling them as the causal effect of agents' (latent) individual contributions. Unlike prior work [2], CAST relaxes the additive assumption and models the team reward as a nonlinear *invertible* mixture of individual rewards, inspired by iVAE [5]. By mapping individual rewards to transformed rewards that sum to the team reward, CAST preserves identifiability of both the causal structure and the unobserved individual reward functions, enabling explicit structural constraints for interpretability.

Our contributions are: (1) a nonlinear invertible mixture framework for spatial-temporal credit assignment with identifiability guarantees; (2) explicit estimation of causal relationships underlying individual reward generation for interpretability; (3) an iVAE-based individual reward predictor under the proposed mixture model; and (4) state-of-the-art results on MPE and variants, with visualizations that reveal meaningful causal structure.

2 CAST: CAUSALITY-INSPIRED SPATIAL-TEMPORAL RETURN DECOMPOSITION

We focus on enhancing policy learning through explicit credit assignments in cooperative games with sparse and delayed team

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/LWGT2184

rewards, especially episodic ones. We begin by describing the generative process in cooperative games, which lays the foundation for our proposed approach. **Generative Model.** CAST exploits a Dynamic Bayesian Network (DBN) [9], $\mathcal{G}$, over a finite number of random variables,

$$\{\underbrace{\{s_{1,t},\cdots,s_{|s|,t}\}}_{\text{joint state}}\cup\underbrace{\{a_t^1,\cdots a_t^N\}}_{\text{agents' action}}\cup\underbrace{\{r_t^1,\cdots,r_t^N\}}_{\text{individual rewards}}\cup \underbrace{\{\tilde{r}_t^1,\cdots,\tilde{r}_t^N\}}_{\text{transformed individual rewards}}\}_{t=1}^{T}\bigcup Q,$$

where $|s|$ and $|a^n|$ are the dimension of s_t and a_t^n, N is the number of agents, and $\mathcal{G}$ characterizes the underlying generative process in MARL as follows,

$$\begin{cases}\text{individual reward:} & r_t^n = f^n(C^n \odot s_t, a_t^n, \epsilon_{r,n,t})\\ \text{team reward:} & R_t = \sum_{n=1}^{N} g^n([r_t^1, r_t^2, \cdots, r_t^N])\\ \text{long-term return:} & Q = \sum_{t=1}^{T}\gamma^{t-1}R_t\end{cases} \tag{1}$$

where $\odot$ is the element-wise product. Note that such causal modeling relaxes the previous strict linear assumption [2], which can be regarded as a special case of our model when g is an identity function. In our experimental environments, where the state is not available during training, we use the agents' observations o_t as a proxy for the environmental state s_t and agents' index n.

Notations. For simplicity, we define,

$$\tilde{r}_t^n = g^n(r_t), \quad r_t = [r_t^1,\cdots,r_t^N], \quad \tilde{r}_t = [\tilde{r}_t^1,\cdots,\tilde{r}_t^N]. \tag{2}$$

We denote by r_t^n the individual reward at time step t of agent n. In the rest of the paper, we call $\tilde{r}_t^n$ as *transformed individual rewards*. Q is the trajectory-wise long-term return. T is the maximum episode length of the environment. $\epsilon_{r,n,t}$ is the *i.i.d.* noise.

Causal structure and interpretability. $C^n, \forall n \in [1,\cdots,N]$ is a binary mask to capture the causal structure between the elements of joint state and individual rewards of agents, with $C^n \in \{0,1\}^{|s|}$. C^n controls if a specific dimension of the state s_t impacts the individual reward r_t^n at timestep t. Let C_k^n be the k-th element in the vector C^n. If there is an edge from the k-th dimension of s_t to the agent n's individual reward r_t^n in $\mathcal{G}$, then $C_k^n = 1$. Given $\mathcal{G}$, we can naturally explain how the individual rewards are generated, *i.e.*, the explicit contribution of each dimension of the joint state towards individual rewards.

Functions in Eq. 1. f is the unknown individual reward function, whose output r_t^n is expected to accurately describe the contribution of agent n, and serves as the reward signals for independent policy learning. g is an invertible function to generate the transformed individual rewards $\tilde{r}_t$ from individual rewards r_t, *i.e.*, $\tilde{r}_t$ is a nonlinear or linear invertible mixture of r_t. We assume that the sum of transformed individual rewards $\sum_{n=1}^{N}\tilde{r}_t^n$ equals Q, where we follow previous work to ignore the discount factor γ [12].

Beyond Linear Summation: A Relaxed Invertibility Condition. We want to highlight the complexity of team reward generation in a multi-agent system, which can range from simple linear sums to complex nonlinear functions. This contrasts with previous work, such as STAS [2], which assumes that the team reward equals the sum of individual rewards. Such a restrictive assumption limits the reasonable and precise credit assignment in the complex MARL system. In contrast, our proposed framework allows the team reward to be a general mixture of individual rewards r_t^n, with the linear assumption being a special case when $r_t = \tilde{r}_t$.

Below we provide 1) the identifiability results of the unknown functions and structures in Eq. 1, which support the estimation of the causal structure from the data, enabling the interpretability of our method; 2) the equivalence of using the decomposed rewards for policy learning in our proposed framework.

Proposition 2.1 (Identifiability for Spatial-Temporal Credit Assignment)**.** *Consider the data generating process in Eq. 1. Suppose the joint state s_t, the action a_t^n for each agent n and the long-term return (can be calculated by the discounted sum of delayed rewards) are observable, while the individual r_t^n for each agent n and team reward R_t are unobserved.*

Under the Markov condition and faithfulness assumption , if the function g for generating the transformed individual rewards is invertible, then the causal mask C^n is identifiable and we can identify the individual rewards r_t^n to their monolithic invertible transformations, e.g. $\log(r_t^n)$.

Remark. The proposition 2.1 shows that we can identify causal structures and individual rewards (up to their invertible nonlinear transformation) from the observed data, as long as **the mixture function g is invertible**.

Proof Sketch. The proof begins by establishing the identifiability of the transformed individual rewards, represented by $\tilde{r}^n$, indicating the possibility of recovering it from the data. The second part of the proof highlights the relationship between our method and nonlinear Independent Component Analysis (ICA), along with confirming the identifiability of individual rewards, r_t^n. The full proof is in Appendix. □

Violation of Identifiability Conditions. Awareness of the consequences of violating the identifiability conditions can help understand the theoretical strength of our work: any violation of these conditions will result in the unidentifiable individual rewards, which means that they cannot be uniquely recovered. Therefore, while adopting the additive assumption in a setting where the individual rewards are mixed in a much more complex way, the individual rewards are not recoverable, thus limiting the applicability of those methods. In contrast, our framework demonstrates that identifiability remains achievable under weaker structural assumptions, as long as the mixing function is invertible. This greatly expands the class of environments and reward structures to which multi-agent credit assignment methods can be applied.

Since the individual rewards are one-dimensional, *i.e.,* they are scalars, their monolithic invertible transformations are equivalent to their monolithic functions. However, monolithity could be positive or negative. Therefore, we give Proposition 2.2 to show that if the individual rewards recovered are positively correlated with their ground truth, learning a policy with estimated individual rewards can induce the same optimal policy as the policy learning under the ground truth individual rewards.

Proposition 2.2. *If $k(\cdot)$ is a monotonically increasing invertible transformation, then it is equivalent to optimizing the policy using the ground individual rewards r_t^n and its k-based transformation $k(r_t^n)$.*

REFERENCES

[1] Jose A Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. 2019. RUDDER: Return decomposition for delayed rewards. *Advances in Neural Information Processing Systems* 32 (2019).

[2] Sirui Chen, Zhaowei Zhang, Yali Du, and Yaodong Yang. 2023. STAS: Spatial-Temporal Return Decomposition for Multi-agent Reinforcement Learning. *arXiv preprint arXiv:2304.07520* (2023).

[3] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[4] Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, and Kun Zhang. 2021. AdaRL: What, Where, and How to Adapt in Transfer Reinforcement Learning. *CoRR* abs/2107.02729 (2021). arXiv:2107.02729 https://arxiv.org/abs/2107.02729

[5] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. 2020. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2207–2217.

[6] Jiahui Li, Kun Kuang, Baoxiang Wang, Furui Liu, Long Chen, Fei Wu, and Jun Xiao. 2021. Shapley counterfactual credits for multi-agent reinforcement learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 934–942.

[7] Boyin Liu, Zhiqiang Pu, Yi Pan, Jianqiang Yi, Yanyan Liang, and D. Zhang. 2023. Lazy Agents: A New Perspective on Solving Sparse Reward Problem in Multi-agent Reinforcement Learning. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 21937–21950. https://proceedings.mlr.press/v202/liu23ac.html

[8] Boyi Liu, Lujia Wang, and Ming Liu. 2019. Lifelong federated reinforcement learning: a learning architecture for navigation in cloud robotic systems. *IEEE Robotics and Automation Letters* 4, 4 (2019), 4555–4562.

[9] Kevin Patrick Murphy. 2002. *Dynamic bayesian networks: representation, inference and learning*. University of California, Berkeley.

[10] Afshin Oroojlooy and Davood Hajinezhad. 2023. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence* 53, 11 (2023), 13677–13722.

[11] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2020. Monotonic value function factorisation for deep multi-agent reinforcement learning. *The Journal of Machine Learning Research* 21, 1 (2020), 7234–7284.

[12] Zhizhou Ren, Ruihan Guo, Yuan Zhou, and Jian Peng. 2022. Learning Long-Term Reward Redistribution via Randomized Return Decomposition. In *International Conference on Learning Representations*.

[13] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philiph H. S. Torr, Jakob Foerster, and Shimon Whiteson. 2019. The StarCraft Multi-Agent Challenge. *CoRR* abs/1902.04043 (2019).

[14] Jennifer She, Jayesh K Gupta, and Mykel J Kochenderfer. 2022. Agent-time attention for sparse rewards multi-agent reinforcement learning. *arXiv preprint arXiv:2210.17540* (2022).

[15] Daigo Shishika, James Paulos, and Vijay Kumar. 2020. Cooperative team strategies for multi-player perimeter-defense games. *IEEE Robotics and Automation Letters* 5, 2 (2020), 2738–2745.

[16] Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. 2019. Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *International conference on machine learning*. PMLR, 5887–5896.

[17] Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, et al. 2017. Starcraft ii: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782* (2017).

[18] Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. 2021. {QPLEX}: Duplex Dueling Multi-Agent Q-Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Rcmk0xxIQV

[19] Deheng Ye, Guibin Chen, Wen Zhang, Sheng Chen, Bo Yuan, Bo Liu, Jia Chen, Zhao Liu, Fuhao Qiu, Hongsheng Yu, et al. 2020. Towards playing full moba games with deep reinforcement learning. *Advances in Neural Information Processing Systems* 33 (2020), 621–632.

[20] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. 2022. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems* 35 (2022), 24611–24624.

[21] Yudi Zhang, Yali Du, Biwei Huang, Ziyan Wang, Jun Wang, Meng Fang, and Mykola Pechenizkiy. 2023. Interpretable Reward Redistribution in Reinforcement Learning: A Causal Approach. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[22] Meng Zhou, Ziyu Liu, Pengwei Sui, Yixuan Li, and Yuk Ying Chung. 2020. Learning implicit credit assignment for cooperative multi-agent reinforcement learning. *Advances in neural information processing systems* 33 (2020), 11853–11864.

[23] Roy Zohar, Shie Mannor, and Guy Tennenholtz. 2022. Locality matters: A scalable value decomposition approach for cooperative multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 9278–9285.