CHAIN: Continual Heterogeneous Cooperation with Information Bottleneck for Multi-Agent Reinforcement Learning

Haowen Dou **Lujuan Dang**⋆

Badong Chen

National Key Laboratory of Human-Machine Hybrid Augmented Intelligence
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
douhaowen@stu.xjtu.edu.cn, danglj@xjtu.edu.cn, chenbd@mail.xjtu.edu.cn

Abstract

Efficient heterogeneous Multi-Agent Reinforcement Learning (MARL) in continuously changing environments is key to advancing MARL from simulation to the real world. However, it remains unclear whether current MARL methods can adapt under such dynamic scenarios. In this work, we show that existing methods face significant performance degradation in a novel continual heterogeneous MARL environment we construct, named slippery Multi-Agent Mujoco. We further demonstrate the connection between this phenomenon and *representation redundancy* as well as *insufficient policy heterogenization*. To address this issue, we propose **C**ontinual **H**eterogeneous Cooper**A**tion with **I**nformation Bottle**N**eck (CHAIN), equipped with Heterogeneous Policy Bottleneck (HPB). In HPB, we extend the traditional information bottleneck into three parts, fitting, compression and heterogenization. Our HPB encourages agents to learn an efficient latent representation to adapt to changed environments, as well as disentangles specializations of agents in state perception, thereby encouraging heterogenization. We evaluate CHAIN on standard MA-Mujoco, our slippery MA-Mujoco and a multi-task continual MARL benchmark, MEAL, with various challenging tasks. The experimental results indicate that our CHAIN outperforms state-of-the-art heterogeneous MARL methods on multiple continual learning metrics.

1 Introduction

Multi-Agent Reinforcement Learning (MARL) with heterogeneous agents extends the standard MARL paradigm and naturally emerges in real-world applications, including multi-robot learning [Bettini et al., 2023, Hou et al., 2026], autonomous driving [Zhou et al., 2021], and large-scale traffic control [Ma et al., 2024]. In these tasks, agents have predefined individual specializations and are required to behave differently under the same observations [Monon et al., 2025]. However, traditional MARL methods are primarily designed for homogeneous agents, typically relying on parameter sharing to ensure monotonic policy improvement. To overcome this limitation, heterogeneous MARL approaches have been proposed, aiming to heterogenize agent behaviors in complex scenarios through sequential policy updates [Li et al., 2024, Zhong et al., 2024]. The key idea behind

⋆ Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

these algorithms is to encourage agents to learn from the preceding partners in a one-by-one updating paradigm. To effectively support sequential updating while preserving the desirable property of monotonic policy improvement, various sequential MARL algorithms have been proposed [Kuba et al., 2022, Wang et al., 2023]. Nevertheless, the heterogenization strategies of existing MARL methods lack active incentive, which may hinder a thorough exploration of the policy space [Dou et al., 2024].

The heterogeneous setting presents a significant step toward enabling MARL methods to transition from simulated environments to real-world applications. However, for real-world heterogeneous scenarios, such as multi-joint collaboration for humanoid walking, the unpredictable change of environment dynamics may occur during training and can not be ignored [Haarnoja et al., 2024, Tang et al., 2025a]. Recent research pointed out that even in simulated environments like Mujoco, with a periodic friction change, state-of-the-art single-agent RL methods show a noticeable performance decline [Dohare et al., 2024]. This performance drop is partly due to the gradual loss of representational capacity in neural networks of RL policies during training, which is caused by changing environment-specific information [Moalla et al., 2024, Sokar et al., 2023]. However, how existing heterogeneous MARL methods perform in such unstable environments remains underexplored.

In this work, we focus on MARL in heterogeneous and frequently changing environments. To this end, we first construct a new slippery Multi-Agent Mujoco environment where the friction between the ground and agents changes periodically. Using this new benchmark, we empirically demonstrate that, existing heterogeneous MARL methods struggle to relearn reliable policies after each friction change and, in some scenarios, suffer complete performance collapse. We further demonstrate that the reason behind this phenomenon is high redundancy of the representations of latent layers in policy networks, and insufficient heterogenization. Motivated by this observation, we propose **C**ontinual **H**eterogeneous Cooper**A**tion with **I**nformation Bottle**N**eck (CHAIN). Our key insight is that, ***trained in environments with changing dynamics, the representation redundancy of existing MARL methods reflects an overemphasis on environment-specific information***, which is harmful to continual learning. Moreover, ***insufficient heterogenization leads to homogenization redundancy, causing widespread performance degradation of homogenized agents when the environment changes.*** To address this issue, we propose an information bottleneck (IB) based method, Heterogeneous Policy Bottleneck (HPB) to compress both the representation redundancy and the homogenization redundancy. In this way, our HPB encourages agents to focus on the invariant and task-irrelevant information before and after the environment change. We theoretically prove that optimizing the proposed HPB steer agents away from the changing environment-specific information and toward environment-invariant information. We also prove that the novel heterogenization term can disentangle multi-agent perception of the environment, and thus promotes policy diversity. The main contributions are summarized as follows:

1. We construct a new environment for benchmarking MARL algorithms in continual learning, named slippery MA-Mujoco. We empirically show that, the baseline method fails to continually relearn reliable policies after each environment change. We further demonstrate that this limitation arises from high representation redundancy and insufficient heterogenization.
2. We develop a novel MARL model, named CHAIN, where the Heterogeneous Policy Bottleneck (HPB) is proposed. Our HPB is an extension of the traditional IB, aiming to learn invariant information of environments for continual heterogeneous scenarios. To our best knowledge, this is the first continual heterogeneous MARL method.
3. We conduct extensive experiments to evaluate our method. The results indicate that CHAIN outperforms state-of-the-art MARL methods in slippery MA-MuJoCo environments, and outperforms other continual methods in MEAL environments.

2 Related Works

2.1 Heterogeneous MARL

Heterogeneous MARL generalizes standard MARL by relaxing the assumption of homogeneous policies with parameter sharing. The foundational concept of heterogeneous MARL stems from the single-agent rollout iteration reinforcement learning framework proposed by Bertsekas [2021], which introduced the idea of sequential policy updates. Based on this powerful framework, various

algorithms are proposed to offer theoretical guarantees for policy improvement in heterogeneous MARL. For example, from the perspective of joint policy convergence, Kuba et al. [2022] prove the multi-agent advantage decomposition lemma, and present HAPPO. On the other hand, Wang et al. [2023] theoretically demonstrate that the individual policy monotonic improvement is crucial as well, and further present A2PO equipped with a special updating order. Nevertheless, most existing methods lack explicit incentives for heterogenization, limiting their performance in complex tasks Dou et al. [2026].

2.2 Continual Learning in RL

In general, training RL policy networks suffers from sample inefficiency since the data from the non-stationary environment is frequently changing [Kumar et al., 2021, 2025]. Thus, in unstable environments, how to enhance the representation efficiency of policy networks within limited training iterations is the key to continual RL [Zhang et al., 2023, Tang et al., 2025b, Elsayed and Mahmood, 2024, Dohare et al., 2024]. Another continual learning paradigm, termed multi-task continual reinforcement learning, involves tasks that emerge sequentially over time [Caccia et al., 2023, Zhang et al., 2025, Pan et al., 2025, Chung et al., 2024]. This task-switching manner is somewhat similar to environmental dynamics changing, since when agents transition into environments with new dynamics, each such transition can be interpreted as encountering a new task.

In multi-agent settings, continual learning remains relatively underexplored, with only a handful of works addressing it. For example, Yuan et al. [2024] propose MACPro, which learns task contexts and progressively expands or reuses task-specific policy heads for continual multi-agent coordination. Yao et al. [2025] also focus on continual learning on SMAC, and propose to decouple general relational reasoning from task-specific decision-making via conditional hypernetworks and dual regularization. These methods address continual coordination primarily through task contextualization and task-specific policy components.

3 Preliminaries

3.1 Heterogeneous MARL Problem Formulation

In this paper, we consider a cooperative multi-agent system, which can be described as a decentralized Markov decision process (DEC-MDP). A DEC-MDP with n agents can be formulated as the tuple: $\langle \mathcal{S}, \boldsymbol{\mathcal{A}}, r, \mathcal{T}, \delta \rangle$, where $\mathcal{S}$ represents the state space. At time step t, each agent i at state $S_t \in \mathcal{S}$ takes an action a_t^i from its policy $\pi^i(\cdot|S_t)$ to form a joint action $\boldsymbol{a}_t = \{a_t^1, \ldots, a_t^n\}$. The joint policy is given by $\boldsymbol{\pi}(\cdot|S_t) = \pi^1 \times \ldots \times \pi^n$. Agents then receive a joint reward $r(S_t, \boldsymbol{a}_t)$ and step to the $(t+1)^{th}$ state S_{t+1} with the probability $\mathcal{T}(S_{t+1}|S_t, \boldsymbol{a}_t)$. The joint action space of n agents is given by $\boldsymbol{\mathcal{A}} = \mathcal{A}^1 \times \ldots \times \mathcal{A}^n$. $\delta \in [0, 1]$ is the discount factor. The original objective is to learn an optimal joint policy by maximizing the cumulative return: $\bar{\boldsymbol{\pi}} = \arg\max_{\boldsymbol{\pi}} \sum_{t=0}^{\infty} \mathbb{E}_{S_t \sim \rho_{\boldsymbol{\pi}}, \boldsymbol{a}_t \sim \boldsymbol{\pi}} [\delta^t r(S_t, \boldsymbol{a}_t)]$, where $\rho_{\boldsymbol{\pi}}$ is the marginal state distribution. Following HAPPO proposed by Kuba et al. [2022], at episode k, i^{th} agent maximizes the multi-agent surrogate objective as follows,

$$\mathcal{J}_{surro} = \mathbb{E}_{S \sim \rho_{\boldsymbol{\pi}_k}, \boldsymbol{a} \sim \boldsymbol{\pi}_k} [\min(\frac{\pi^i(a^i|S)}{\bar{\pi}_k^i(a^i|S)} A_{joint}^{1:i}(S, \boldsymbol{a}), clip \left(\frac{\pi^i(a^i|S)}{\bar{\pi}_k^i(a^i|S)}, 1 \pm \epsilon \right) A_{joint}^{1:i}(S, \boldsymbol{a}))], \quad (1)$$

$A_{joint}^{1:i}(S, \boldsymbol{a})$ is the joint advantage of the first to i^{th} agents, which is given by $A_{joint}^{1:i} = \frac{\bar{\boldsymbol{\pi}}^{1:i-1}}{\boldsymbol{\pi}^{1:i-1}} \hat{A}(S, \boldsymbol{a})$, where $\boldsymbol{\pi}^{1:i-1} = \prod_{p=1}^{i-1} \pi^{i_p}$ is the joint policy of the first to i^{th} agents.

3.2 Information Bottleneck

The information bottleneck principle is based on the insight that the best representation should preserve sufficient yet minimal information required for prediction Zhai and Zhang [2022]. To this end, Tishby et al. Tishby et al. [2000] propose to maximize the prediction accuracy with the constraint of representation complexity. Formally, given a latent representation Z to be learned, the input data X and the prediction target Y, the information bottleneck optimizes the following Lagrangian objective:

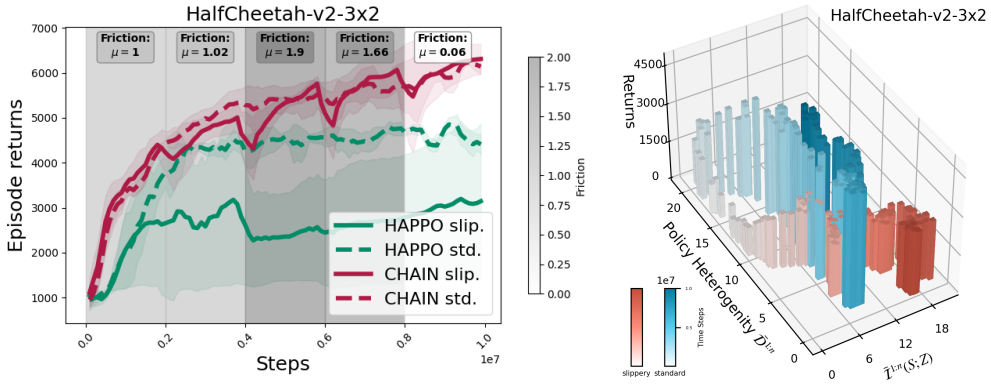
$$\max \mathcal{I}(Z; Y) - \beta \mathcal{I}(Z; X), \quad (2)$$

where $\mathcal{I}(\cdot; \cdot)$ represents the mutual information. The objective of information bottleneck contains maximizing $\mathcal{I}(Z; Y)$ for fitting and minimizing $\mathcal{I}(Z; X)$ for compression.

4 Method

4.1 Problems of Existing Heterogeneous MARL Methods in Continual Settings

Existing heterogeneous MARL methods have demonstrated strong performance in complex tasks. However, how they perform in environments that are more dynamic and unstable remains underexplored. To investigate this, we introduce a new multi-agent Mujoco environment, **slippery MA-Mujoco**, where the friction between agents and the ground changes episodically. In slippery MA-Mujoco, the friction coefficient μ is randomly changed every 2 million time steps, with the new value randomly chosen from the range $[0.01, 2]$. Such variations in friction can significantly affect the learned motor strategies. For instance, consider a fixed policy trained in low-friction settings. When friction increases, the same action may become insufficient for effective locomotion. On the other hand, a policy trained in high-friction conditions may produce unreliable actions when friction decreases, as the agent becomes prone to slipping if it applies the same level of force on a more slippery surface.



(a) (b)

Figure 1: (a) Illustration of the slippery MA-Mujoco. (b) Performance comparison between standard and slippery MA-Mujoco in terms of returns, representation redundancy and policy heterogeneity, where colors of bars are changed from light to dark, representing training from start to end.

Fig. 1a shows the return comparison between training in standard and slippery environments. Note that in Fig. 1a, friction has 4 random changes per run, with the values shared across 5 runs. And in the Sec. 5, the friction μ has different values for each run. Fig. 1a shows that, in such an unstable environment, HAPPO fails to learn continuously and shows a significant improvement slowdown after 4 million steps and two friction changes.

We attribute this phenonmenon to the fact that existing methods in continually changing environments fit environment-specific information the same way as in standard settings. As tasks or dynamics change, these once-useful representations they learned become misaligned with the demand of the current environment, leading to performance degradation. What is truly needed in these environments is the representation of invariant, generalizable information that remains consistent across changes.

We argue, therefore, that frequent environment changing or task switching naturally induces higher representation redundancy for MARL methods. To investigate this, we conduct a comprehensive comparison experiment on between standard MA-Mujoco and our slippery MA-Mujoco. We measure the representation redundancy by computing the mutual information between the input S and the latent representation Z of n policy networks $\pi^{1:n}$, *i.e.* the output of the latent latent layers in the policy network, as follows,

$$\bar{\mathcal{I}}^{1:n}(S;Z) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{I}(S;z^i). \tag{3}$$

This metric measures the information gained from the current environment by the latent representations. And we measure the policy heterogeneity by computing the mean KL divergence of policy

pairs as follows,

$$\bar{\mathcal{D}}^{1:n} = \frac{2}{n(n-1)} \sum_{i<j} \mathrm{KL}(\pi^i; \pi^j). \tag{4}$$

Here, we choose HAPPO as the testing algorithm.

Fig. 1b demonstrates that, compared to the standard scenario, policies learned in the slippery environments have higher $\bar{\mathcal{I}}^{1:n}(S; Z)$, indicating that latent representations obtain more information from the current environment. However, a higher $\bar{\mathcal{I}}^{1:n}(S; Z)$ does not lead to better performance, and instead, it degrades performance. In continual MARL, without compression, the latent representation tends to capture three types of information from the current environment, forming $\bar{\mathcal{I}}^{1:n}(S; Z)$: (1) totally redundant information that provides no benefit to learning, (2) environment-specific information tied to changing dynamics, and (3) environment-invariant information that remain useful across tasks or environments. Unfortunately, the first two types are not only unhelpful but potentially detrimental for continual learning, leading to ***Representation Redundancy***. Environmental instability in slippery settings amplifies the capture of redundant information, which we identify as a primary cause of the performance drop in Fig. 1b.

On the other hand, assuming that the heterogeneity requirements of MA-Mujoco for the agents are consistent across different environmental dynamics to some extent, Fig. 1b shows that policies learned in the slippery environment have lower policy heterogeneity, failing to meet this requirement. We term this effect as ***Homogenization Redundancy***, which increases the risk of systemic failure of homogenized agents when facing environmental changes. These are the reasons why returns in the slippery environment fail to make further progress. Consequently, this reveals a potential optimization direction, *i.e.*, reducing representation redundancy and homogenization redundancy, for addressing the continual heterogeneous MARL challenge.

4.2 Heterogeneous Policy Bottleneck

Motivated by the observation from Fig. 1b, we pursue an MARL method that can enhance adaptability in new heterogeneous environments. Information bottleneck-based methods show efficiency in enhancing generalization for neural networks in supervised learning Yu et al. [2022b]. However, directly using IB for MARL is not the best choice. Solely compressing representations in a frequently changing scenario may encourage agents to learn generic yet conservative policies. Moreover, the homogenization redundancy tends to cause collective performance collapse when the environment changes. Based on the sequential updating scheme, we present Heterogeneous Policy Bottleneck (HPB), an extension of the traditional information bottleneck. In the policy network of agent i, consider the state S, the latent representation z^i and the action a^i, our HPB optimizes the following objective,

$$\max \mathcal{I}(z^i; a^i) - \beta \mathcal{I}(S; z^i) - \gamma \mathcal{I}(z^i; \bar{z}^{i-1}), \tag{5}$$

where β and γ are the coefficients to control the temperature of the two terms.

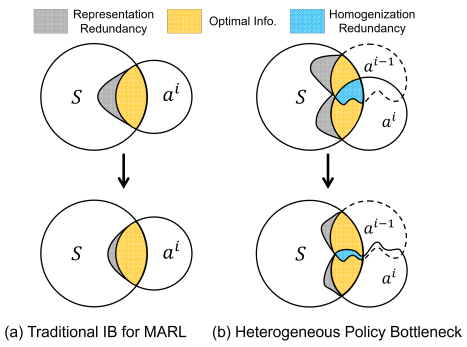


Figure 2: Comparison of traditional IB and our HPB.

As shown in Fig. 2, different from traditional IB, our HPB for MARL contains three parts, fitting, compression and heterogenization, which correspond to the three terms in Eq. 5. The first two terms in Eq. 5 aim to learn an optimal latent layer by limiting learning from the current environment. The third term in Eq. 5 bottlenecks the homogenization redundancy of agents by minimizing the mutual information between the latent representations of the current agent and the previous one. This term can also be seen as a complementary incentive towards heterogenization, which can help agents to escape the local optima induced by the compression term. We discuss the rationality of this pair-wise design of HPB in Appendix 7.8. We then provide some desirable properties of Eq. 5.

Proposition 4.1. *At episode t, suppose the current state $S \sim \mathcal{T}_{current}(\cdot|s_{t-1}, a_{t-1}, \phi)$ is derived from the former state $S_{former} \sim$*

$\mathcal{T}_{former}(\cdot|s_{t-1}, a_{t-1})$ through the environment change ϕ. Define the bias induced by the environment change as $\tilde{S} = S - S_{former}$. Then, the mutual information of $\tilde{S}$ and z_t^i is upper bounded as follows,

$$\mathcal{I}(\tilde{S}; z_t^i) \leq \mathcal{I}(S; z_t^i) - \mathcal{I}(a_{t-1}^i; z_t^i). \tag{6}$$

Proofs can be found in the Appendix 7.1. Prop. 4.1 indicates that, the information that z_t^i obtains from the changed dynamics $\tilde{S}$ is upper bounded by the IB objective. Thus, for continual learning, to steer z_t^i away from the changing environment-specific information and toward environment-invariant information, we can minimize the right of Eq. 6.

Proposition 4.2. *Given state S, a latent representation of current agent to be updated z^i, and a fixed latent representation of one preceding agent $\bar{z}^{i-1}$, then we have,*

$$\mathcal{I}(z^i; \bar{z}^{i-1}) = \underbrace{\mathcal{I}(\bar{z}^{i-1}; S)}_{\textit{Static}} + \underbrace{\mathcal{I}(z^i; S)}_{\substack{\textit{Retain Info.} \\ \textit{Individually}}} \underbrace{- \mathcal{I}(z^i, \bar{z}^{i-1}; S)}_{\substack{\textit{Retain Info.} \\ \textit{Jointly}}} + \underbrace{\mathcal{I}(z^i; \bar{z}^{i-1}|S)}_{\textit{Heterogenization}}. \tag{7}$$

Proofs can be found in the Appendix 7.1. Prop. 4.2 indicates that, minimizing the mutual information between latent representations z^i and $\bar{z}^{i-1}$ is equal to jointly minimizing $\mathcal{I}(z^i; S)$, maximizing $\mathcal{I}(z^i, \bar{z}^{i-1}; S)$, and minimizing $\mathcal{I}(z^i; \bar{z}^{i-1}|S)$. Here, $\mathcal{I}(\bar{z}^{i-1}; S)$ is static when training policy π^i.

In Eq. 7, the second term represents the correlation between the learned representations of two agents and the state. Therefore, minimizing this term encourages z^i to coordinate with a fixed agent to retain information about the environment jointly, enhancing joint perception of the environment. On the other hand, individual perception information of the current agent is compressed by minimizing the $\mathcal{I}(z^i; S)$. Therefore, given a fixed preceding policy $\bar{\pi}^{i-1}$, jointly minimizing $-\mathcal{I}(z^i, \bar{z}^{i-1}; S)$ and $\mathcal{I}(z^i; S)$ encourages the current policy π^i to extract valuable yet sparse information, *i.e.* complementary information to $\bar{\pi}^{i-1}$ from state S. In this way, agents are disentangled to perceive the environment exclusively, thereby promoting heterogeneity. Besides, minimizing the third term $\mathcal{I}(z^i; \bar{z}^{i-1}|S)$ further enhances the discrepancy between agents given state S. Consequently, Prop. 4.2 provides an intuitive explanation of how $\mathcal{I}(z^i; \bar{z}^{i-1})$ can heterogenize agents.

4.3 Continual Heterogeneous Cooperation with Information Bottleneck (CHAIN)

Fig. 3 shows the overall framework of our proposed CHAIN. We combine HPB with the sequential updating paradigm, which provide the access to the information of predecessor agents, enabling learning from them. Compared to traditional heterogeneous MARL methods, our CHAIN encourages policy heterogeneity and low redundancy of latent representations by optimizing HPB.

For CHAIN, directly optimizing Eq. 5 is intractable since the mutual information terms are non-differentiable and hard to compute. Using variational methods Alemi et al. [2017], we derive the variational lower bound of the first terms $\mathcal{I}(z^i; a^i)$ in Eq. 5 as follows,

$$\mathcal{I}(z^i; a^i) \geq \mathbb{E}_{a^i, z^i \sim p}\left[\log q(a^i|z^i; \theta)\right]. \tag{8}$$

where $q(a^i|z^i; \theta)$ is variational approximation to $p(a^i|z^i)$. Maximizing ineq. 8 can be seen as maximizing the log likelihood of the learned policy network, which shares the same effect as Eq. 1. Therefore, we replace $\mathbb{E}_{a^i, z^i \sim p}\left[\log q(a^i|z^i; \theta)\right]$ with the surrogate loss, and define the fitting part of HPB as follows,

$$\mathcal{J}_{fit} \triangleq \mathcal{J}_{surro}. \tag{9}$$

Further, we define and derive the tractable objective of $\mathcal{I}(S; z^i)$ by using the matrix-based α-order Rényi entropy mutual information [Yu et al., 2020] as follows,

$$\mathcal{I}(S; z^i) = \cancel{\mathcal{H}_\alpha(S)} + \mathcal{H}_\alpha(z^i) - \mathcal{H}_\alpha(S, z^i), \tag{10}$$

$$= \frac{1}{1-\alpha} \log_2(\text{Tr}(\mathbf{K}^{z^i})) - \mathcal{H}_\alpha\left(\frac{\mathbf{K}^S \circ \mathbf{K}^{z^i}}{\text{Tr}(\mathbf{K}^S \circ \mathbf{K}^{z^i})}\right), \tag{11}$$

$$=: \mathcal{J}_{compress}. \tag{12}$$

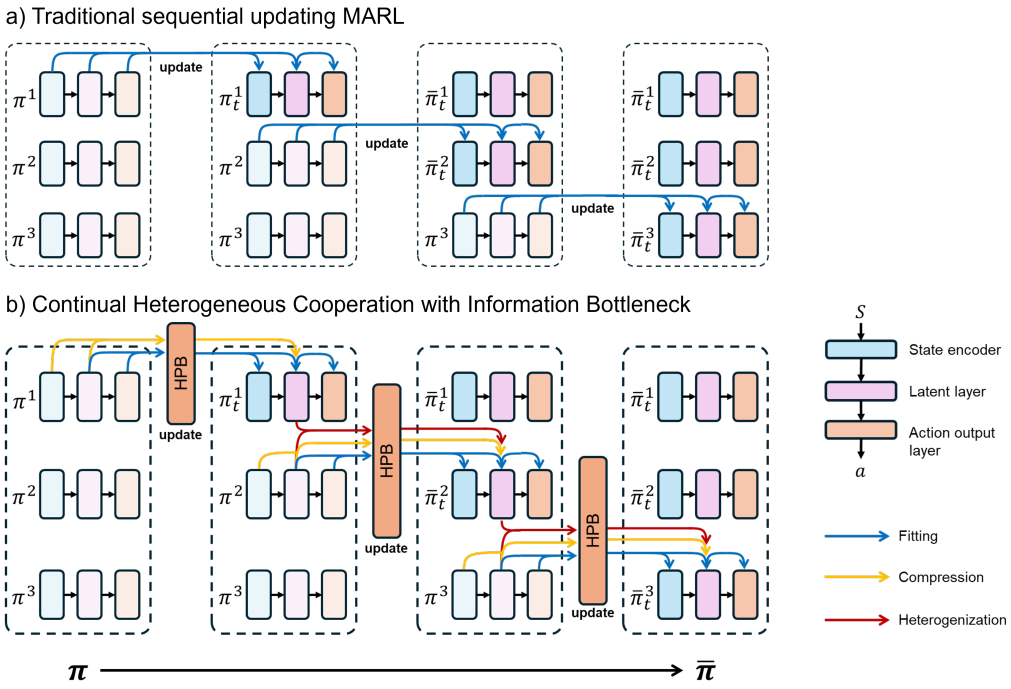


Figure 3: Framework comparison between traditional sequential updating MARL and our CHAIN.

where $\circ$ is the Hadamard product. $\mathcal{H}_\alpha$ is the α-order Rényi entropy, and in this paper, we set $\alpha = 1.001$ to approximate the Shannon's mutual information. $\mathbf{K}^*$ is the Gram matrix of $*$, which can be computed as $\mathbf{K}^*_{ij} = \kappa(*_i, *_j)$ using kernel function κ with finite samples $\{*_i\}_{i=1}^N$. In this paper, we use the Gaussian kernel $\kappa(x, y) = \exp(-\|x - y\|^2/\sigma^2)$, where σ is the kernel bandwidth. Similarly, we obtain the tractable objective of $\mathcal{I}(z^i; \bar{z}^{i-1})$ as follows,

$$\mathcal{I}(z^i; \bar{z}^{i-1}) = \cancel{\mathcal{H}_\alpha(\bar{z}^{i-1})} + \mathcal{H}_\alpha(z^i) - \mathcal{H}_\alpha(z^i, \bar{z}^{i-1}), \tag{13}$$

$$= \frac{1}{1-\alpha}\log_2(\mathrm{Tr}(\mathbf{K}^{z^i})) - \mathcal{H}_\alpha\left(\frac{\mathbf{K}^{\bar{z}^{i-1}} \circ \mathbf{K}^{z^i}}{\mathrm{Tr}(\mathbf{K}^{\bar{z}^{i-1}} \circ \mathbf{K}^{z^i})}\right), \tag{14}$$

$$=: \mathcal{J}_{heterogenize}. \tag{15}$$

Combining Eq. 9, Eq. 12 and Eq. 15, we obtain the total maximization objective of CHAIN as follows,

$$\mathcal{J}_{total} \triangleq \mathcal{J}_{fit} - \beta\mathcal{J}_{compress} - \gamma\mathcal{J}_{heterogenize}. \tag{16}$$

5 Experiments

5.1 Experiment Settings

Benchmarks. To evaluate the performance of CHAIN under changing environment dynamics, we conduct experiments on the proposed **slippery MA-Mujoco** and **MEAL** [Tomilin et al., 2025]. MEAL comprises over 10 handcrafted and procedurally generated Overcooked environments [Carroll et al., 2019b], across three difficulty levels. It involves two heterogeneous agents and emphasizes long-term continual learning capabilities. On MEAL, we compare HPB with continual learning baselines including EWC [Kirkpatrick et al., 2017], MAS [Aljundi et al., 2018], and TRAC [Muppidi et al., 2024], using HAPPO as the backbone algorithm. We also evaluate CHAIN on the standard MA-Mujoco, and the results are presented in the Appendix 7.4.

Baselines. We compare CHAIN with state-of-the-art MARL methods: MAPPO [Yu et al., 2022a], HAPPO [Kuba et al., 2022], A2PO [Wang et al., 2023], MADPO [Dou et al., 2024], and

MACPro Yuan et al. [2024]. We follow the standard MA-Mujoco hyperparameter settings, disabling parameter-sharing for HAPPO, A2PO, MADPO, and CHAIN, while keeping it enabled for MAPPO. All results average over 5 runs with different random seeds. Shaded areas represent 95% confidence intervals.

5.2 Main Results

Training curves in Slippery MA-Mujoco. Fig. 4 compares CHAIN with baselines in the proposed slippery MA-Mujoco environments. After each friction change, evaluated baselines face a serious performance degradation, subsequently struggling to recover their peak returns. In contrast, compared to these methods, CHAIN shows a rapid re-adaptation ability and outperforms baselines in terms of both maximum returns and sample efficiency. Additionally, we observe that after the environmental change, our method still outperforms baselines, indicating that our method successfully learns environment-invariant information by bottlenecking learning from the current environment. Fig. 5 further shows that CHAIN achieves statistically significant improvements over baselines in terms of aggregate IQM performance across 10 tasks of slippery MA-Mujoco.

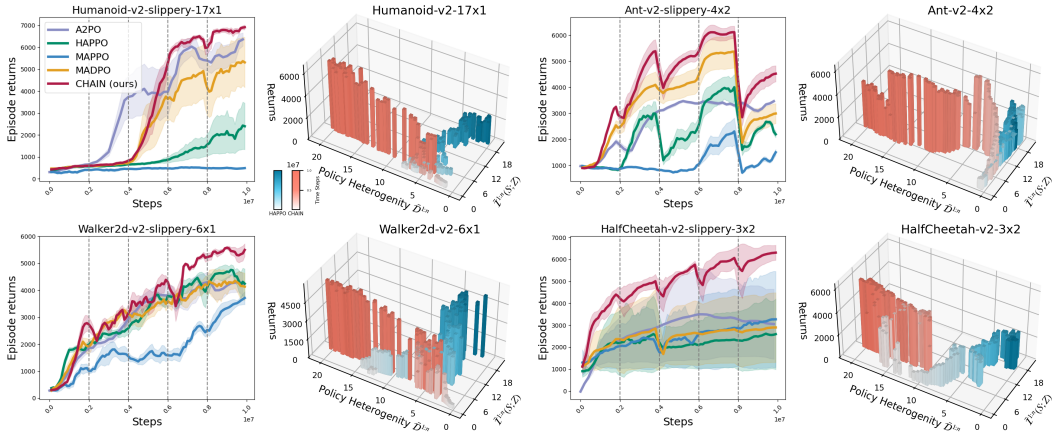


Figure 4: Performance comparison with MARL baselines in the slippery MA-Mujoco environments.

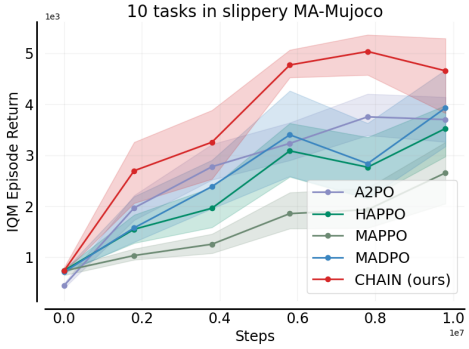


Figure 5: Aggregate IQM performance comparison against baseline methods in 10 tasks of slippery MA-Mujoco. CHAIN achieves statistically significant improvements over baselines.

Comprehensive comparison. Results in column 2 and 4 further reveal that the policies learned by CHAIN exhibit significantly higher policy heterogeneity and impose a greater restriction on learning from the current environment, compared to the baseline. Additionally, these results confirm our key insight: reducing representation redundancy and promoting heterogenization can be the key to enhance the continual learning ability of MARL methods. Interestingly, we also find from the `Ant-slippery` environment that, as training of CHAIN progresses, the trend of $\bar{\mathcal{I}}^{1:n}(S;Z)$ is consistent with the information plane reported in IB-based supervised learning [Tishby et al., 2000]. The information gained from the environment increases first to fit the target, and then decreases to compress redundant information. This finding provides an explanation, theoretically grounded by IB, of the representation learning process in MARL.

Training curves in MEAL. Fig. 6 compares HPB against continual RL baselines in the 15-task `Level 1` and `Level 2` MEAL environment. We can observe that HPB achieves higher overall returns than baselines in long-term continual learning, with a steadily rising training curve indicating

better forward transfer ability. These results confirm that HPB, benefiting from compression and heterogenization, can continually learn reliable policies in long-term multi-task environments.

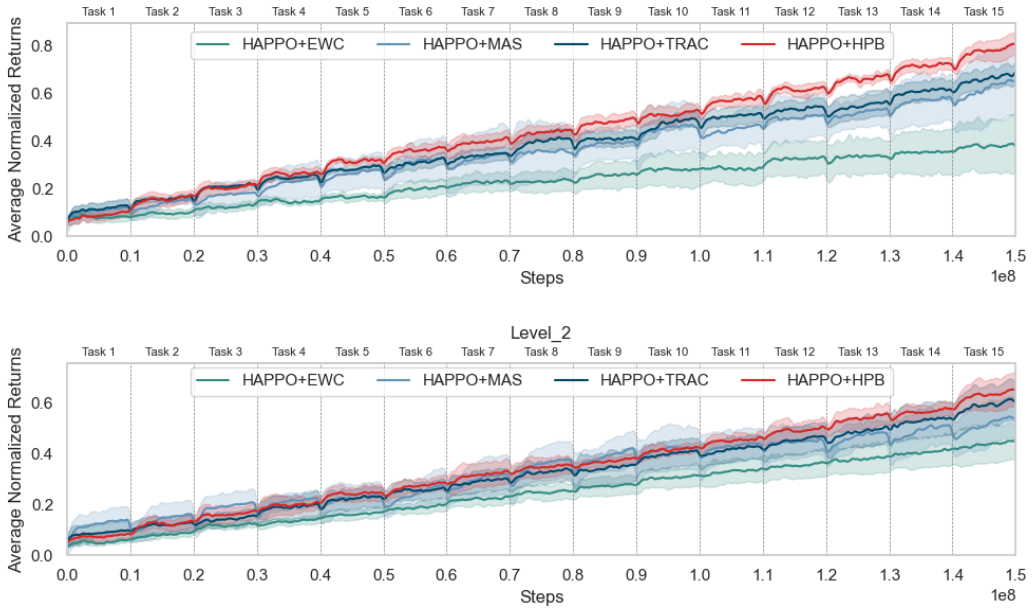


Figure 6: Performance comparison with continual learning baselines in MEAL environment.

Evaluation on Continual Learning Metrics. In this section, we evaluate CHAIN and baselines using continual leanring metrics. Since learning effective policies on standard MA-Mujoco is already challenging for existing MARL methods, we evaluate CHAIN and baselines on slippery MA-Mujoco using two metrics: mean maximum return $\bar{R}_{\max} = \frac{1}{N}\sum_{i=1}^{N} R_{i,i}$ and mean performance drop after environment changes $\bar{P}_{\text{drop}} = \frac{1}{N}\sum_{i=1}^{N} R_{i,i+1}/R_{i,i}$, where $R_{i,j}$ denotes the maximum return of policies trained on task i and evaluated on task j. For MEAL tasks, we measure forward transfer $FT = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N-i}\sum_{j=i+1}^{N} R_{i,j} - R_{i-1,j}$ and forgetting $FG = \frac{1}{N}\sum_{i=1}^{N}\max(R_{i,i} - R_{N,i}, 0)$.

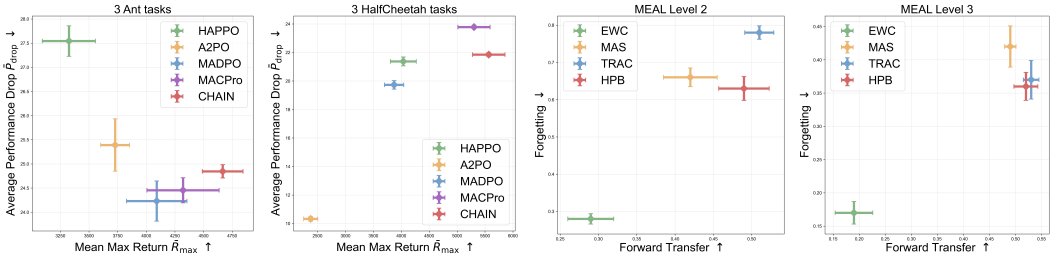


Figure 7: Continual learning performance comparison.

The results in Fig. 7 show that CHAIN achieves higher $\bar{R}_{\max}$ and lower $\bar{P}_{\text{drop}}$ compared to heterogeneous MARL baselines in slippery MA-Mujoco environments, and compared to continual learning baselines in MEAL environments. These results confirm that by combining compression and heterogenization, CHAIN achieves a better balance between learning environment-invariant and environment-specific dynamics.

5.3 Ablation Study

In CHAIN, hyperparameters β and γ control the temperature of compression and heterogenization respectively, and we study their effects in this part. The results in Fig. 8 reveal complementary trade-offs

between the two components. Increasing compression β significantly reduces $\bar{P}_{\text{drop}}$ but moderately decreases $\bar{R}_{\text{max}}$, as compressing information forces the policy to focus on environment-invariant information that generalize across tasks, sacrificing task-specific peak performance. Conversely, increasing heterogenization γ substantially improves $\bar{R}_{\text{max}}$, but slightly increases $\bar{P}_{\text{drop}}$ due to reduced flexibility. Critically, disabling both ($\beta = 0, \gamma = 0$) or reversing them ($\beta < 0, \gamma < 0$) causes severe degradation, confirming that both components are essential and effective. This confirms that both compression and heterogenization are essential: compression focuses on generalizable representations, while heterogenization maximizes performance through specialization.

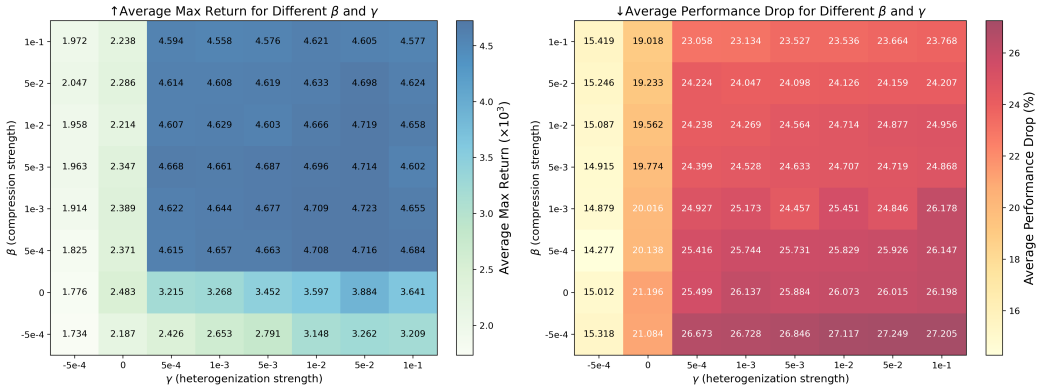


Figure 8: The effects of heterogenization strength and compression strength of CHAIN.

6 Conclusion

In this work, we present CHAIN, a continual heterogeneous MARL method, equipped with Heterogeneous Policy Bottleneck (HPB). We first develop a new MA-Mujoco environment, named slippery MA-Mujoco, where the environment dynamics change periodically. We demonstrate that existing MARL methods can fail in such an unstable scenario due to the representation redundancy and insufficient heterogenization. To this end, we extend the traditional IB to HPB, which contains three parts, fitting, compression and heterogenization. Our HPB can bottleneck both the representation redundancy and the homogenization. Empirical results indicate that CHAIN outperforms state-of-the-art MARL methods on our slippery MA-Mujoco and MEAL.

Acknowledgments and Disclosure of Funding

This work is supported by the National Key R&D Program of China (2023YFB4704900) and National Natural Science Foundation of China (U21A20485, 32541016, 62436005). The authors declare that they have no conflict of interest.

References

R. Agarwal, M. Schwarzer, P. S. Castro, A. C. Courville, and M. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in Neural Information Processing Systems*, 34:29304–29320, 2021.

A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–154, 2018.

C. Bai, L. Wang, L. Han, A. Garg, J. Hao, P. Liu, and Z. Wang. Dynamic bottleneck for robust self-supervised exploration. *Advances in Neural Information Processing Systems*, 34:17007–17020, 2021.

D. Bertsekas. Multiagent reinforcement learning: Rollout and policy iteration. *IEEE/CAA Journal of Automatica Sinica*, 8(2):249–272, 2021.

M. Bettini, A. Shankar, and A. Prorok. Heterogeneous multi-robot reinforcement learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 1485–1494, 2023.

M. Caccia, J. Mueller, T. Kim, L. Charlin, and R. Fakoor. Task-agnostic continual reinforcement learning: Gaining insights and overcoming challenges. In *Conference on Lifelong Learning Agents*, pages 89–119. PMLR, 2023.

M. Carroll, R. Shah, M. K. Ho, T. Griffiths, S. Seshia, P. Abbeel, and A. Dragan. On the utility of learning about humans for human-ai coordination. In *Advances in Neural Information Processing Systems*, volume 32, 2019a.

M. Carroll, R. Shah, M. K. Ho, T. Griffiths, S. Seshia, P. Abbeel, and A. Dragan. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32, 2019b.

W. Chung, L. Cherif, D. Meger, and D. Precup. Parseval regularization for continual reinforcement learning. *Advances in Neural Information Processing Systems*, 37:127937–127967, 2024.

T. M. Cover. *Elements of information theory*. John Wiley & Sons, 1999.

S. Ding, W. Du, L. Ding, J. Zhang, L. Guo, and B. An. Multiagent reinforcement learning with graphical mutual information maximization. *IEEE Transactions on neural networks and learning systems*, 2023a.

S. Ding, W. Du, L. Ding, J. Zhang, L. Guo, and B. An. Robust multi-agent communication with graph information bottleneck optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3096–3107, 2023b.

S. Dohare, J. F. Hernandez-Garcia, Q. Lan, P. Rahman, A. R. Mahmood, and R. S. Sutton. Loss of plasticity in deep continual learning. *Nature*, 632(8026):768–774, 2024.

H. Dou, L. Dang, Z. Luan, and B. Chen. Measuring mutual policy divergence for multi-agent sequential exploration. *Advances in Neural Information Processing Systems*, 37:76265–76288, 2024.

H. Dou, L. Dang, M. Lu, L. Chen, and B. Chen. Efficient heterogeneous exploration with mutual policy divergence maximization for multiagent reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

M. Elsayed and A. R. Mahmood. Addressing loss of plasticity and catastrophic forgetting in continual learning. In *International Conference on Learning Representations*, 2024.

J. Fan and W. Li. Dribo: Robust deep reinforcement learning via multi-view information bottleneck. In *International Conference on Machine Learning*, pages 6074–6102. PMLR, 2022.

T. Haarnoja, B. Moran, G. Lever, S. H. Huang, D. Tirumala, J. Humplik, M. Wulfmeier, S. Tunyasuvunakool, N. Y. Siegel, R. Hafner, et al. Learning agile soccer skills for a bipedal robot with deep reinforcement learning. *Science Robotics*, 9(89):eadi8022, 2024.

J. Hou, H. Dou, L. Dang, L. Chen, and C. Ge. Gradient-protected value decomposition for cooperative multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 21779–21787, 2026.

H. Hwang, G.-H. Kim, S. Hong, and K.-E. Kim. Variational interaction information maximization for cross-domain disentanglement. *Advances in Neural Information Processing Systems*, 33: 22479–22491, 2020.

M. Igl, K. Ciosek, Y. Li, S. Tschiatschek, C. Zhang, S. Devlin, and K. Hofmann. Generalization in reinforcement learning with selective noise injection and information bottleneck. *Advances in Neural Information Processing Systems*, 32, 2019.

Y. Jin, S. Wei, J. Yuan, and X. Zhang. Information-bottleneck-based behavior representation learning for multi-agent reinforcement learning. In *2021 IEEE International Conference on Autonomous Systems (ICAS)*, pages 1–5. IEEE, 2021.

J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

P. Knott, M. Carroll, S. Devlin, K. Ciosek, K. Hofmann, A. Dragan, and R. Shah. Evaluating the robustness of collaborative agents. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1560–1562, 2021.

J. Kuba, R. Chen, M. Wen, Y. Wen, F. Sun, J. Wang, and Y. Yang. Trust region policy optimisation in multi-agent reinforcement learning. In *International Conference on Learning Representations*, page 1046, 2022.

A. Kumar, R. Agarwal, D. Ghosh, and S. Levine. Implicit under-parameterization inhibits data-efficient deep reinforcement learning. In *International Conference on Learning Representations*, 2021.

S. Kumar, H. Marklund, A. Rao, Y. Zhu, H. J. Jeon, Y. Liu, B. Van Roy, et al. Continual learning as computationally constrained reinforcement learning. *Foundations and Trends® in Machine Learning*, 18(5):913–1053, 2025.

S. Li, R. Xu, J. Xiu, Y. Zheng, P. Feng, Y. Ma, B. An, Y. Yang, and X. Liu. Robust multi-agent reinforcement learning by mutual information regularization. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

X. Li, L. Pan, and J. Zhang. Kaleidoscope: Learnable masks for heterogeneous multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 37:22081–22106, 2024.

C. Ma, A. Li, Y. Du, H. Dong, and Y. Yang. Efficient and scalable reinforcement learning for large-scale network control. *Nature Machine Intelligence*, 6(9):1006–1020, 2024.

S. Moalla, A. Miele, D. Pyatko, R. Pascanu, and C. Gulcehre. No representation, no trust: Connecting representation, collapse, and trust issues in ppo. *Advances in Neural Information Processing Systems*, 37:69652–69699, 2024.

J. S. Monon, D. D. Barua, and M. M. Khan. Learning heterogeneous agent collaboration in decentralized multi-agent systems via intrinsic motivation. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pages 2681–2683, 2025.

A. Muppidi, Z. Zhang, and H. Yang. Fast trac: A parameter-free optimizer for lifelong reinforcement learning. *Advances in Neural Information Processing Systems*, 37:51169–51195, 2024.

C. Pan, L. Ren, Y. Feng, L. Xiong, W. Wei, Y. Li, and X. Yang. Multi-granularity knowledge transfer for continual reinforcement learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 6012–6020, 2025.

G. Sokar, R. Agarwal, P. S. Castro, and U. Evci. The dormant neuron phenomenon in deep reinforcement learning. In *International Conference on Machine Learning*, pages 32145–32168. PMLR, 2023.

C. Tang, B. Abbatematteo, J. Hu, R. Chandra, R. Martín-Martín, and P. Stone. Deep reinforcement learning for robotics: A survey of real-world successes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28694–28698, 2025a.

H. Tang, J. Obando-Ceron, P. S. Castro, A. Courville, and G. Berseth. Mitigating plasticity loss in continual reinforcement learning by reducing churn. In *International Conference on Machine Learning*, 2025b.

N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

T. Tomilin, L. v. d. Boogaard, S. Garcin, B. Grooten, M. Fang, and M. Pechenizkiy. Meal: A benchmark for continual multi-agent reinforcement learning. *arXiv preprint arXiv:2506.14990*, 2025.

R. Wang, X. He, R. Yu, W. Qiu, B. An, and Z. Rabinovich. Learning efficient multi-agent communication: An information bottleneck approach. In *International conference on machine learning*, pages 9908–9918. PMLR, 2020.

X. Wang, Z. Tian, Z. Wan, Y. Wen, J. Wang, and W. Zhang. Order matters: Agent-by-agent policy optimization. In *International Conference on Learning Representations*, pages 1–35, 2023.

J. Yang, H. Dou, L. Dang, L. Chen, and B. Chen. Experience replay based on reducible policy entropy in off-policy reinforcement learning. *Neurocomputing*, page 135417, 2026.

C. Yao, Y. Lin, S. Song, H. Wu, Y. Ma, S. Han, and K. Lv. From general relation patterns to task-specific decision-making in continual multi-agent coordination. In *Proceedings of the Thirty-Forth International Joint Conference on Artificial Intelligence*, 2025.

J. You, Y. Wu, D. Wen, Y. Zhou, Y. Jiang, and Y. Shi. Dynamic communication in multi-agent reinforcement learning via information bottleneck. In *GLOBECOM 2024-2024 IEEE Global Communications Conference*, pages 4884–4889. IEEE, 2024.

C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu. The surprising effectiveness of ppo in cooperative multi-agent games. In *Advances in Neural Information Processing Systems*, volume 35, pages 24611–24624, 2022a.

J. Yu, T. Xu, Y. Rong, Y. Bian, J. Huang, and R. He. Graph information bottleneck for subgraph recognition. In *International Conference on Learning Representations*, 2021.

J. Yu, J. Cao, and R. He. Improving subgraph recognition with variational graph information bottleneck. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19396–19405, 2022b.

S. Yu, L. G. S. Giraldo, R. Jenssen, and J. C. Príncipe. Multivariate extension of matrix-based rényi's α-order entropy functional. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42 (11):2960–2966, 2020.

L. Yuan, L. Li, Z. Zhang, F. Zhang, C. Guan, and Y. Yu. Multiagent continual coordination via progressive task contextualization. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):6326–6340, 2024.

P. Zhai and S. Zhang. Adversarial information bottleneck. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):221–230, 2022.

T. Zhang, Z. Lin, Y. Wang, D. Ye, Q. Fu, W. Yang, X. Wang, B. Liang, B. Yuan, and X. Li. Dynamics-adaptive continual reinforcement learning via progressive contextualization. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10):14588–14602, 2023.

X. Zhang, X. Wang, X. Liu, W. Wang, X. Fan, and D. Zhao. Task-agnostic learning to accomplish new tasks. *IEEE Transactions on Cognitive and Developmental Systems*, 2025.

Y. Zhong, J. G. Kuba, X. Feng, S. Hu, J. Ji, and Y. Yang. Heterogeneous-agent reinforcement learning. *Journal of Machine Learning Research*, 25(32):1–67, 2024.

M. Zhou, J. Luo, J. Villella, Y. Yang, D. Rusu, J. Miao, W. Zhang, M. Alban, I. Fadakar, Z. Chen, et al. Smarts: An open-source scalable multi-agent rl training school for autonomous driving. In *Conference on Robot Learning*, pages 264–285. PMLR, 2021.

7 Appendix

7.1 Theoretical Proofs

Proposition 7.1. *At episode t, suppose the current state $S \sim \mathcal{T}_{current}(\cdot|s_{t-1}, a_{t-1}, \phi)$ is derived from the former state $S_{former} \sim \mathcal{T}_{former}(\cdot|s_{t-1}, a_{t-1})$ through environment change ϕ. Define the bias induced by the environment change as $\tilde{S} = S - S_{former}$. Then, the mutual information of $\tilde{S}$ and z_t^i, $\mathcal{I}(\tilde{S}; z_t^i)$, is upper bounded as follows,*

$$\mathcal{I}(\tilde{S}; z_t^i) \leq \mathcal{I}(S; z_t^i) - \mathcal{I}(a_{t-1}^i; z_t^i). \tag{17}$$

Proof. Inspired by Yu et al. [2021], suppose that z_t^i obtains information from $\tilde{S}$ only through S, and S is defined by a_{t-1}^i and $\tilde{S}$, we can define the following Markov chain $(a_{t-1}^i, \tilde{S}) \to S \to z_t^i$. Using the Data Processing Inequality, we obtain:

$$\begin{aligned}
\mathcal{I}(S; z_t^i) &\geq \mathcal{I}(a_{t-1}^i, \tilde{S}; z_t^i) && (18)\\
&= \mathcal{I}(\tilde{S}; z_t^i) + \mathcal{I}(a_{t-1}^i; z_t^i|\tilde{S}) && (19)\\
&= \mathcal{I}(\tilde{S}; z_t^i) + \mathcal{H}(a_{t-1}^i|\tilde{S}) - \mathcal{H}(a_{t-1}^i|\tilde{S}, z_t^i), && (20)
\end{aligned}$$

Consider a multi-task continual MARL setting (*e.g.* the MEAL environment), where the positions of immovable entities changes across tasks, we can define the states before and after the environmental change as S_{former} and S. Then, we can decompose the state S into the entity state S^e predefined by the current task, and the agent state $S^a \sim \mathcal{T}^a(\cdot|a_{t-1,S_{t-1}^a})$ as $S = \begin{pmatrix} S^e \\ S^a \end{pmatrix}$. Since the environment changes only affect the entity state, we have $\mathcal{T}^a(\cdot|a_{t-1,S_{t-1}^a}) = \mathcal{T}^a_{former}(\cdot|a_{t-1,S_{t-1}^a})$. Therefore, we have $\tilde{S} = \begin{pmatrix} S^e - S^e_{former} \\ S^a - S^a_{former} \end{pmatrix} = \begin{pmatrix} S^e - S^e_{former} \\ 0 \end{pmatrix}$, which is independent of the action a_{t-1}^i. Thus, we obtain $\mathcal{H}(a_{t-1}^i|\tilde{S}) = \mathcal{H}(a_{t-1}^i)$. Moreover, we have $\mathcal{H}(a_{t-1}^i|\tilde{S}, z_t^i) \leq \mathcal{H}(a_{t-1}^i|z_t^i)$. Combine them with Eq. 20, and we obtain:

$$\begin{aligned}
\mathcal{I}(S; z_t^i) &\geq \mathcal{I}(\tilde{S}; z_t^i) + \mathcal{H}(a_{t-1}^i) - \mathcal{H}(a_{t-1}^i|z_t^i) && (21)\\
&= \mathcal{I}(\tilde{S}; z_t^i) + \mathcal{I}(a_{t-1}^i; z_t^i), && (22)
\end{aligned}$$

thus we obtain $\mathcal{I}(\tilde{S}, z_t^i) \leq \mathcal{I}(S, z_t^i) - \mathcal{I}(a_{t-1}^i, z_t^i)$, which completes the proof. □

Proposition 7.2. *Given state S, a latent representation of current agent to be updated z^i, and a fixed latent representation of one preceding agent $\bar{z}^{i-1}$, then we have,*

$$\mathcal{I}(z^i; \bar{z}^{i-1}) = \underbrace{\mathcal{I}(\bar{z}^{i-1}; S)}_{\textit{Static}} + \underbrace{\mathcal{I}(z^i; S)}_{\substack{\textit{Retain Info.}\\ \textit{Individually}}} - \underbrace{\mathcal{I}(z^i, \bar{z}^{i-1}; S)}_{\substack{\textit{Retain Info.}\\ \textit{Jointly}}} + \underbrace{\mathcal{I}(z^i; \bar{z}^{i-1}|S)}_{\textit{Heterogenization}}. \tag{23}$$

Proof. Inspired by Hwang et al. Hwang et al. [2020], we first introduce the state S and the interaction information Cover [1999] between z^i, $\bar{z}^{i-1}$ and S as follows,

$$\begin{aligned}
\mathcal{I}(z^i; \bar{z}^{i-1}; S) =& \mathcal{I}(z^i; \bar{z}^{i-1}) - \mathcal{I}(z^i; \bar{z}^{i-1}|S), && (24)\\
=& \mathcal{I}(z^i; S) - \mathcal{I}(z^i; S|\bar{z}^{i-1}), && (25)\\
=& \mathcal{I}(\bar{z}^{i-1}; S) - \mathcal{I}(\bar{z}^{i-1}; S|z^i). && (26)
\end{aligned}$$

Combining Eq. 24 and Eq. 25, we obtain the following equation,

$$\mathcal{I}(z^i; \bar{z}^{i-1}) = \mathcal{I}(z^i; S) - \mathcal{I}(z^i; S|\bar{z}^{i-1}) + \mathcal{I}(z^i; \bar{z}^{i-1}|S). \tag{27}$$

We also have $\mathcal{I}(z^i; S|\bar{z}^{i-1}) = \mathcal{I}(z^i, \bar{z}^{i-1}; S) - \mathcal{I}(\bar{z}^{i-1}; S)$, and thus,

$$\mathcal{I}(z^i; \bar{z}^{i-1}) = \mathcal{I}(\bar{z}^{i-1}; S) + \mathcal{I}(z^i; S) - \mathcal{I}(z^i, \bar{z}^{i-1}; S) + \mathcal{I}(z^i; \bar{z}^{i-1}|S), \tag{28}$$

which completes the proof.

7.2 Pseudocode of CHAIN

Algo. 1 shows the total algorithm of CHAIN. In Algo 1, the sequential update method in CHAIN is the same as that in HAPPO. Specifically, at each iteration, CHAIN maximizes the first two terms of HPB in Eq. 16 when updating the first agent, since it has no preceding partner. After that, CHAIN evaluates the previous policy networks to obtain the latent representation $\bar{z}^{i-1}$ and computes Eq. 16 for updating.

Algorithm 1 Continual Heterogeneous Cooperation with Information Bottleneck (CHAIN)

Require: Initial joint policy $\boldsymbol{\pi}_0 = \pi_0^1 \times ... \times \pi_0^n$, initial global value network V^{ψ}, parameters of HPB, β, γ and σ.

for iteration $k = 1, ..., K$ **do**
  Collection trajectories $\mathbf{T}_k = \{\boldsymbol{\tau}_k^1, ..., \boldsymbol{\tau}_k^n\}$ by running $\bar{\boldsymbol{\pi}}_k = \bar{\pi}_k^1 \times ... \times \bar{\pi}_k^n$.
  Restore $\mathbf{T}_k$ into the buffer.
  Compute the advantage $\hat{A}(s, \boldsymbol{a})$ by using the value network and GAE.
  for agent $i = 1, ..., n$ **do**
    Evaluate the policy network of the current agent π^i using the minibatch to obtain z^i.
    Compute $\mathcal{J}_{compress} = \mathcal{I}(S; z^i)$ via Eq. 11.
    if $i > 1$ **then**
      Evaluate the policy network of the preceding agent $\bar{\pi}^{i-1}$ using the minibatch to obtain $\bar{z}^{i-1}$.
      Compute the $\mathcal{J}_{heterogenize} = \mathcal{I}(z^i; \bar{z}^{i-1})$ via Eq. 14.
    end if
    Compute $\mathcal{J}_{fit}$ via Eq. 9.
    Compute $\mathcal{J}_{total} = \mathcal{J}_{fit} - \beta\mathcal{J}_{compress} - \gamma\mathcal{J}_{heterogenize}$
    Update the actor network $\pi_{k+1}^i = \arg\max_{\pi_k^i} \mathcal{J}_{total}$
  end for
  Compute the value target $V_{target} = \hat{A}(s_t, \boldsymbol{a}_t) + V(s_t)$.
  Update the value network $V^{\psi k+1} = \arg\min_{\psi_k} \mathbb{E}_{\mathbf{T}_k} \|V^{\phi_k}(s_t) - V_{target}\|^2$.
end for

7.3 Experiment Details

Baselines and Benchmarks In this paper, we compare CHAIN with heterogeneous MARL methods, following their official implementations, including MAPPO [1] [Yu et al., 2022a], HAPPO [2] [Kuba et al., 2022], A2PO [3] [Wang et al., 2023], MADPO [4] [Dou et al., 2024]. We also compare CHAIN with continual learning baselines, including EWC [5] [Kirkpatrick et al., 2017], MAS [5] [Aljundi et al., 2018], and TRAC [6] [Muppidi et al., 2024]. We evaluate CHAIN using a popular multi-joint control environments, MA-Mujoco [7], and a long-term multi-task continual MARL benchmark, MEAL [5] Tomilin et al. [2025].

Parameter settings Tab. 1 presents the common hyperparameter settings of CHAIN in all the scenarios used in this paper. Tab. 2 presents the hyperparameter settings of CHAIN in different scenarios of standard MA-Mujoco, while Tab. 3 presents the hyperparameter settings in slippery MA-Mujoco. Here, we keep the parameter settings consistent with the official implementations of baselines. Our method adopts HAPPO as the backbone. To ensure a controlled comparison, we use the same parameter settings as HAPPO for the shared components and only tune the parameters introduced by our method. All experiments were carried out on a workstation equipped with four NVIDIA RTX 3090 GPUs, one Intel Xeon 64-core CPU, and 128 GB of RAM.

[1] https://github.com/marlbenchmark/on-policy
[2] https://github.com/PKU-MARL/HARL
[3] https://github.com/xihuai18/A2PO-ICLR2023
[4] https://github.com/hwdou6677/MADPO
[5] https://github.com/TTomilin/MEAL
[6] https://github.com/ComputationalRobotics/TRAC
[7] https://github.com/schroederdewitt/multiagent_mujoco

MEAL environment details MEAL [Tomilin et al., 2025] builds upon the Overcooked environment [Carroll et al., 2019a]. MEAL generates a sequence of tasks with random positions of interactive stations and obstacles of Overcooked before training. Recent study has shown that agents in small layout variations of Overcooked exhibit poor generalization ability [Knott et al., 2021]. Thus, it is well-suited to use MEAL for multi-task continual MARL evaluation. Tasks with different difficulties in MEAL have different map sizes. Specifically, `Level_1`, `Level_2` and `Level_3` have maximum 7×7, 9×9 and 11×11 maps respectively, as well as 15%, 25% and 35% obstacle density respectively. In general, `Level_1`, `Level_2` and `Level_3` are of increasing difficulty: easy, medium, and hard. In this study, we set the number of sequential tasks as 15.

Table 1: Common hyperparameters in both standard and slippery MA-Mujoco

parameters	batch size	gamma	gain	episode length	rollout threads
value	4000	0.99	0.01	200	20

Table 2: Different hyperparameters in standard MA-Mujoco

Tasks	latent layers	actor lr	critic lr	clip	β	γ	σ
Ant-v2-2x4	[64,64]	5e-4	5e-4	0.2	1e-3	5e-2	5e-3
Ant-v2-4x2	[64,64]	5e-4	5e-4	0.2	1e-3	5e-2	5e-3
Ant-v2-8x1	[64,64]	5e-4	5e-4	0.2	1e-3	5e-2	5e-3
HalfCheetah-v2-2x3	[64,64]	5e-4	5e-4	0.2	1e-3	1e-2	5e-3
HalfCheetah-v2-3x2	[64,64]	5e-4	5e-4	0.2	1e-3	1e-2	5e-3
HalfCheetah-v2-6x1	[64,64]	5e-4	5e-4	0.2	1e-3	1e-2	5e-3
Walker2d-v2-2x3	[256,256]	1e-3	1e-3	0.05	5e-2	1e-2	5e-3
Walker2d-v2-3x2	[256,256]	1e-3	1e-3	0.05	5e-2	1e-2	5e-3
Walker2d-v2-6x1	[256,256]	1e-3	1e-3	0.05	5e-2	1e-2	5e-3
Humanoid-v2-17x1	[256,256]	5e-4	5e-4	0.1	5e-2	1e-2	5e-2

Table 3: Different hyperparameters in slippery MA-Mujoco

Tasks	latent layers	actor lr	critic lr	clip	β	γ	σ
Ant-v2-2x4	[64,64]	5e-4	5e-4	0.2	1e-3	5e-2	1e-3
Ant-v2-4x2	[64,64]	5e-4	5e-4	0.2	1e-3	5e-2	1e-3
Ant-v2-8x1	[64,64]	5e-4	5e-4	0.2	1e-3	5e-2	1e-3
HalfCheetah-v2-2x3	[64,64]	5e-4	5e-4	0.2	1e-3	1e-3	5e-3
HalfCheetah-v2-3x2	[64,64]	5e-4	5e-4	0.2	1e-3	1e-3	5e-3
HalfCheetah-v2-6x1	[64,64]	5e-4	5e-4	0.2	1e-3	1e-3	5e-3
Walker2d-v2-2x3	[256,256]	1e-3	1e-3	0.05	5e-2	1e-3	5e-3
Walker2d-v2-3x2	[256,256]	1e-3	1e-3	0.05	5e-2	1e-3	5e-3
Walker2d-v2-6x1	[256,256]	1e-3	1e-3	0.05	5e-2	1e-3	5e-3
Humanoid-v2-17x1	[256,256]	5e-4	5e-4	0.1	5e-2	1e-3	5e-2

Table 4: Hyperparameters in MEAL

Difficulty	lr anneal	latent layers	rollout threads	use multi-head	task id input	β	γ	σ
Level 1	False	[128,128]	16	True	False	1e-2	1e-2	1e-3
Level 2	True	[128,128]	16	True	False	1e-2	1e-2	1e-3
Level 3	True	[128,128]	16	True	False	1e-3	1e-3	1e-3

7.4 Complete Results in Slippery and Standard MA-Mujoco

Fig. 9 shows the results of complete 10 scenarios of slippery MA-Mujoco. It is clear that the proposed CHAIN consistently outperforms other baselines in terms of both maximum returns and sample efficiency. Notably, our proposed method, which explicitly compresses both task-irrelevant information and heterogenize agents, achieves the lowest measured redundancy while showing state-of-the-art performance. The results reveal a strong inverse correlation between overall performance

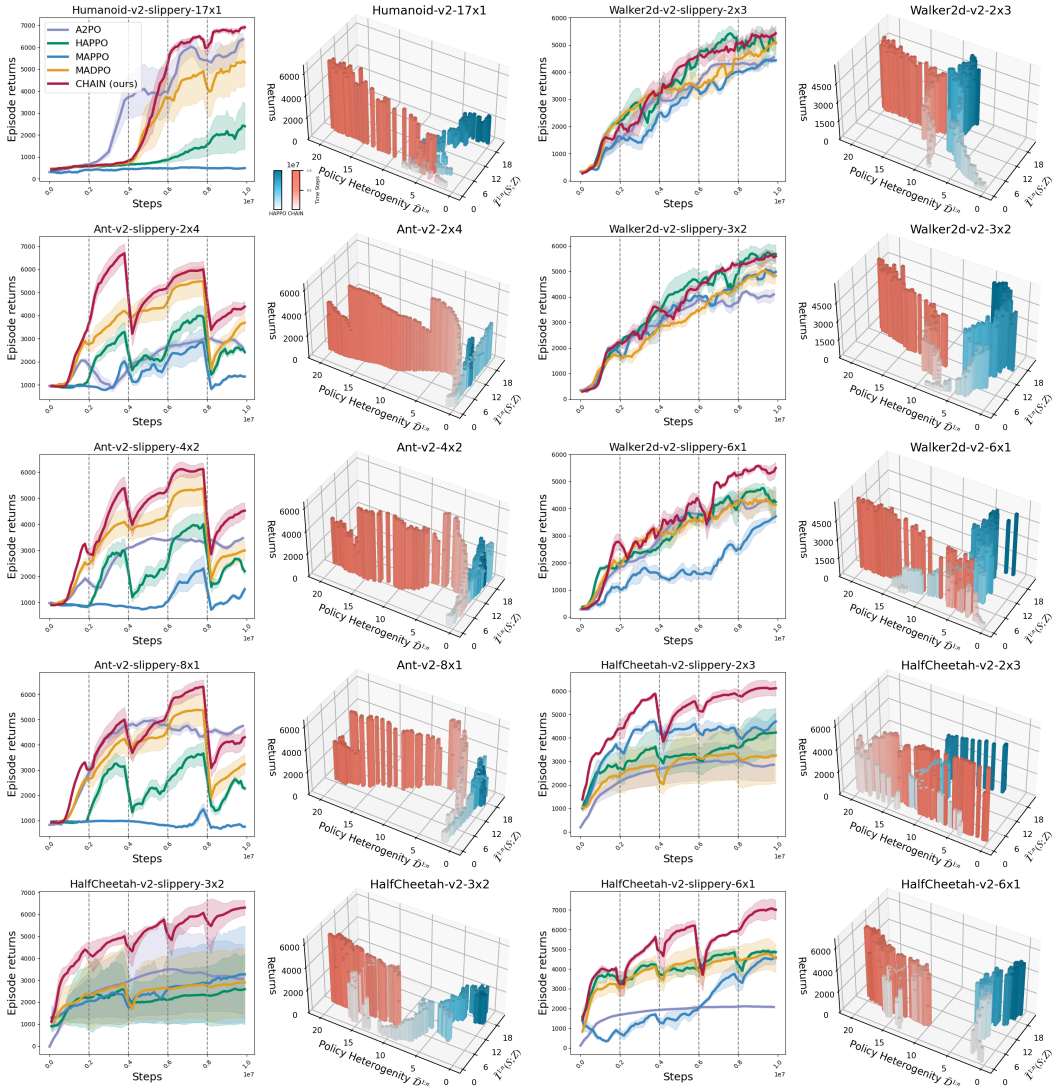


Figure 9: Performance comparison with MARL baselines in the slippery MA-Mujoco environments.

and the degrees of homogenization redundancy and representation redundancy, which is, more importantly, a potential optimization direction for continual heterogeneous MARL.

Although our method is primarily designed for continual reinforcement learning, we additionally evaluate its performance in standard non-continual environments. Fig. 10 shows the performance comparison in complete 10 scenarios of standard MA-Mujoco. Across a diverse set of 10 environments, our approach achieves comparable or slightly improved performance relative to state-of-the-art methods in 8 cases. This suggests that the proposed method maintains competitive performance across a relatively large set of environments. While our primary goal focuses in continual environments, we also report the correlation between representation redundancy, policy heterogenization, and overall performance in standard MA-Mujoco. Given that standard environments can be considered a simplified variant of continual environments, such correlation may still exist, albeit to a lesser extent. Empirically, we observe weaker correlations from Fig. 10, which are aligned with the trends identified in the slippery setting.

To ensure statistical significance, we also evaluate the aggregate Interquartile Mean (IQM) [Agarwal et al., 2021] of CHAIN and baselines across multiple scenarios. Fig. 11 presents the aggregate performance across 10 tasks in standard MA-Mujoco. These results indicate that, in the standard scenarios, CHAIN obtains a slight performance improvement compared to the state-of-the-art ex-

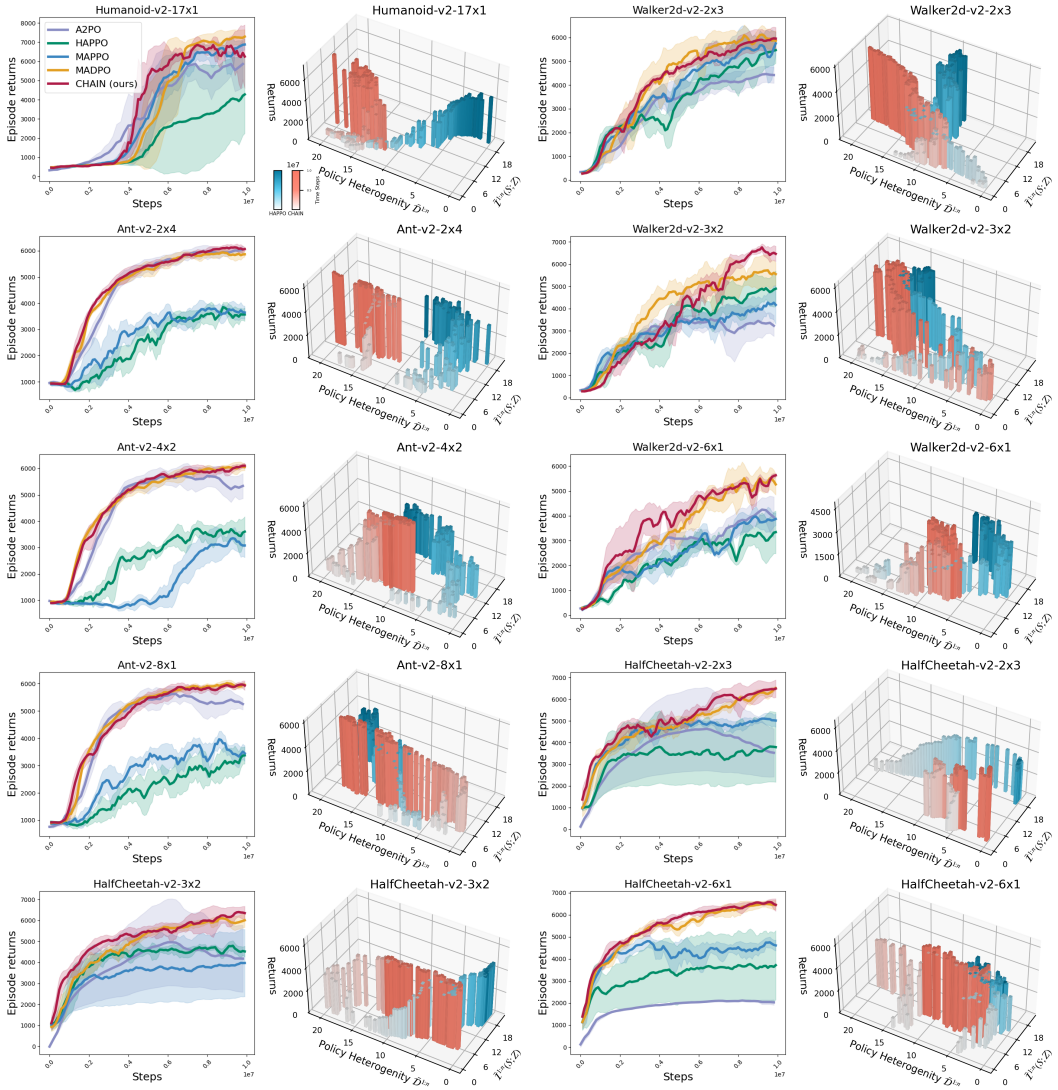


Figure 10: Complete performance comparison with MARL baselines in the standard Multi-Agent Mujoco environments. In standard scenarios, CHAIN achieves improved performance over the baselines, primarily due to its efficient latent representation and heterogenization incentive.

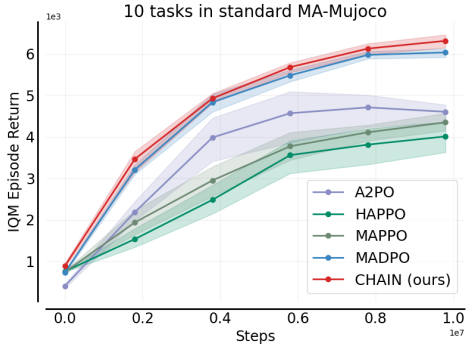


Figure 11: Aggregate IQM performance comparison against baseline methods in 10 tasks of standard MA-Mujoco. CHAIN achieves statistically significant improvements over baselines.

ploration strategy. On the other hand, in scenarios that require continual heterogeneous learning in Fig. 5, CHAIN obtains a notable improvement. Overall, CHAIN demonstrates its superiority with statistical significance in heterogeneous MARL environments, particularly where continual learning are required.

7.5 More Analysis on the Effects of Representation Redundancy and Heterogenization

We provide a deeper analysis of CHAIN's ability to compress redundancy and enhance agent heterogeneity, as illustrated in Fig. 12. We observe that our method consistently outperforms the baseline in terms of returns.

Specifically, our method has two key advantages. First, after an environmental change, the performance drop is smaller than that of the baseline, indicating better stability. Second, after the environmental change, the performance improvement is faster than the baseline, showing a quicker adaptation. We can alse observe that the mutual information term $\bar{\mathcal{I}}^{1:n}(S;Z)$, which indicates the degree of learning of latent representations from current environment, shows a decline after the environmental change in our method, while the baseline method maintained a higher level. Additionally, the degree of policy heterogeneity in our method consistently remained higher than that of the baseline. We attribute the lower performance decline after the environmental change to our method's optimization of representation redundancy. This reduction indicates that the agents are no longer focusing to environment-specific dynamics, but instead are focusing on more stable, environment-agnostic information. This shift is crucial for transferring knowledge from past environments to new ones. On the other hand, we attribute the faster re-adaptation to environmental changes to the higher degree of policy heterogeneity in our method. The increased heterogeneity promotes diverse exploration of the policy space, allowing our method to recover and adapt more quickly after an environmental change. Specifically, in Fig. 12, before 0.4M steps, benefiting from the heterogenization incentive, our method show better sample efficiency. In this stage, both methods are learning from the environment, with almost no difference in $\bar{\mathcal{I}}^{1:n}(S;Z)$. At 0.4M steps, an environmental change is introduced to evaluate the continual learning capabilities of the methods. After that, both the two methods exhibit a significant performance drop. However, following this drop, the performance of our method remains consistently higher than that of the baseline. This indicates that our method has successfully learned environment-invariant information that helps mitigate the performance decline. The experimental

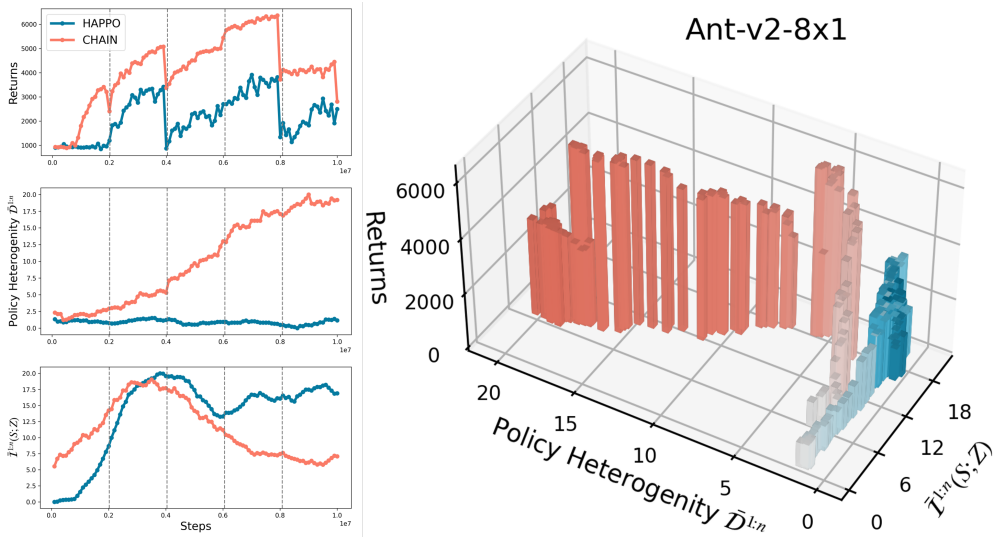


Figure 12: Comparison between our CHAIN and the baseline in terms of returns, compression and heterogenization on slippery `Ant-8x1`. Each dashed line on the left represents a friciton change.

results suggest a correlation between the overall performance and the representation redundancy as well as the policy heterogeneity. These results further emphasize the importance of considering both representation redundancy and policy heterogeneity in continual MARL for better performance.

7.6 Additional Results in MEAL

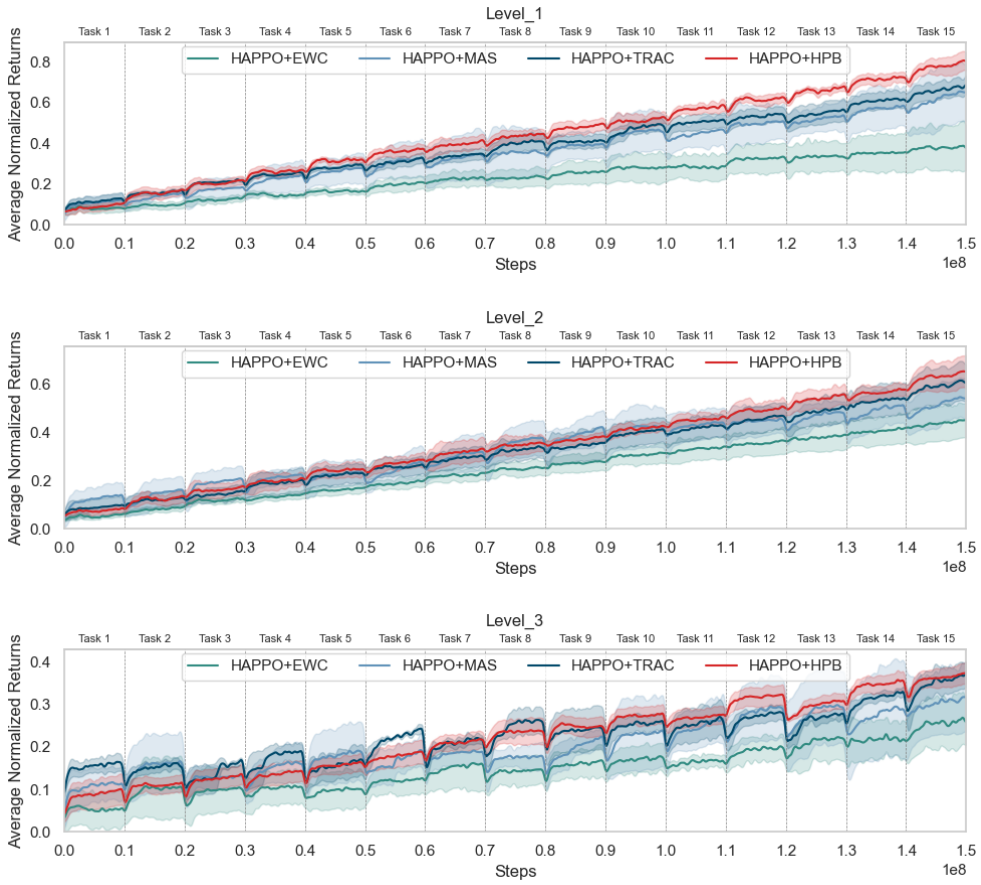


Figure 13: Performance comparison in MEAL environments.

Complete performance comparison on MEAL Fig. 13 shows the complete performance comparison against continual RL baselines in three levels of MEAL environments. Our method exhibits a consistent and monotonic improvement across the sequence of environments, indicating effective knowledge accumulation and higher *FT* during continual learning. Consequently, the performance gap between our method and the baselines progressively widens, highlighting its superior long-term learning capability. While MAS shows similar *FT* as our method, our HPB exhibits smaller performance decline after task switches, demonstrating its enhanced ability to retain knowledge and adapt to new tasks with minimal forgetting. This indicates that our method is better at balancing *FT* and *FG* in multi-task scenarios.

Performance comparison on partially observable MEAL We also evaluate our CHAIN on partially observable MEAL, where agents have a 2×3, 3×3, and 3×5 observation window for `Level 1`, `Level 2`, and `Level 3` respectively. Here, following existing DEC-MDP methods, we employ an RNN-based policy network to approximate the global state from multi-step observations. The results are presented in Fig. 14. When trained under partial observability, all evaluated methods exhibit performance degradation. Nevertheless, CHAIN consistently achieves superior performance across all three difficulty levels compared to the baselines, demonstrating its ability of adapting to scenarios with limited state information.

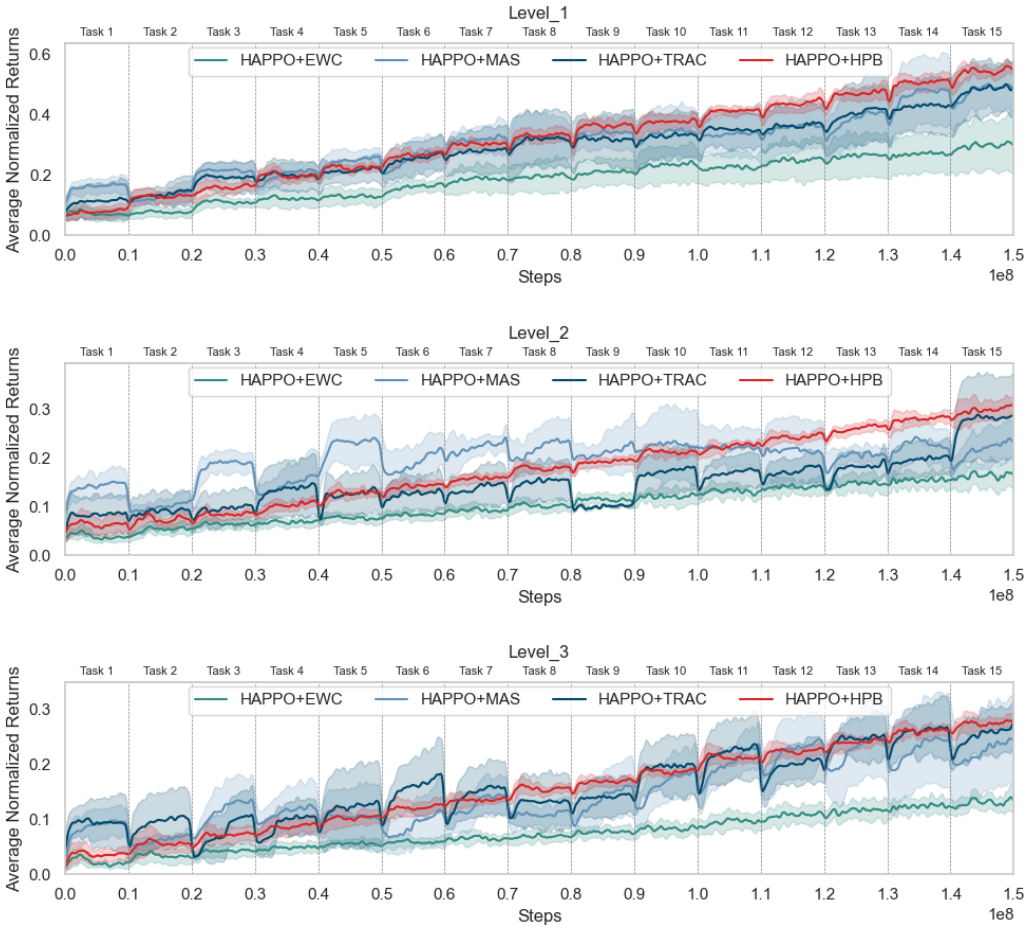


Figure 14: Performance comparison in partially observable MEAL environments.

7.7 Additional Parameter Sensitivity Study

In this part, we present the ablation study on the kernel parameter σ. Results in Fig. 8 show that CHAIN is relatively insensitive to γ and σ within a reasonable range, which is friendly to parameter tuning. Specifically, setting σ below 1e-2 yields consistently good performance.

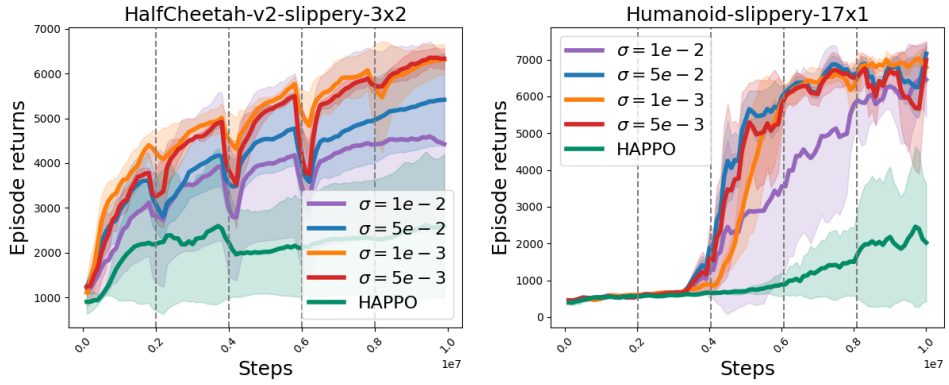


Figure 15: Parameter sensitivity of σ for CHAIN.

7.8 Updating Order Ablation Analysis

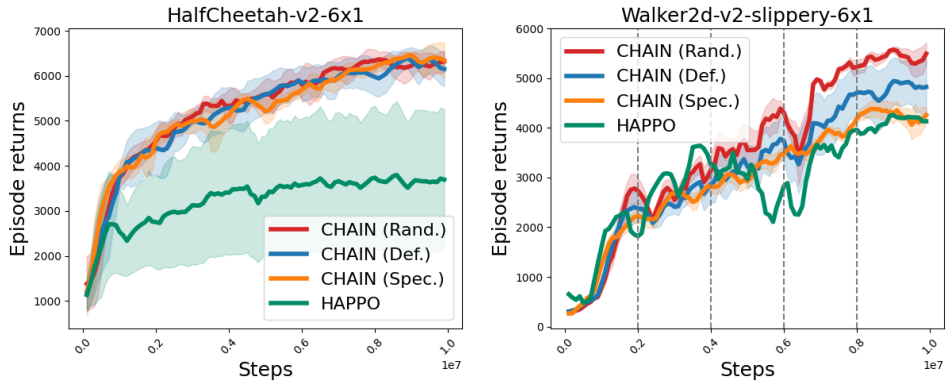


Figure 16: Updating order study for CHAIN.

Since our HPB measures the heterogeneity between the current agent and the preceding one, the updating order can influence performance. To investigate this effect, we conduct an ablation study on the updating order using standard and slippery environments of `HalfCheetah-6x1`. The results are shown in Fig. 16. We consider three updating orders, random order, default order and special order, denoted as *Rand.*, *Def.*, and *Spec.* Here, *Rand.* chooses a random order of agents. *Def.* is the default order of MA-Mujoco, which sorts agents by the part they belong to. For example, in `HalfCheetah-6x1`, *Def.* order can be right thigh joint, right leg joint, right foot joint, left thigh joint, and so on. *Spec.* order sorts agents according to their specializations, and can be right thigh joint, left thigh joint, right leg joint, left thigh joint, and so on.

From Fig. 16, we can observe that all three updating orders yield comparable performance. This is because, under our HPB framework, agents can benefit from preceding agents, regardless of whether those agents are similar to or different from themselves. Specifically, if the adjacent two agents share the same specialization, minimizing the heterogenization term of HPB encourages exploration. On the other hand, if two agents play different roles, minimizing this term enhances agent heterogeneity. Consequently, designing a specialized learning-based update order is largely unnecessary, since it may introduce extra computational cost while offering no substantial performance improvement.

7.9 Training Time Analysis

As shown in Table 5, our method introduces only marginal computational overhead compared to the strongest baseline. This is because the matrix-based Rényi entropy mutual information we use is computationally efficient. More importantly, this minor time increase is accompanied by substantial performance gains. Our approach outperforms the baseline by a large margin in both final reward and sample efficiency. These results demonstrate that our method achieves a highly favorable trade-off between efficiency and effectiveness, delivering state-of-the-art performance with minimal extra computational overhead.

Table 5: Traing time comparison between CHAIN and baselines on `HalfCheetah-6x1`

Envs	MAPPO	HAPPO	A2PO	MADPO	CHAIN
HalfCheetah-6x1-standard	3h23min	3h45min	3h51min	4h55min	4h25min
HalfCheetah-6x1-slippery	3h18min	3h59min	3h41min	4h49min	4h19min
Humanoid-17x1-standard	6h13min	6h09min	6h31min	7h13min	6h55min
Humanoid-17x1-slippery	6h11min	6h10min	6h17min	7h19min	7h01min

7.10 Additional Related Works

Information Bottleneck in RL Recently, information-theoretic principles have gained significant attention as effective tools for measuring reward-irrelevant regularization metrics Yang et al. [2026],

serving as an intrinsic incentive to accelerate exploration in reinforcement learning (RL) Ding et al. [2023a], Li et al. [2025]. Furthermore, the information bottleneck (IB) framework has been widely applied in neural networks for RL, addressing challenges such as representation efficiency and balancing information redundancy Igl et al. [2019], enabling efficient environment perception Bai et al. [2021], and facilitating multi-view observation fusion Fan and Li [2022]. In multi-agent settings, there has been growing interest in integrating IB and Graph IB to enhance communication efficiency. Notable approaches include IB-based communication at the single-message level Wang et al. [2020], systematic-level communication You et al. [2024], and graph IB-based message representation Ding et al. [2023b]. Moreover, IB has also been utilized to compress observations, refining the information that is most relevant to other agents Jin et al. [2021]. However, these methods are primarily designed for relatively stable environments with homogeneous agents and lack the adaptability needed for continual and heterogeneous settings. In this work, we focus on continual heterogeneous multi-agent cooperation and extend the traditional IB framework to Heterogeneous Policy Bottleneck (HPB), which encourages agents to reduce the representation redundancy and homogenization redundancy.

Limitations

While our method demonstrates strong empirical performance across various continual heterogeneous MARL scenarios, it introduces a kernel parameter σ in the mutual information estimation. Although our ablation studies show that the performance is relatively robust to this parameter within a reasonable range, tuning σ for optimal performance in new environments can still be challenging.

Impact Statement

This paper presents work whose goal is to advance the field of continual MARL. The primary focus of our research is to improve agent cooperation and adaptability in controlled environments. While there are numerous potential applications for these methods, including autonomous systems and collaborative robotics, we believe that the ethical implications and societal consequences of our work are not of immediate concern. The algorithms proposed here do not involve real-world deployment and are intended for research purposes in safe, simulated environments. We believe that this work, when deployed responsibly in future applications, can contribute positively to fields such as robotics and autonomous systems, with careful attention to safety and ethical guidelines.

Usage of LLMs

We used LLMs solely for language polishing and grammar correction of the manuscript text. LLMs were not involved in the design of methods, generation of experimental results, data analysis, or any core scientific contributions of this work.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims are included in the introduction section.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of this work are discussed in the Appendix section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The complete proofs of the propositions in this paper are included in the Appendix section.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The pseudocode and detailed parameter settings are included in the Appendix section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The detail pseudocode is included in the Appendix, and the source code will be released after the review.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental details are included in the Experiment section and the Appendix section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The results in the paper are reported with confidence intervals.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The computational cost and the hardware of the experiments in this paper are included in the Appendix section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The societal impacts are discussed in the Appendix section.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The assets used in this paper are correctly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper currently does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The usage of LLMs are described in the Appendix.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.