cMALC-D: Contextual Multi-Agent LLM-Guided Curriculum Learning with Diversity-Based Context Blending

Anirudh Satheesh
anirudhs@terpmail.umd.edu
University of Maryland
College Park, Maryland, USA

Keenan Powell
kpowell1@terpmail.umd.edu
University of Maryland
College Park, Maryland, USA

Hua Wei
hua.wei@asu.edu
Arizona State University
Tempe, Arizona, USA

Abstract

Many multi-agent reinforcement learning (MARL) algorithms are trained in fixed simulation environments, making them brittle when deployed in real-world scenarios with more complex and uncertain conditions. Contextual MARL (cMARL) addresses this by parameterizing environments with context variables and training a context-agnostic policy that performs well across all environment configurations. Existing cMARL methods attempt to use curriculum learning to help train and evaluate context-agnostic policies, but they often rely on unreliable proxy signals, such as value estimates or generalized advantage estimates that are noisy and unstable in multi-agent settings due to inter-agent dynamics and partial observability. To address these issues, we propose Contextual Multi-Agent LLM-Guided Curriculum Learning with Diversity-Based Context Blending (cMALC-D), a framework that uses Large Language Models (LLMs) to generate semantically meaningful curricula and provide a more robust evaluation signal. To prevent mode collapse and encourage exploration, we introduce a novel diversity-based context blending mechanism that creates new training scenarios by combining features from prior contexts. Experiments in traffic signal control domains demonstrate that cMALC-D improves both generalization and sample efficiency compared to existing curriculum learning baselines.

CCS Concepts

• **Computing methodologies → Multi-agent reinforcement learning**; *Partially-observable Markov decision processes*; *Cooperation and coordination*; *Multi-agent systems*; • **Applied computing** → *Transportation.*

Keywords

multi agent, contextual reinforcement learning, generalization, Large Language Models

ACM Reference Format:
Anirudh Satheesh, Keenan Powell, and Hua Wei. 2025. cMALC-D: Contextual Multi-Agent LLM-Guided Curriculum Learning with Diversity-Based Context Blending. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3746252.3760863

CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3760863

1 Introduction

Multi-Agent Reinforcement Learning (MARL) has shown promising results across diverse applications, including real-time strategy games [15, 22], supply chain management [17, 19], navigation and pathfinding [24, 35], and traffic signal control [3, 13, 23]. These successes are largely attributed to the ability of MARL algorithms, such as Independent Proximal Policy Optimization (IPPO) [4] and Multi-Agent Proximal Policy Optimization (MAPPO) [32], to train agents capable of coordination and cooperation.

Despite this progress, generalization remains a key challenge. Most MARL algorithms are trained in simulation environments with fixed or limited variability, making them brittle when deployed in real-world scenarios where conditions are more complex and uncertain. External factors such as noise [2, 10] and dynamic changes [34] can degrade MARL performance substantially. These issues are amplified in multi-agent settings due to the combinatorial explosion of agent interactions, which can destabilize learned policies and exacerbate overfitting to training conditions.

To address the challenge of poor generalization to unseen or out-of-distribution environments, we build on the contextual MARL (cMARL) framework [11], which explicitly represents environment variability through a context variable c [9]. Generalization in cMARL is commonly improved via curriculum learning [1], where agents are trained on contexts that gradually increase in difficulty or novelty [7, 12, 14, 20, 27]. This allows agents to incrementally acquire transferable skills and improves robustness at test time.

While curriculum learning improves generalization in contextual MARL by ordering training environments by difficulty or novelty, existing approaches often depend on hand-crafted heuristics or static curriculum schedules. These strategies may struggle to adapt to the evolving agents or the complex dependencies among context variables in dynamic environments. To address these limitations, we explore the use of Large Language Models (LLMs) as high-level curriculum designers. Recent advancements in LLMs, such as GPT-4o, Qwen, and Gemini, have shown strong capabilities in reasoning, planning, and abstraction [6, 18, 25, 29, 31]. These models can operate in diverse domains through in-context learning, suggesting their potential for adaptively generating and sequencing environment contexts based on the agent's current state and performance.

In this work, we propose Contextual Multi-Agent LLM-Guided Curriculum Learning with Diversity-Based Context Blending (cMALC-D), a novel framework that integrates LLMs into contextual MARL to dynamically generate training curricula. Specifically, in cMALC-D, the LLM acts as a high-level controller that observes the agent's learning progress and adaptively proposes new environment contexts by reasoning over the space of context variables. To enhance context coverage and prevent overfitting to narrow distributions, we introduce a diversity-based blending mechanism that mixes previously sampled contexts to construct novel yet meaningful training conditions. This LLM-guided process allows the curriculum to evolve in tandem with agent capabilities, providing more targeted and generalizable training experiences. We evaluate cMALC-D in multiple traffic control scenarios, where environments are naturally high-dimensional and dynamic. Results show that our approach improves generalization to unseen environments with higher sample efficiency compared to existing self-paced or handcrafted curriculum strategies.

Our contributions are outlined as follows:

- We introduce Contextual Multi-Agent LLM-Guided Curriculum Learning with Diversity-Based Context Blending (cMALC-D), a framework that leverages Large Language Models (LLMs) to generate semantically meaningful context-based curricula for training MARL agents, improving generalization to unseen environment configurations.
- Experiments in multiple traffic-based environments demonstrate that our approach achieves better generalization compared to other self-paced curricula, with higher sample efficiency.

2 Methodology

2.1 Problem Definition

We formulate cMALC-D as a *contextual decentralized partially observable Markov decision process* (cDec-POMDP). A cDec-POMDP is parameterized by the tuple $M_c = (\mathcal{N}, \mathcal{S}, \mathcal{A}, \mathcal{T}_c, R, \Omega, O, \gamma, \mu)$, where $\mathcal{N}$ is the set of agent indices, denoting a system of $n = |\mathcal{N}|$ cooperative agents. $\mathcal{S}$ is the joint state space shared across all agents, and $\mathcal{A} = \prod_{i\in\mathcal{N}} \mathcal{A}^i$ is the joint action space, where $\mathcal{A}^i$ is the action space for agent i. The transition function $\mathcal{T}_c : \mathcal{S}\times\mathcal{A} \longrightarrow \Delta(\mathcal{S})$ determines the next state distribution given the current state and joint action under context c, while the reward function $R : \mathcal{S}\times\mathcal{A} \longrightarrow \mathbb{R}$ maps state-action pairs to a scalar reward. $\Omega = \prod_{i\in\mathcal{N}} \Omega^i$ denotes the joint observation function, where $\Omega^i : \mathcal{S} \longrightarrow O^i$ provides a private observation to agent i, and $O = \prod_{i\in\mathcal{N}} O^i$ is the joint observation space. The discount factor $\gamma \in [0, 1)$ specifies the importance of future rewards, and $\mu : \mathcal{S} \longrightarrow [0, 1]$ is the initial state distribution.

To model task variation, we define a distribution over contexts $c \in C$, where each context specifies a different context instance by altering the transition functions. This induces a set $\mathcal{M}_C = \{M_c | c \in C\}$ of decentralized POMDPs, each corresponding to a distinct environment. Each M_c encodes a different context instantiation with a different transition function, and we assume that the reward function and the state, action, and observation spaces remain fixed across all contexts.

The objective of a policy π in a cDec-POMDP is to maximize the expected return over the context distribution and over finite horizon H:

$$\pi^* = \arg\max_{\pi\in\Pi} \mathbb{E}_{c\sim C}\left[\mathbb{E}_{\pi}\left[\sum_{t=0}^{H-1} \gamma^t \mathcal{R}(s_t, a_t)\middle|\pi, c\right]\right]$$

where $\Pi = \{\pi = (\pi^1, \ldots, \pi^n) \mid \pi^i : o^i_t \longrightarrow a^i_t\}$ is the set of decentralized policies, with each agent $i \in \mathcal{N}$ selecting actions based only on its local observation o^i_t. The goal is to learn a context-agnostic policy that generalizes well across the set of Dec-POMDPs $\mathcal{M}_C$.

2.2 Related Works on Curriculum Learning

Curriculum learning [1] aims to improve generalization by structuring agents' learning process, progressing from easier to more challenging tasks. While some methods employ manual curricula designed by human experts, others explore automated curriculum generation to eliminate the need for manual design. For example, Sukhbaatar et al. [26] leverages self-play to minimize the number of training episodes by generating progressively harder tasks through agent interactions. Dendorfer et al. [5] and Florensa et al. [8] use Generative Adversarial Networks (GANs) to create challenging goals tailored to the agent's capabilities. Portelas et al. [21] uses a Gaussian Mixture Model (GMM) to model the task space and align a student's learning trajectory with a teacher-generated curriculum. However, each of these requires an auxiliary model to determine the learnability of a task. Instead, Eimer et al. [7], Klink et al. [14], Parker-Holder et al. [20] employ self-paced learning, which orders the curriculum based on agents' performance. Thus, each task is tailored to each agent's abilities and ensures that the learning progress is more self-contained.

2.3 LLM-based Self-Paced Curriculum

The two main limitations of current self-paced curriculum learning algorithms for contextual MARL are random task sampling and unreliable proxy evaluations. Most existing approaches generate new contexts by randomly sampling from the context space, without considering any meaningful progression between sampled environments. This can lead to large variations between contexts across training episodes, making learning unstable and inefficient. Additionally, current methods typically rely on policy metrics like the value estimate or the Generalized Advantage Estimate (GAE) [12, 20] to evaluate agent performance and determine subsequent contexts to train on. These can be unreliable during early training, under sudden domain transfer, and noisy in multi-agent settings.

To address these limitations, we propose cMALC-D, a novel curriculum learning strategy for contextual MARL that combines structured reasoning capabilities of large language models (LLMs) with an exploration mechanism based on task arithmetic. This approach improves both the generation of semantically meaningful environment contexts and the robustness of policy evaluation under limited feedback. We present the full algorithm in Algorithm 1.

LLM-Guided Context Generation Instead of randomly sampling from a context space C, we leverage an LLM to reason over a sliding window of past training results and generate new contexts that reflect a meaningful progression in difficulty or diversity. At each curriculum step, the LLM receives a window of most recent contexts $\{c_{t-w}, \cdots, c_t\}$ and their associated performance metrics from each agent $\{m_{t-w}, \cdots m_t\}$ when trained on the MARL algorithm A. It then leverages this history to propose a new context that either incrementally challenges the current multi-agent policy or targets known weaknesses observed in recent episodes.

Diversity-Based Context Blending To avoid curriculum stagnation and encourage exploration of the context space, we monitor the similarity between successive contexts. If the LLM repeatedly generates highly similar contexts, indicating potential mode collapse in curriculum progression, we enable a diversity mechanism.

Specifically, when the number of consecutive similar contexts exceeds a threshold, we blend the current LLM-proposed context with a randomly sampled context from the history. This interpolation helps inject novelty into the curriculum while avoiding sudden changes in curricula. We measure similarity between contexts by treating each context as a vector and measuring the normalized cosine similarity between them.

Alternating Policy Training and Context Generation Similar to [18], we alternate between policy training and context generation. After each training phase, the agent's performance on the current context is recorded and passed to the LLM, which conditions on a sliding window of past evaluations to generate the next context. This approach, in-context context generation, enables the LLM to implicitly reason about task difficulty and progression without gradient updates or handcrafted reward shaping.

Algorithm 1 Contextual Multi-Agent LLM-Guided Curriculum Learning with Diversity Based Context Blending (cMALC-D)

```
Require: MARL algorithm $A$, context space $C$, LLM $M$, blending factor $\alpha$, sliding window size $w$, similarity threshold $\delta$, max similar count $k$, initial context $c_0$
Initialize context buffer $H \leftarrow []$, similarity counter $s \leftarrow 0$
Set current context $c_0$
for curriculum step $t = 0, 1, \dots, T$ do
    Train $\pi_t$ on $c_t$ via $A$, collect performance metric $m_t$
    Append $(c_t, m_t)$ by algorithm $A$ to context buffer $H$
    Construct window $H_w = \{(c_{t-w}, m_{t-w}), \dots, (c_t, m_t)\}$
    Query $M$ with $H_w$ to generate new context $c^M_{t+1}$
    Compute similarity $\sigma \leftarrow \mathrm{Sim}(\{c_{t-w}, \cdots, c_t\}, c^M_{t+1})$
    if $\sigma \geq \delta$ then
        Increment similarity counter $s \leftarrow s + 1$
    else
        Reset similarity counter $s \leftarrow 0$
    if $s \geq k$ then
        Blend: $c_{t+1} \leftarrow \alpha c_r + (1 - \alpha) c^M_{t+1}, c_r \sim \mathrm{Unif}(H_w)$
        Reset similarity counter $s \leftarrow 0$
    else
        Set $c_{t+1} \leftarrow c^M_{t+1}$
```

3 Experiments

In this section, we conduct experiments to answer two main questions: What is the generalization performance of the algorithm? How does the diversity mechanism influence context generation?

3.1 Experimental Setup

We evaluate cMALC-D on three autonomous traffic signal control environments based on real-world data, where the context parameters of the environment are defined in Table 1. We run our experiments with the CityFlow environment [33] over 5 different random seeds. We train all policies with MAPPO [32], but any MARL algorithm will work; we choose to use MAPPO due to its efficiency compared to off-policy algorithms. The LLM used is Qwen2.5-7B-Instruct with activation-aware weight quantization [16] to reduce memory usage.

Table 1: Context parameters used for curriculum learning and their specified ranges.

Parameter	Description	Range
`length`	Length of each car	1–10 m
`width`	Width of each car	1–5 m
`maxPosAcc`	Max acceleration when speeding up	0.5–5 m/s^2
`maxNegAcc`	Max deceleration when braking	0.5–5 m/s^2
`usualPosAcc`	Default acceleration when speeding up	1–5 m/s^2
`usualNegAcc`	Default deceleration when braking	1–5 m/s^2
`minGap`	Minimum gap between cars	1–10 m
`maxSpeed`	Maximum speed a car can travel	3–15 m/s
`headwayTime`	Time to reach the vehicle in front	1–5 s (int)

For all experiments, we alternate between expanding the curriculum and training the MARL policy for 500 episodes, where each episode is 360 timesteps, resulting in 180,000 trajectories per training phase. We reserve a held-out test set of 5 contexts and evaluate the current policy every 5 episodes using greedy action sampling. After training, we generate 10 additional random contexts to assess generalization performance after a brief finetuning phase of 5 episodes. In addition to the test reward (which is the total time vehicles are moving), we report the throughput and the average total time. We evaluate cMALC-D against 5 baselines: No Curriculum (using the initial context), Domain Randomization [28], PLR [12], ACCEL [20], and SPACE [7].

3.2 Generalization Performance

We show the generalization performance of cMALC-D against the baseline algorithms in Table 2. Across all three environments, JN 1×3, HZ, and JN 3×4, cMALC-D consistently outperforms or matches all other curriculum strategies on the test reward and specific traffic policy metrics, such as average delay and throughput. For example, in JN 1×3, it achieves the highest test reward (29.01 ± 0.35) and throughput (3073.22 ± 114.06) while reducing wait time by 2% over the second-best algorithm. Similar trends hold for the HZ and JN 3×4 environments.

Structured curricula are necessary to learn generalizable policies. In contrast, Domain Randomization underperforms compared to cMALC-D, often giving 3rd or 4th place results across performance metrics (e.g., 4th place in average delay in HZ with 241.79 ± 35.60 vs. cMALC-D's 146.96 ± 19.93). While it occasionally yields high throughput or test rewards over other algorithms (e.g., 2nd place test reward of 27.55 ± 0.41 in JN 1×3), these gains are unreliable and highly environment-dependent. This inconsistency highlights a fundamental limitation of randomization-based strategies: while they expose agents to a wide range of environments, they do so without considering progression or context relevance. As a result, agents may struggle to learn the high-level coordination skills necessary for generalization due to rapid context switching in the curriculum.

Original context can be a useful prior, but may encourage overfitting. Training without a curriculum can yield strong performance, particularly in the JN environments, where No Curriculum frequently ranks second after cMALC-D (e.g., throughput of 3704.17 ± 147.39 vs. cMALC-D's 3795.46 ± 159.50 in JN 3×4). This suggests that the original context provides a good prior, enabling agents to learn basic coordination strategies. However, its effectiveness diminishes in more diverse settings (most notably in the HZ environment, where its average delay of 339.09 ± 53.28 is worse than cMALC-D's 146.96 ± 19.93), where it performs significantly worse than cMALC-D and exhibits high variance even in the JN environments (e.g., test reward standard deviation of 2.44 vs. 1.53 in JN 3 × 4). This drop shows that without curriculum learning, agents overfit to the original context features, which limits generalizability.

LLM-based context evaluation provides a robust signal for effective curriculum learning. While some methods like ACCEL, PLR, and SPACE incorporate similar automatic curriculum schemes, they rely heavily on policy evaluation signals, such as value functions or generalized advantage estimates, to select and schedule tasks. While these signals can be highly effective in single-agent domains with millions of environment updates, they can be noisy or unreliable in MARL due to non-stationarity, partial observability, and inter-agent dependencies (e.g., SPACE's inconsistent rankings from 5th place in JN 1 × 3 to 1st place in HZ). On the other hand, cMALC-D's context selection strategy promotes gradual skill acquisition that transfers well across diverse contexts. This is due to using language-based evaluations that can capture qualitative improvements that traditional metrics might overlook (demonstrated by cMALC-D's top performance across all environments with test rewards of 29.01 ± 0.35, 172.87 ± 1.03, and 116.57 ± 1.53 in JN 1 × 3, HZ, and JN 3 × 4, respectively).

Table 2: Performance metrics across all environments. Best results per metric are shown in bold and second-best results are underlined. We include uncertainty within one standard deviation of the mean, averaged over 5 seeds.

Curriculum	Average Time	Throughput	Test Reward
	JN 1 × 3		
No Curriculum	816.00 ± 35.63	3032.72 ± 114.20	27.56 ± 0.46
Domain Randomization	865.64 ± 43.08	2807.22 ± 141.63	27.55 ± 0.41
PLR	841.74 ± 40.63	2955.79 ± 127.53	27.61 ± 0.44
ACCEL	860.81 ± 39.57	2813.86 ± 130.81	26.55 ± 0.45
SPACE	939.39 ± 39.96	2544.24 ± 133.01	26.89 ± 0.43
cMALC-D	**809.39 ± 36.37**	**3073.22 ± 114.06**	**29.01 ± 0.35**
	HZ 4 × 4		
No Curriculum	710.01 ± 42.78	2188.94 ± 74.78	164.04 ± 2.68
Domain Randomization	637.52 ± 31.61	2318.41 ± 53.49	168.53 ± 1.44
PLR	611.26 ± 26.48	2390.56 ± 38.11	171.85 ± 1.02
ACCEL	615.02 ± 25.46	2393.30 ± 35.37	171.07 ± 1.17
SPACE	588.84 ± 25.11	**2440.18 ± 32.02**	**172.90 ± 1.05**
cMALC-D	**586.86 ± 25.50**	2440.09 ± 35.45	172.87 ± 1.03
	JN 3 × 4		
No Curriculum	829.92 ± 36.98	3704.17 ± 147.39	115.77 ± 2.44
Domain Randomization	976.73 ± 37.67	3108.67 ± 151.36	112.38 ± 1.52
PLR	992.79 ± 47.30	3071.38 ± 188.12	111.98 ± 1.74
ACCEL	1077.27 ± 42.13	2699.78 ± 166.09	110.28 ± 1.60
SPACE	899.94 ± 43.34	3447.63 ± 172.27	114.47 ± 1.64
cMALC-D	**815.81 ± 39.51**	**3795.46 ± 159.50**	**116.57 ± 1.53**

3.3 Influence of the Diversity Mechanism

To evaluate the impact of the diversity mechanism, we compare three variants of our method: the full version with similarity-based diversity (**cMALC-D**), a baseline without the diversity mechanism (**cMALC**), and a variant that applies task arithmetic with random probability $\epsilon = 0.1$ instead of using similarity checks (**cMALC-ϵ**).

Figure 1 shows the mean test reward across the three datasets, each averaged over 5 seeds. On all three datasets, cMALC-D outperforms the baseline cMALC and the random-diversity variant cMALC-ϵ. On the simpler Jinan 1 × 3 dataset, cMALC-D achieves clear gains, maintaining a reward of nearly 30.5, about two points higher on average than the other variants. On the medium-complexity Hangzhou dataset, cMALC-D exhibits faster convergence and improved stability, while cMALC-ϵ abruptly drops in performance after 170,000 timesteps. Finally, while performance is generally lower on the JN 3 × 4 dataset, cMALC-D remains the top-performing variant. Notably, cMALC test reward declines significantly, which suggests mode collapse and poor generalization to test contexts, further highlighting the need for context diversity during training.

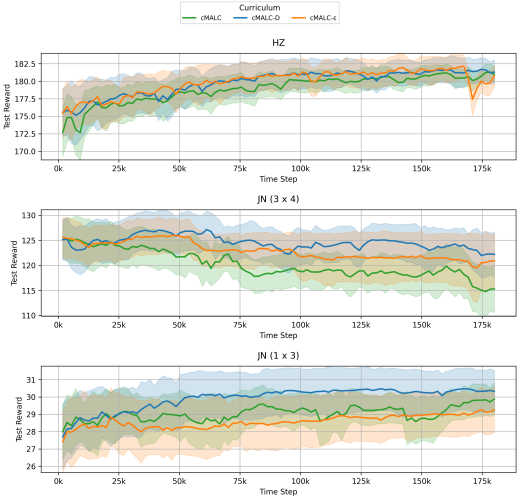

Figure 1: Mean Test Reward over the traffic datasets.

4 Conclusion and Future Work

In this paper, we develop cMALC-D, an LLM-based curriculum learning algorithm for contextual MARL. Our method leverages the reasoning capabilities of LLMs to generate semantically meaningful curricula. We also introduce a novel diversity-based mechanism based on task arithmetic from continual learning to encourage exploration in the context space and avoid mode collapse. Our experiments on three real-world traffic environments show that cMALC-D enhances MARL policy generalization and sample efficiency over a variety of environment configurations. Future work includes extending our formulation to handle noisy or malfunctioning environments [30], and exploring semantic feature relationships to enhance self-paced curricula.

Acknowledgments

The work was partially supported by NSF awards #2421839, #2442477, NAIRR #240120 and used AWS through the CloudBank project, which is supported by NSF grant #1925001. The views and conclusions in this paper are those of the authors and should not be interpreted as representing any funding agencies. We thank OpenAI for providing API credits under the Researcher Access program and Amazon Research Awards.

GenAI Disclosure

Generative AI tools were only used for light editing of the manuscript.

References

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning* (Montreal, Quebec, Canada) *(ICML '09)*. Association for Computing Machinery, New York, NY, USA, 41–48. doi:10.1145/1553374.1553380

[2] Alexander Bukharin, Yan Li, Yue Yu, Qingru Zhang, Zhehui Chen, Simiao Zuo, Chao Zhang, Songan Zhang, and Tuo Zhao. 2023. Robust multi-agent reinforcement learning via adversarial regularization: Theoretical foundation and stable algorithms. *Advances in Neural Information Processing Systems* 36 (2023), 68121–68133.

[3] Tianshu Chu, Jie Wang, Lara Codecà, and Zhaojian Li. 2019. Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE transactions on intelligent transportation systems* 21, 3 (2019), 1086–1095.

[4] Christian Schroeder De Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip HS Torr, Mingfei Sun, and Shimon Whiteson. 2020. Is independent learning all you need in the starcraft multi-agent challenge? *arXiv preprint arXiv:2011.09533* (2020).

[5] Patrick Dendorfer, Aljosa Osep, and Laura Leal-Taixé. 2020. Goal-gan: Multimodal trajectory prediction based on goal position estimation. In *Proceedings of the Asian Conference on Computer Vision*.

[6] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A Survey on In-context Learning. arXiv:2301.00234 [cs.CL] https://arxiv.org/abs/2301.00234

[7] Theresa Eimer, André Biedenkapp, Frank Hutter, and Marius Lindauer. 2021. Self-paced context evaluation for contextual reinforcement learning. In *International Conference on Machine Learning*. PMLR, 2948–2958.

[8] Carlos Florensa, David Held, Xinyang Geng, and Pieter Abbeel. 2018. Automatic goal generation for reinforcement learning agents. In *International conference on machine learning*. PMLR, 1515–1528.

[9] Assaf Hallak, Dotan Di Castro, and Shie Mannor. 2015. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259* (2015).

[10] Sihong He, Songyang Han, Sanbao Su, Shuo Han, Shaofeng Zou, and Fei Miao. 2023. Robust multi-agent reinforcement learning with state uncertainty. *arXiv preprint arXiv:2307.16212* (2023).

[11] Vindula Jayawardana, Baptiste Freydt, Ao Qu, Cameron Hickert, Zhongxia Yan, and Cathy Wu. 2024. Intersectionzoo: Eco-driving for benchmarking multi-agent contextual reinforcement learning. *arXiv preprint arXiv:2410.15221* (2024).

[12] Minqi Jiang, Edward Grefenstette, and Tim Rocktäschel. 2021. Prioritized level replay. In *International Conference on Machine Learning*. PMLR, 4940–4950.

[13] Qize Jiang, Minhao Qin, Shengmin Shi, Weiwei Sun, and Baihua Zheng. 2022. Multi-agent reinforcement learning for traffic signal control through universal communication method. *arXiv preprint arXiv:2204.12190* (2022).

[14] Pascal Klink, Carlo D'Eramo, Jan R Peters, and Joni Pajarinen. 2020. Self-paced deep reinforcement learning. *Advances in Neural Information Processing Systems* 33 (2020), 9216–9227.

[15] Karol Kurach, Anton Raichuk, Piotr Stańczyk, Michał Zając, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, and Sylvain Gelly. 2020. Google Research Football: A Novel Reinforcement Learning Environment. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 04 (Apr. 2020), 4501–4510. doi:10.1609/aaai.v34i04.5878

[16] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of Machine Learning and Systems* 6 (2024), 87–100.

[17] Xiaotian Liu, Ming Hu, Yijie Peng, and Yaodong Yang. 2022. Multi-agent deep reinforcement learning for multi-echelon inventory management. *Production and Operations Management* (2022), 10591478241305863.

[18] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Eureka: Human-level reward design via coding large language models. *arXiv preprint arXiv:2310.12931* (2023).

[19] Marwan Mousa, Damien van de Berg, Niki Kotecha, Ehecatl Antonio del Rio Chanona, and Max Mowbray. 2024. An analysis of multi-agent reinforcement learning for decentralized inventory control systems. *Computers & Chemical Engineering* 188 (2024), 108783.

[20] Jack Parker-Holder, Minqi Jiang, Michael Dennis, Mikayel Samvelyan, Jakob Foerster, Edward Grefenstette, and Tim Rocktäschel. 2023. Evolving Curricula with Regret-Based Environment Design. arXiv:2203.01302 [cs.LG] https://arxiv.org/abs/2203.01302

[21] Rémy Portelas, Cédric Colas, Katja Hofmann, and Pierre-Yves Oudeyer. 2020. Teacher algorithms for curriculum learning of deep rl in continuously parameterized environments. In *Conference on Robot Learning*. PMLR, 835–853.

[22] 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).

[23] Anirudh Satheesh and Keenan Powell. 2025. A Constrained Multi-Agent Reinforcement Learning Approach to Autonomous Traffic Signal Control. *arXiv preprint arXiv:2503.23626* (2025).

[24] Alexey Skrynnik, Anton Andreychuk, Maria Nesterova, Konstantin Yakovlev, and Aleksandr Panov. 2024. Learn to follow: Decentralized lifelong multi-agent pathfinding via planning and learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17541–17549.

[25] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M. Sadler, Wei-Lun Chao, and Yu Su. 2023. LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[26] Sainbayar Sukhbaatar, Ilya Kostrikov, Arthur Szlam, and Rob Fergus. 2017. Intrinsic Motivation and Automatic Curricula via Asymmetric Self-Play. (03 2017). doi:10.48550/arXiv.1703.05407

[27] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 23–30.

[28] Eric Tzeng, Coline Devin, Judy Hoffman, Chelsea Finn, Pieter Abbeel, Sergey Levine, Kate Saenko, and Trevor Darrell. 2020. Adapting deep visuomotor representations with weak pairwise constraints. In *Algorithmic Foundations of Robotics XII: Proceedings of the Twelfth Workshop on the Algorithmic Foundations of Robotics*. Springer, 688–703.

[29] Yudong Xu, Wenhao Li, Pashootan Vaezipoor, Scott Sanner, and Elias B Khalil. 2023. Llms and the abstraction and reasoning corpus: Successes, failures, and the importance of object-based representations. *arXiv preprint arXiv:2305.18354* (2023).

[30] Qinchen Yang, Zejun Xie, Hua Wei, Desheng Zhang, and Yu Yang. 2024. MalLight: Influence-Aware Coordinated Traffic Signal Control for Traffic Signal Malfunctions. arXiv:2408.09768 [cs.AI] https://arxiv.org/abs/2408.09768

[31] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems* 35 (2022), 20744–20757.

[32] 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.

[33] Huichu Zhang, Siyuan Feng, Chang Liu, Yaoyao Ding, Yichen Zhu, Zihan Zhou, Weinan Zhang, Yong Yu, Haiming Jin, and Zhenhui Li. 2019. Cityflow: A multi-agent reinforcement learning environment for large scale city traffic scenario. In *The world wide web conference*. 3620–3624.

[34] Kaiqing Zhang, TAO SUN, Yunzhe Tao, Sahika Genc, Sunil Mallya, and Tamer Basar. 2020. Robust Multi-Agent Reinforcement Learning with Model Uncertainty. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 10571–10583. https://proceedings.neurips.cc/paper_files/paper/2020/file/774412967f19ea61d448977ad9749078-Paper.pdf

[35] Xiao-Yi Zhang, Yang Liu, Paolo Arcaini, Mingyue Jiang, and Zheng Zheng. 2024. Met-mapf: A metamorphic testing approach for multi-agent path finding algorithms. *ACM Transactions on Software Engineering and Methodology* 33, 8 (2024), 1–37.