Diverse Heterogeneous Graph Conditioned Diffusion for Multi-Agent Teaming

Extended Abstract

Luis Pimentel
Georgia Institute of Technology
Atlanta, USA
lpimentel3@gatech.edu

Sean Ye
Georgia Institute of Technology
Atlanta, USA
seancye@gatech.edu

James Ellis Grant Pagan
Sandia National Laboratories
Albuquerque, New Mexico, USA
jepagan@sandia.gov

Matthew Gombolay
Georgia Institute of Technology
Atlanta, USA
matthew.gombolay@cc.gatech.edu

ABSTRACT

Diverse multi-agent teams have the potential to solve complex tasks by learning effective teaming through reinforcement learning (RL). The high variability of interactions across team compositions poses scalability and real-world applicability challenges for online methods, highlighting the need for offline approaches that learn from pre-collected datasets. However, it is challenging to effectively leverage diverse data, adapt across team compositions using only offline data, and maintain decentralization during online deployment. To address these challenges, we present **H**eterogeneous **G**raph **C**onditioned **D**iffusion (**HGCD**), a multi-agent diffusion model that leverages the conditional generative modeling abilities of diffusion and heterogeneous multi-agent communication to learn generalizable policies offline, while ensuring decentralized execution online. We demonstrate the effectiveness of our method on StarCraft II Multi-Agent Challenge v2 (SMACv2) tasks, achieving superior generalization performance over prior state-of-the-art.

KEYWORDS

offline reinforcement learning; diffusion models; multi-agent communication; multi-agent generalization

ACM Reference Format:
Luis Pimentel, Sean Ye, James Ellis Grant Pagan, and Matthew Gombolay. 2025. Diverse Heterogeneous Graph Conditioned Diffusion for Multi-Agent Teaming: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

Heterogeneous teaming is essential learning to solve complex real-world problems where agents with diverse capabilities must work together to achieve shared objective [10, 16]. One prominent area is in search and rescue missions [7], where different types of robots can be deployed to leverage their unique sensing and mobility capabilities: drones for surveillance [23], ground robots for reconnaissance, and marine robots for underwater missions. In critical settings like this, directly interacting with the environment poses safety risks, deployment can be costly [14, 17], and decentralization capabilities are essential for robustness. In these scenarios, offline multi-agent reinforcement learning (MARL) is valuable for enabling agents to learn effective teaming from datasets. However, generalization remains a challenge, as the offline data must sufficiently capture the diversity of interactions across various heterogeneous team compositions [12]. When the data lacks this diversity, learning becomes significantly more difficult, highlighting the need for methods that can adapt policies to unseen team compositions. We address these challenges through a diffusion-based offline Meta-MARL architecture, integrating heterogeneous communication for diverse and decentralized multi-agent coordination.

2 METHOD

Communication Message Encoding. We integrate multi-agent communication [4, 18, 21] with graph-based mechanisms [15, 16] into diffusion, learning to extract task-relevant information from the heterogeneous graph structure, $\mathcal{G}_\tau^m$. We aim to learn from offline data samples $d_m = \left[\mathcal{G}_\tau^m, \mathcal{R}_\tau^m, \bar{\mathbf{a}}_\tau^m, \bar{\mathbf{o}}_\tau^m\right]$, containing reward, action and observation trajectories where m specifies the team composition. We follow the Denoising Diffusion Probabilistic Model (DDPM) formulation [8, 19, 20] extended to the reinforcement learning [1, 13] and multi-agent setting [24]. At each diffusion time-step k, each agent i of class j encodes its observation histories as a set of messages, $m_r^{i_j} = \text{MessageEncoder}_r^i(\mathbf{x}_k^{i_j})$, across r layers where $\mathbf{x}_k^{i_j} = \left[o_{1:h}^{i_j} || \tilde{x}_{h+1:h+H}^{i_j}\right]$. Here, each $\tilde{x}_t^{i_j}$ is a time-step in the diffusion horizon to be noised during training and denoised during sampling, extending H steps beyond its history h. Agents undergo message-passing via heterogeneous graph-based communication, then apply Heterogeneous Graph Attention [16] with class based node parameters, W_j, class-to-class edge parameters, $W_{l\to j}$, and attention parameters, $W_{l\to j}^{\text{att}}$. First, the normalized attention coefficients, $\alpha_{ik}^{l\to j} = \text{softmax}_k\left(\sigma'\left(W_{l\to j}^{att}\left[W_j m^{i_j} \parallel W_{l\to j} m^{k_l}\right]\right)\right)$, are computed to weigh each neighbor k's message. Then, the communication embeddings, $z_r^{i_j} = \sigma\left(W_j m_r^{i_j} + \sum_{l\in C}\sum_{k\in N_l(i)} \alpha_{ik}^{l\to j} m_r^{k_l}\right)$, are computed. These embeddings function as learned representations of

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA.

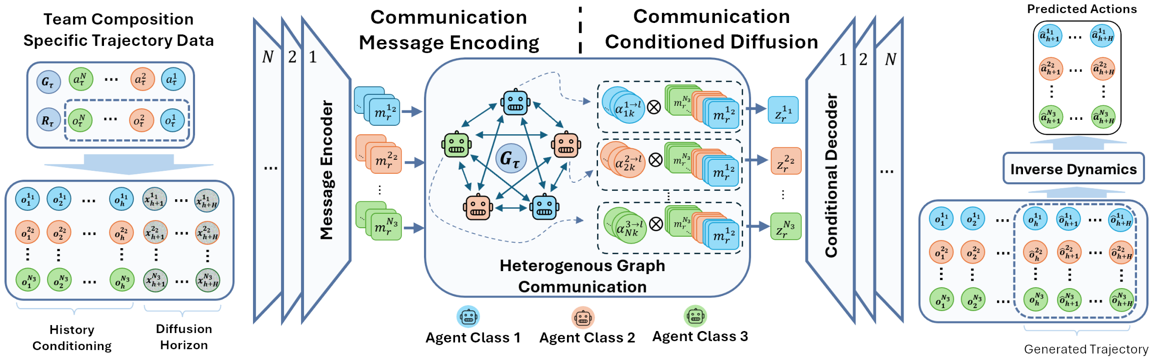

Figure 1: Overview of the architecture for Heterogenous Graph Conditioned Diffusion (HGCD).

the task, as heterogeneous communication provides a suitable basis of task-relevant information about the team composition.

Communication Conditioned Diffusion. We leverage the conditional generative modeling ability of diffusion models to condition not only on constraints like high-returns, as in prior work [1, 24], but on the communication embeddings to provide better grounding for trajectory generation. This is achieved through classifier-free guidance sampling [9], with the predicted noise computed as shown in Eq. 1, where ω is the guidance scale coefficient, and θ parametrizes both unconditional and conditional diffusion models.

$$\hat{\epsilon} := \epsilon_\theta\left(\mathbf{x}_k^{i_j}, \emptyset, k\right) + \omega\left(\epsilon_\theta\left(\mathbf{x}_k^{i_j}, (z_\tau^{i_j}, R_\tau), k\right) - \epsilon_\theta\left(\mathbf{x}_k^{i_j}, \emptyset, k\right)\right) \quad (1)$$

During the denoising process, Eq. 1 is applied iteratively until $\mathbf{x}_0^{i_j} := \left[o_{1:h}^{i_j} \| \hat{o}_{h+1:h+H}^{i_j}\right]$ retrieves each agent's planned trajectory. This guides each agent's diffusion process towards sequences that satisfy the condition of achieving high return within their team composition. Decision-making is inferred via an inverse dynamics model [1], parameterized by ϕ, which estimates each agent's action as $\hat{a}_t^{i_j} := f_\phi(o_t^{i_j}, o_{t+1}^{i_j})$, enabling the transition from $o_t^{i_j}$ to $o_{t+1}^{i_j}$.

Diverse Offline Meta-MARL. Conditionally sampling trajectories requires learning the conditional data distributions across team compositions, which is enabled through the loss objective $\mathcal{L}_D(\epsilon, \bar{\mathbf{x}}_k^m, \bar{\mathbf{z}}_\tau^m, R_\tau^m, \beta, k; \theta) = \left\|\epsilon - \epsilon_\theta(\bar{\mathbf{x}}_k^m, (1-\beta)(\bar{\mathbf{z}}_\tau^m, R_\tau^m) + \beta\emptyset, k)\right\|^2$ of the diffusion model. Additionally, the inverse dynamics model is learned through the loss objective $\mathcal{L}_I(\bar{\mathbf{a}}_\tau^m, \bar{\mathbf{o}}_{\tau-1}^m, \bar{\mathbf{o}}_\tau^m; \phi) = \left\|\bar{\mathbf{a}}_\tau^m - f_\phi(\bar{\mathbf{o}}_{\tau-1}^m, \bar{\mathbf{o}}_\tau^m)\right\|^2$. Both are combined and applied across team compositions, formulating the offline meta-reinforcement learning objective with both offline outer and inner loops [2], $\mathcal{L}_{\text{train}}(\theta, \phi) = \mathbb{E}_{d_m \sim \mathcal{D}_m}\left[\mathbb{E}_{\bar{\mathbf{z}}_\tau^m, \epsilon, k, \beta}\left[\mathcal{L}_D(\epsilon, \bar{\mathbf{x}}_k^m, \bar{\mathbf{z}}_\tau^m, R_\tau^m, \beta, k)\right] + \mathcal{L}_I(\bar{\mathbf{a}}_\tau^m, \bar{\mathbf{o}}_{\tau-1}^m, \bar{\mathbf{o}}_\tau^m)\right]$. The *outer-loop*, consists of the overall architecture, learning to generate high-return trajectories across team compositions, and the *inner-loop* consists of the HetGAT layers that learn to produce communication embeddings for enhanced adaptability.

3 EVALUATION AND RESULTS

We evaluate on two SMACv2 [5] scenarios with 5 agent teams spanning 20 distinct team compositions across 3 classes. Offline training uses OG-MARL [6] datasets with Percentage Filtering [3] to address the abundance of suboptimal trajectories [11]. Win rates are measured over fifty online episodes per seed for each team composition, using three seeds. Baseline comparisons include: Behavior Cloning (BC) and Implicit Constraint Q-Learning [22] which assume decentralized execution, and Multi-Agent Diffusion [24] which can be executed both centralized and decentralized. Experiments cover centralized (C) and decentralized (D) execution modes, as well as full data (F) and limited data (L) training with seen and unseen team compositions. As shown Table 1, our method results in considerable performance improvements over baselines.

	terran 5 vs. 5				zerg 5 vs. 5			
Baseline	C/F	C/L	D/F	D/L	C/F	C/L	D/F	D/L
BC	-	-	185.68	157.86	-	-	141.87	164.24
IQL	-	-	2.87	9.06	-	-	78.79	26.85
MAD	25.39	87.58	76.44	229.30	18.70	17.20	38.65	82.20
All	25.39	87.58	2.13	1.97	18.70	17.20	25.67	1.10

Table 1: Mean percent improvement of HGCD over baselines.

4 DISCUSSION AND CONCLUSION

The increased performance of HGCD highlights its ability to adapt decision-making for more effective trajectory across diverse team compositions. In limited data settings, HGCD leverages the structural information in the heterogeneous graph network to capture relation interactions among different agent classes, facilitating generalization beyond the compositions encountered during training. This is particularly important in heterogeneous multi-agent systems where exhaustive data collection is impractical. Additionally, improvements in decentralized settings demonstrate effective handling of partial observability and limited global information. Notably, surpassing MAD, the closest decentralized baseline, underscores the advantage of communication over teammate modeling, advancing state-of-the-art decentralized diffusion methods. These results highlight the potential of integrating heterogeneous graph-based communication within diffusion models to enhance coordination and generalization in diverse multi-agent systems.

ACKNOWLEDGMENTS

This work was completed under the Laboratory Directed Research and Development program at Sandia National Laboratories and the Naval Research Lab under grant N00173-21-1-G009.

REFERENCES

[1] Anurag Ajay, Yilun Du, Abhi Gupta, Joshua Tenenbaum, Tommi Jaakkola, and Pulkit Agrawal. 2022. Is conditional generative modeling all you need for decision-making? *arXiv preprint arXiv:2211.15657* (2022).

[2] Jacob Beck, Risto Vuorio, Evan Zheran Liu, Zheng Xiong, Luisa Zintgraf, Chelsea Finn, and Shimon Whiteson. 2023. A survey of meta-reinforcement learning. *arXiv preprint arXiv:2301.08028* (2023).

[3] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems* 34 (2021), 15084–15097.

[4] Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Mike Rabbat, and Joelle Pineau. 2019. Tarmac: Targeted multi-agent communication. In *International Conference on Machine Learning*. PMLR, 1538–1546.

[5] Benjamin Ellis, Jonathan Cook, Skander Moalla, Mikayel Samvelyan, Mingfei Sun, Anuj Mahajan, Jakob Foerster, and Shimon Whiteson. 2024. Smacv2: An improved benchmark for cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2024).

[6] Claude Formanek, Asad Jeewa, Jonathan Shock, and Arnu Pretorius. 2023. Off-the-Grid MARL: Datasets with Baselines for Offline Multi-Agent Reinforcement Learning. *arXiv preprint arXiv:2302.00521* (2023).

[7] Tyler Gunn and John Anderson. 2015. Dynamic heterogeneous team formation for robotic urban search and rescue. *J. Comput. System Sci.* 81, 3 (2015), 553–567.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[9] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

[10] Lu Hong and Scott E Page. 2001. Problem solving by heterogeneous agents. *Journal of economic theory* 97, 1 (2001), 123–163.

[11] Zhang-Wei Hong, Aviral Kumar, Sathwik Karnik, Abhishek Bhandwaldar, Akash Srivastava, Joni Pajarinen, Romain Laroche, Abhishek Gupta, and Pulkit Agrawal. 2023. Beyond uniform sampling: Offline reinforcement learning with imbalanced datasets. *Advances in Neural Information Processing Systems* 36 (2023), 4985–5009.

[12] Christopher D Hsu, Heejin Jeong, George J Pappas, and Pratik Chaudhari. 2021. Scalable reinforcement learning policies for multi-agent control. In *2021 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 4785–4791.

[13] Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. 2022. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991* (2022).

[14] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643* (2020).

[15] Yaru Niu, Rohan R Paleja, and Matthew C Gombolay. 2021. Multi-Agent Graph-Attention Communication and Teaming.. In *AAMAS*. 964–973.

[16] Esmaeil Seraj, Rohan Paleja, Luis Pimentel, Kin Man Lee, Zheyuan Wang, Daniel Martin, Matthew Sklar, John Zhang, Zahi Kakish, and Matthew Gombolay. 2024. Heterogeneous policy networks for composite robot team communication and coordination. *IEEE Transactions on Robotics* (2024).

[17] Tianyu Shi, Dong Chen, Kaian Chen, and Zhaojian Li. 2021. Offline reinforcement learning for autonomous driving with safety and exploration enhancement. *arXiv preprint arXiv:2110.07067* (2021).

[18] Amanpreet Singh, Tushar Jain, and Sainbayar Sukhbaatar. 2018. Learning when to communicate at scale in multiagent cooperative and competitive tasks. *arXiv preprint arXiv:1812.09755* (2018).

[19] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*. PMLR, 2256–2265.

[20] Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* 32 (2019).

[21] Sainbayar Sukhbaatar, Rob Fergus, et al. 2016. Learning multiagent communication with backpropagation. *Advances in neural information processing systems* 29 (2016).

[22] Yiqin Yang, Xiaoteng Ma, Chenghao Li, Zewu Zheng, Qiyuan Zhang, Gao Huang, Jun Yang, and Qianchuan Zhao. 2021. Believe What You See: Implicit Constraint Approach for Offline Multi-Agent Reinforcement Learning. arXiv:2106.03400 [cs.AI] https://arxiv.org/abs/2106.03400

[23] Sean Ye, Manisha Natarajan, Zixuan Wu, Rohan Paleja, Letian Chen, and Matthew C. Gombolay. 2023. Learning Models of Adversarial Agent Behavior Under Partial Observability. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 3688–3695. https://doi.org/10.1109/IROS55552.2023.10341378

[24] Zhengbang Zhu, Minghuan Liu, Liyuan Mao, Bingyi Kang, Minkai Xu, Yong Yu, Stefano Ermon, and Weinan Zhang. 2023. Madiff: Offline multi-agent learning with diffusion models. *arXiv preprint arXiv:2305.17330* (2023).