HDCF: Bi-level Diversity Coordination for Hierarchical Cooperative Multi-Agent Reinforcement Learning

Zhou Wang, Mengke Wang, Xiangfeng Luo, Shaorong Xie
School of Computer Engineering and Science, *Shanghai University*
{shushuwang,21820237,luoxf,srxie}@shu.edu.cn

***Abstract*—Hierarchical multi-agent reinforcement learning is a promising approach for complex cooperative tasks, since temporally extended macro-actions can support diverse behaviours over long horizons. However, high-level macro-action learning is often weakly grounded in behavioural outcomes, which may cause a few macro-actions to dominate while many become redundant, thereby limiting the behavioural diversity available for cooperation. To address this, we propose HDCF, a novel framework that formulates diversity learning as a bi-level optimisation process coordinating high-level macro-action regulation with low-level execution. At the high level, HDCF introduces counterfactual similarity-based shaping to regulate macro-action learning using behaviour-level signals. At the low level, HDCF incorporates an advantage-based inter-level incentive to strengthen hierarchical alignment, so that different macro-actions translate into reliably differentiated executions. Experiments on SMAC and GRF show that HDCF outperforms representative baselines, and ablation studies validate that both components contribute to the overall improvement.**

***Index Terms*—Multi-Agent Systems, Hierarchical Reinforcement Learning, Exploration Incentive, Behaviour Diversity.**

I. Introduction

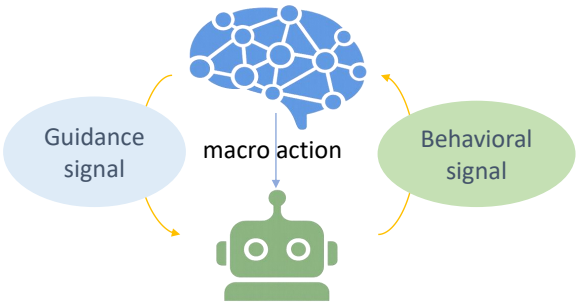
Cooperative multi-agent reinforcement learning (MARL) has become a promising framework for real-world systems in which multiple decision-makers must coordinate under partial observability and shared constraints. It has been studied in a range of coordination-intensive applications, such as multi-robot collaboration [1], traffic signal control [2] and autonomous driving interactions systems [3], where effective cooperation is critical to overall performance [4]. To improve training stability and scalability in such settings, a dominant paradigm is Centralized Training with Decentralized Execution (CTDE) [5], which leverages global information during training to mitigate non-stationarity while preserving decentralized decision-making at execution time [6]. Within CTDE, value-decomposition methods [7], [8] further address the credit assignment problem by factorizing a global value function into agent-wise components, enabling more scalable learning in cooperative settings.

Despite these advances, cooperative MARL algorithms still face significant challenges in complex tasks, particularly with respect to policy homogenization and limited long-horizon planning capability. Parameter sharing is commonly adopted to reduce model complexity, but it often drives agents toward similar behaviors [9], diminishing the diversity required for effective exploration and complementary cooperation [10]. At the same time, insufficient long-term planning capability can cause agents to prioritize short-term local objectives over global coordination, leading to suboptimal performance [11].

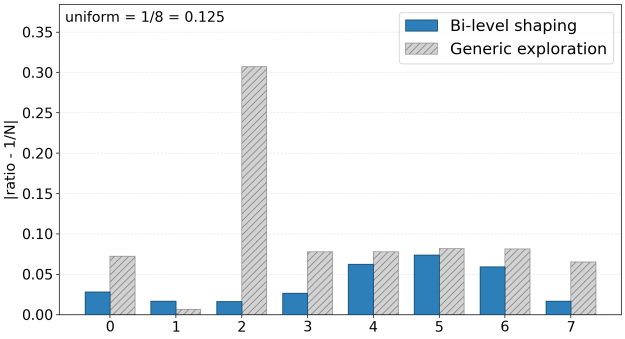
Hierarchical reinforcement learning offers a principled approach to these challenges by decomposing complex tasks into multiple levels of abstraction [12]. In this paradigm, a high-level policy selects abstract macro-actions or sub-policies [13], [14], while low-level policies execute primitive actions to realise the selected behaviours [15]. Such temporal abstraction reduces the effective complexity of long-horizon decision making and can support complementary cooperation by allowing different agents to commit to different abstract behaviours [13], [16]. Motivated by this potential, a growing body of work has focused on promoting diversity in hierarchical MARL. Existing methods introduce various mechanisms primarily at the low level, such as mutual information maximisation to encourage skills to visit distinct regions of the state space [17], [18], clustered or role-based action representations [19], [20], dynamically differentiated subtask encodings [21], or advantage-based alignment incentives to coordinate hierarchical levels [14]. These approaches can effectively shape diverse low-level behaviours and improve hierarchical controllability.

However, an important aspect of hierarchical coordination remains insufficiently addressed. The effectiveness of a macro-action set hinges on whether its macro-actions can consistently induce distinguishable low-level behaviours under the same local observation [15]. In many hierarchical methods, behavioural diversity is promoted through low-level mechanisms, for example through skill discovery or inter-level coordination [14], [17], [18], whereas high-level macro-action selection is often still driven by generic, behaviour-agnostic exploration such as uniform sampling or ϵ-greedy. This means that the high level may not be guided by an explicit objective that links macro-action choices to their behavioural consequences at the low level. Moreover, generic exploration in multi-agent settings is already known to be challenging due to non-stationarity [22], and this difficulty can interact with the intrinsic instability of hierarchical training [23], making it harder for different macro-actions to develop clearly differentiated behavioural effects.

Intuitively, when macro-actions are selected in a behaviour-agnostic manner, learning signals can concentrate on a small



(a) Bi-level coordination overview.



(b) Macro-action ratio comparison.

Fig. 1: Motivation and overview. (a) HDCF performs bi-level diversity coordination. (b) Macro-action usage on `3s_vs_5z`.

subset of macro-actions that are easier to differentiate or happen to yield higher returns early. In contrast, macro-actions whose behavioural differences have not yet emerged can remain data-scarce and difficult to learn, which can slow down their behavioural formation and reduce their eventual contribution. A common outcome is that low-level behaviours under a few macro-actions become increasingly clear, while behaviours induced by many other macro-actions remain similar and receive weaker learning signals, thereby limiting the behavioural diversity that can be exploited for cooperation. Fig. 1 provides an illustrative example, where generic high-level exploration is associated with a strongly skewed macro-action usage pattern (Fig. 1b), motivating an explicit bi-level mechanism that coordinates high-level selection with low-level behavioural differentiation (Fig. 1a). Meanwhile, simply increasing exploration randomness may be insufficient, because it does not allocate learning effort according to whether a macro-action still induces behaviours similar to other macro-actions under the current local observation.

Our principal contributions are as follows:

(i) We introduce a counterfactual similarity based high-level shaping objective for bi-level diversity coordination in hierarchical cooperative MARL, encouraging an effective and diverse macro-action set during learning.

(ii) We propose **HDCF**, a bi-level diversity coordination framework that explicitly couples high-level shaping with a low-level inter-level incentive in a unified training formulation, enabling diversity regulation at the decision level and reliable realisation at the execution level.

(iii) We provide a comprehensive empirical evaluation of HDCF on SMAC [24] and GRF [25], showing consistent improvements over strong baselines, and conduct ablations to isolate the contributions of each proposed component.

II. Related Work

A. Value-Function Decomposition in MARL

Value-decomposition methods constitute a core component of the CTDE paradigm in cooperative MARL. Value Decomposition Networks (VDN) [7] first decomposed the joint action-value function into a sum of per-agent utilities, enabling scalable learning with decentralized execution. QMIX [8] further introduced a mixing network that constrains the joint action-value to be a monotonic function of individual utilities, preserving consistency between individual and global action selections while improving representational capacity. QTRAN [26] reformulated value factorisation to guarantee optimal joint action selection without monotonicity constraints, but can be challenging to scale to complex environments.

Subsequent studies have revisited and extended QMIX's assumptions. QPLEX [27] adopts a dueling-value factorisation to retain the practical benefits of monotonic mixing while improving expressiveness for more general value functions. WQMIX [28] improves learning stability by re-weighting the Bellman error projection, especially useful in settings where the data distribution is limited or biased. These methods collectively advance the theoretical and practical foundations of value decomposition. Our framework inherits a QMIX-style backbone [8] and extends it to a hierarchical cooperative setting, enabling the integration of additional objectives required for coordinated bi-level learning.

B. Hierarchical MARL

Hierarchical reinforcement learning has been widely adopted in multi-agent systems to address long-horizon coordination via temporal abstraction and multi-level decision making [29]. Beyond improving planning tractability, a key motivation in cooperative settings is to mitigate policy homogenization by enabling agents to acquire diverse and complementary behaviours at different levels of abstraction. Early work such as H-QMIX [30] integrates hierarchical control with cooperative MARL and shows advantages in temporally extended tasks, but it often relies on manually designed high-level behaviours or pre-defined abstractions, which may limit adaptability in complex environments.

A broad body of subsequent research explores how to learn and maintain behavioural diversity within hierarchical MARL. ROMA [31], RODE [19], and SIRD [32] investigate mechanisms for discovering differentiated coordination patterns that support complementary cooperation. HSD [17] and HSL [16] promote distinguishable behaviours through intrinsic objectives that encourage diversity. Other works, such

as LDSA [21], DT2GS [33], MASER [34], and HiSOMA [35], introduce task decomposition or subgoal-oriented hierarchies to improve long-horizon performance, thereby providing distinct behavioural modes for agents to exploit. More recently, HAVEN [14] and DCC [18] advance hierarchical coordination by strengthening inter-level consistency and cooperation among learned behaviours, achieving strong empirical performance on cooperative benchmarks.

While these hierarchical methods can enrich low-level behaviours and improve inter-level coordination, high-level exploration is typically handled by generic schemes without an explicit objective to encourage adequate and balanced coverage of the macro-action space. In contrast, our approach treats diversity as a coordinated bi-level optimisation problem.

C. Behaviour Similarity Estimation Using KL Divergence

KL divergence is a common tool in reinforcement learning for comparing policy behaviours. It has been used for inter-agent knowledge distillation, enabling weaker agents to learn from stronger ones [36], and for encouraging exploratory diversity by differentiating current behaviours from historical policies [37]. In hierarchical and multi-agent settings, KL-based measures have also been used to construct influence-related rewards that promote behavioural distinction across decision levels [38], and to quantify inter-agent influence under counterfactual formulations [39].

These studies show that KL divergence provides a practical signal for shaping behavioural dynamics in multi-agent systems. Motivated by this, we use KL divergence to quantify behavioural dissimilarity between macro-actions.

III. Preliminary

A. Decentralized Partially Observable Markov Decision Process (Dec-POMDP)

We model cooperative multi-agent interaction as a Dec-POMDP $G = \langle S, \{\mathcal{A}^i\}_{i=1}^N, P, R, \{\Omega^i\}_{i=1}^N, O, \gamma\rangle$, where N is the number of agents. At time t, the environment is in a global state $s_t \in S$ and agents choose a joint action $\mathbf{a}_t = (a_t^1, \dots, a_t^N)$ with $a_t^i \in \mathcal{A}^i$. The next state is sampled from $P(s_{t+1} \mid s_t, \mathbf{a}_t)$ and all agents receive a shared reward $r_t^e = R(s_t, \mathbf{a}_t)$. Each agent i receives a local observation $o_t^i \in \Omega^i$ generated by the observation function $O(\mathbf{o}_t \mid s_t)$, where $\mathbf{o}_t = (o_t^1, \dots, o_t^N)$. Under decentralized execution, agent i conditions its decision on its information state (action-observation history or recurrent hidden state) τ_t^i.

The objective is to learn a joint policy $\boldsymbol{\pi} = (\pi^1, \dots, \pi^N)$ that maximizes $J(\boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\pi}}\big[\sum_{t=0}^{\infty} \gamma^t r_t^e\big]$. Under CTDE, global information (e.g., s_t) may be used for training value functions, while each agent acts based only on τ_t^i at test time.

B. Fixed-duration Hierarchical MARL

We consider a two-level hierarchy with a fixed temporal abstraction factor K. High-level (macro) decisions are indexed by T, and each macro interval corresponds to environment steps $t \in [TK, (T+1)K-1]$. Each agent selects a macro-action $z_T^i \in \mathcal{Z}$ from a discrete set $\mathcal{Z}$ (with $|\mathcal{Z}| = Z$), forming a joint macro-action $\mathbf{z}_T = (z_T^1, \dots, z_T^N)$. The selected macro-actions remain fixed within the macro interval.

We denote the high-level and low-level information states separately as $\tau_T^{h,i}$ and $\tau_t^{l,i}$, respectively. The hierarchical policy consists of a high-level policy $\pi_h^i(z_T^i \mid \tau_T^{h,i})$ and a low-level policy $\pi_l^i(a_t^i \mid \tau_t^{l,i}, z_T^i)$.

IV. Method

This section details the proposed Hierarchical Diversity Coordination Framework (HDCF). We first describe the overall architectural design, then introduce the core component, and finally present the training objectives.

A. Framework overview

HDCF employs a two-level hierarchical architecture with a fixed temporal abstraction factor K, as shown in Fig. 2. Both levels adopt value-function decomposition implemented via QMIX-style mixing networks: each level maintains per-agent utility estimates which are combined by a parameterised mixing network to produce a level-specific global action-value. This supports centralized training while preserving decentralized execution.

At the high level (macro time scale), decisions are indexed by T, and each macro interval corresponds to environment steps $t \in [TK, (T+1)K-1]$. At the beginning of each macro interval, each agent $i \in \{1, \dots, N\}$ selects a macro-action

$$z_T^i \in \mathcal{Z}, \qquad Z \triangleq |\mathcal{Z}|, \tag{1}$$

forming a joint macro-action $\mathbf{z}_T = (z_T^1, \dots, z_T^N)$, which remains fixed within the interval. The cumulative extrinsic reward collected over this macro interval is

$$R_T = \sum_{t=TK}^{(T+1)K-1} r_t^e, \tag{2}$$

where r_t^e denotes the environment reward at timestep t.

At the low level (environment time scale), each agent issues primitive actions at every timestep conditioned on its low-level information state and the selected macro-action

$$a_t^i \sim \pi_1^i\big(a_t^i \mid \tau_t^{l,i}, z_T^i\big), \qquad t \in [TK, (T+1)K-1]. \tag{3}$$

The key distinction of HDCF is a *bi-level diversity coordination* objective that explicitly couples high-level macro-action learning with low-level behavioural differentiation. At the high level, HDCF introduces a behaviour-grounded intrinsic shaping term to regulate macro-action optimisation, mitigating behaviour-level degeneracy where many macro-actions remain redundant. At the low level, HDCF employs an inter-level incentive to strengthen hierarchical alignment, ensuring that different macro-actions translate into reliably differentiated execution.

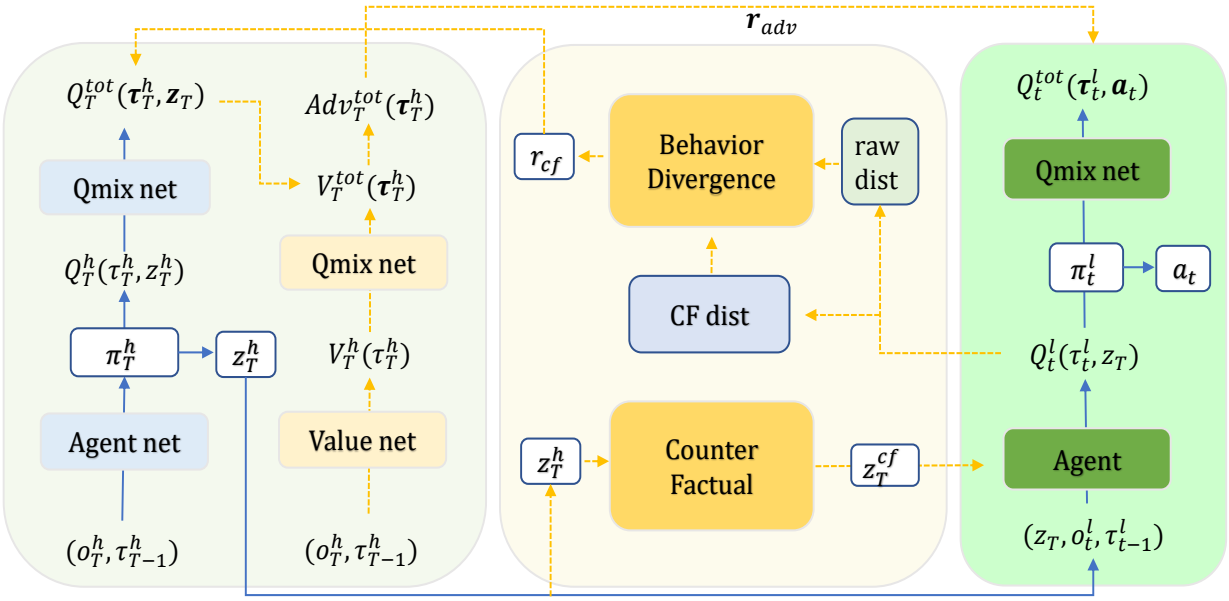


Fig. 2: The framework of the proposed method.

B. Counterfactual policy-similarity shaping

To enable *bi-level diversity coordination*, we augment the high-level objective with a behaviour-grounded intrinsic shaping term based on *counterfactual policy similarity*. This term provides a state-conditional signal for macro-action optimisation by measuring how behaviourally separated the selected macro-actions are relative to their counterfactual alternatives under the current context. Our incentive is intentionally inverse to standard novelty bonuses: we penalise macro-actions that are already strongly differentiated, thereby shifting optimisation pressure towards macro-actions that remain under-differentiated and improving the effective diversity of the macro-action set during learning. The resulting high-level regulation is complementary to the advantage-based inter-level incentive, which helps translate different macro-actions into reliably differentiated low-level execution.

1) Macro-action similarity metric: We measure behavioural distinctiveness at the beginning of each macro step, i.e., at $t = TK$. For agent i, we compare the low-level action distribution induced by the selected macro-action z_T^i with the distributions that would be induced by alternative counterfactual macro-actions $z \in \mathcal{Z} \setminus \{z_T^i\}$ under the same low-level information state $\tau_{TK}^{l,i}$.

Let $Q(\cdot \mid \tau_{TK}^{l,i}, z)$ denote the low-level action-value outputs over primitive actions $a \in \mathcal{A}^i$ produced when conditioning on macro-action z and $\tau_{TK}^{l,i}$. We convert these outputs to a probability distribution via softmax

$$\pi(\cdot \mid \tau_{TK}^{l,i}, z) \;=\; \mathrm{Softmax}\big(Q(\cdot \mid \tau_{TK}^{l,i}, z)\big). \tag{4}$$

When computing the KL-based intrinsic term, we stop gradients through the computation of $\pi(\cdot \mid \tau_{TK}^{l,i}, z)$.

Conditioned on $\tau_{TK}^{l,i}$, the distinctiveness of z_T^i is defined as

$$D_{z_T^i}(\tau_{TK}^{l,i}) \;\triangleq\; \sum_{\substack{z \in \mathcal{Z} \\ z \neq z_T^i}} D_{\mathrm{KL}}\Big(\pi(\cdot \mid \tau_{TK}^{l,i}, z_T^i) \,\|\, \pi(\cdot \mid \tau_{TK}^{l,i}, z)\Big). \tag{5}$$

Normalising the sum in Eq. (5) by $|\mathcal{Z}| - 1$ is generally preferable to keep the intrinsic term on a consistent scale. We use the summed form in Eq. (5) because the macro-action set size $|\mathcal{Z}|$ is fixed in our experimental setup, so the averaged version would only rescale r_T^{cf} by a constant factor and does not affect relative comparisons; this scaling can be absorbed into β.

2) Inverse incentive design: We form a macro-step intrinsic term by averaging distinctiveness across agents

$$r_T^{\mathrm{cf}} \;=\; \frac{1}{N} \sum_{i=1}^{N} D_{z_T^i}(\tau_{TK}^{l,i}), \tag{6}$$

and incorporate it into a shaped high-level return

$$R_{\mathrm{total}} \;=\; R_T \;-\; \beta \cdot r_T^{\mathrm{cf}}, \tag{7}$$

where β is a scaling coefficient. The negative sign penalises macro-actions that are already strongly differentiated under the current context, thereby shifting optimisation pressure towards macro-actions that remain under-differentiated. In this way, the shaping term acts as a behaviour-grounded regulariser for

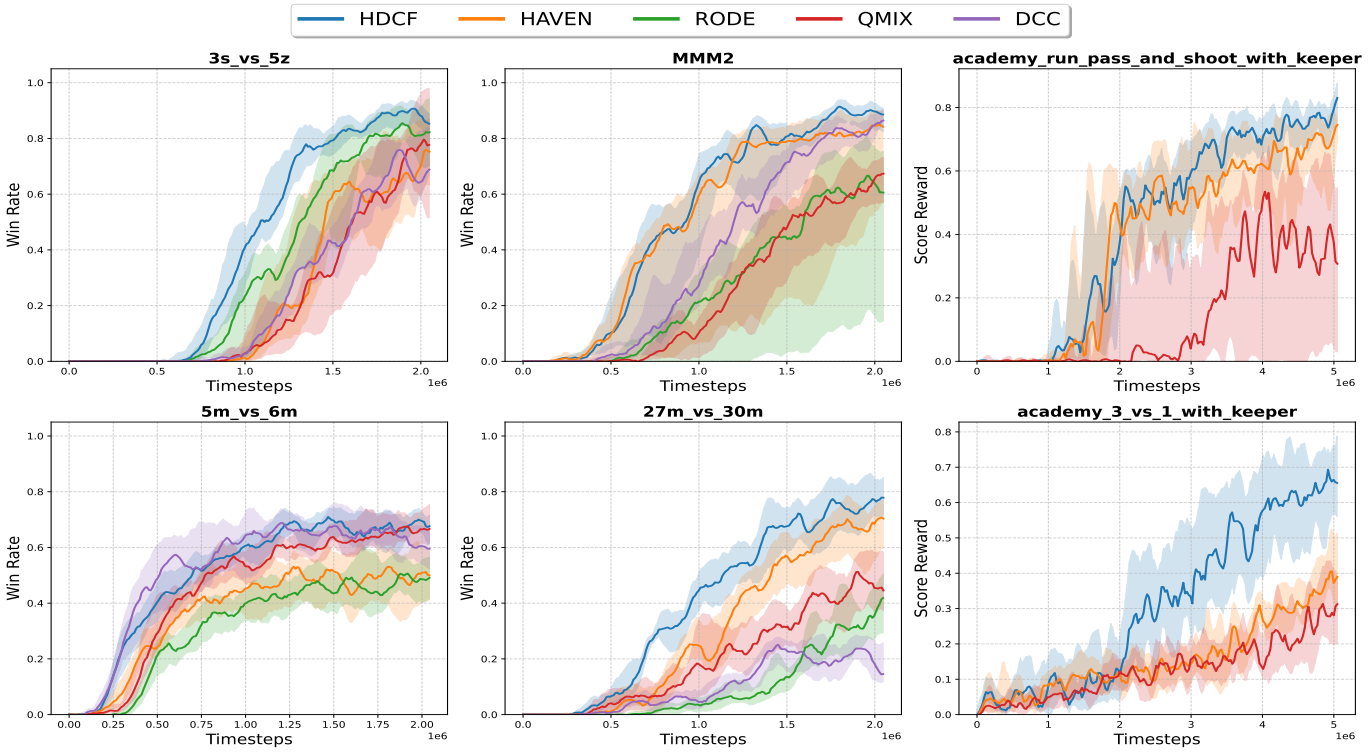


Fig. 3: Overall performance comparison across benchmarks.

the high-level objective, supporting bi-level diversity coordination by discouraging behavioural redundancy among macro-actions.

C. Advantage-Incentive Based Control Enhancement

To strengthen inter-level coordination and ensure that high-level macro-action choices exert significant influence on low-level behaviour, we synchronise the optimisation objectives across levels. Inspired by HAVEN [14], we train an additional high-level state-value network V^{h} to compute an advantage signal and distribute it equally to the low-level policies as an inter-level coordination reward.

Let $V^{\mathrm{h}}(\tau_T^{\mathrm{h}})$ denote the high-level state-value, and let $V^{\mathrm{h}}(\tau_{T+1}^{\mathrm{h}})$ be the value at the next macro step. The high-level advantage is defined as the temporal-difference error

$$A_T \;=\; R_T \;+\; \gamma^{\mathrm{h}}\, V^{\mathrm{h}}(\tau_{T+1}^{\mathrm{h}}) \;-\; V^{\mathrm{h}}(\tau_T^{\mathrm{h}}), \tag{8}$$

where γ^{h} is the high-level discount factor. We then augment the low-level reward by distributing this advantage uniformly across the K steps in macro interval T

$$r_{\mathrm{total}} \;=\; r_t^e \;+\; \lambda \cdot \frac{A_T}{K}, \qquad t \in [TK, (T+1)K-1], \tag{9}$$

where λ controls the strength of inter-level alignment.

D. Training Objectives

Based on the above design, the training objective of HDCF comprises three core components.

1) High-level state-value loss:

$$\mathcal{L}_V^{\mathrm{h}}(\theta) = \mathbb{E}\Big[\big(R_T + \gamma \max_{\mathbf{z}_{T+1}} Q_{\mathrm{tot}}^{\mathrm{h}}(\tau_{T+1}^{\mathrm{h}}, \mathbf{z}_{T+1}) - V^{\mathrm{h}}(\tau_T^{\mathrm{h}} \mid \theta)\big)^2\Big]. \tag{10}$$

2) High-level action-value loss:

$$\mathcal{L}_Q^{\mathrm{h}}(\varphi) = \mathbb{E}\Big[\Big(R_{\mathrm{total}} + \gamma \max_{\mathbf{z}_{T+1}} Q_{\mathrm{tot}}^{\mathrm{h}}(\tau_{T+1}^{\mathrm{h}}, \mathbf{z}_{T+1} \mid \varphi^-) - Q_{\mathrm{tot}}^{\mathrm{h}}(\tau_T^{\mathrm{h}}, \mathbf{z}_T \mid \varphi)\Big)^2\Big]. \tag{11}$$

3) Low-level action-value loss:

$$\mathcal{L}_Q^{\mathrm{l}}(\psi) = \mathbb{E}\Big[\Big(r_{\mathrm{total}} + \gamma \max_{\mathbf{a}_{t+1}} Q_{\mathrm{tot}}^{\mathrm{l}}(\tau_{t+1}^{\mathrm{l}}, \mathbf{z}_{\lfloor (t+1)/K \rfloor}, \mathbf{a}_{t+1} \mid \psi^-) - Q_{\mathrm{tot}}^{\mathrm{l}}(\tau_t^{\mathrm{l}}, \mathbf{z}_{\lfloor t/K \rfloor}, \mathbf{a}_t \mid \psi)\Big)^2\Big]. \tag{12}$$

Here φ^- and ψ^- denote the parameters of the corresponding target networks.

V. Experiment

In this section, we evaluate the performance of the proposed algorithm on two widely used multi-agent reinforcement learning benchmarks: the StarCraft Multi-Agent Challenge (SMAC) and Google Research Football (GRF). To assess the practical utility of the counterfactual behavioural-similarity incentive, we perform ablation studies on SMAC and examine the effect of varying the weight of this intrinsic incentive.

TABLE I: Mean of the median curve over the last 10% timesteps ± SD across seeds

Maps	Metric	HDCF	HAVEN	RODE	QMIX	DCC
3s_vs_5z	**Win**	0.882 ± 0.106	0.698 ± 0.207	0.834 ± 0.250	0.744 ± 0.194	0.689 ± 0.122
MMM2	**Win**	0.896 ± 0.197	0.833 ± 0.291	0.627 ± 0.234	0.637 ± 0.214	0.826 ± 0.135
5m_vs_6m	**Win**	0.678 ± 0.100	0.494 ± 0.136	0.490 ± 0.125	0.660 ± 0.099	0.623 ± 0.133
27m_vs_30m	**Win**	0.750 ± 0.116	0.682 ± 0.152	0.344 ± 0.128	0.479 ± 0.133	0.207 ± 0.080
GRF/2v1	**Score**	0.764 ± 0.193	0.702 ± 0.280	—	0.337 ± 0.250	—
GRF/3v1	**Score**	0.656 ± 0.260	0.340 ± 0.115	—	0.283 ± 0.123	—

A. Experimental Setup

The SMAC environment is a suite of cooperative micromanagement tasks built on StarCraft II. It offers diverse maps for testing cooperative behaviour and long-horizon control. We evaluate our method on two Hard maps (`5m_vs_6m` and `3s_vs_5z`) and two Super Hard maps (`MMM2` and `27m_vs_30m`). GRF is an open-source environment in a simulated soccer domain. We include GRF to verify the generality of our approach across domains with different dynamics and cooperation patterns.

We compare our method to several representative baselines in six maps, including the general value-decomposition method QMIX [8] and hierarchical methods RODE [19], HAVEN [14], and DCC [18]. Since RODE and DCC do not report hyperparameter configurations for GRF, comparisons on GRF focus on QMIX and HAVEN. Across all maps, we use 8 macro-actions and set k to 3 for our method, follow the standard QMIX training setup [8] for hyperparameters and optimization with the RMSProp optimizer for all methods.

For each map, we run 5 independent trials with five random seeds, using different seed sets for different maps; within each map, all methods are evaluated using the same set of five seeds for a fair comparison.

B. Performance Evaluation

Fig. 3 compares learning curves of HDCF and baselines on SMAC and GRF, where solid lines denote median performance and shaded regions indicate the 25–75% percentiles. To complement the curves with a concise summary, Table I reports final performance and stability statistics. Across most maps, HDCF attains the best or competitive best final performance and is typically associated with lower variance, indicating improved learning stability.

As summarized in Table I, HDCF achieves consistently strong results across the evaluated SMAC tasks. The learning curves in Fig. 3 further indicate that HDCF maintains steady progress on challenging maps compared with other hierarchical baselines, reflecting more reliable optimization dynamics. In particular, HDCF shows better robustness on larger-scale maps where coordination becomes harder, suggesting that its hierarchical control remains effective as the joint decision space grows. Additionally, HDCF consistently achieves stronger final performance than the baselines on both GRF tasks.

C. Ablation Study

We compare HDCF with three variants: **HDCF_CF**, **HDCF_WB**, and **HDCF_WA**. HDCF_CF changes the similarity-penalty coefficient β (0.1 and 1.0) to test sensitivity to shaping strength. HDCF_WB computes the counterfactual similarity term r_T^{cf} for logging (to plot Fig. 4) but removes it from the high-level objective, isolating the effect of high-level regulation. HDCF_WA removes the advantage-based incentive to evaluate the role of inter-level alignment in realising behaviourally differentiated execution. In our main setting, β is set to 0.1 for `3s_vs_5z` and `MMM2`, and 1.0 for `5m_vs_6m` and `27m_vs_30m`. Results are reported in Fig. 4 and Table II.

1) Influence of the shaping strength: HDCF_CF evaluates sensitivity to the shaping strength β, which controls how strongly the high-level objective regularises macro-action learning for diversity coordination. On three maps (`5m_vs_6m`, `3s_vs_5z`, and `27m_vs_30m`), HDCF_CF achieves performance close to HDCF, suggesting that the high-level shaping is reasonably robust within the tested range. In contrast, on the heterogeneous map `MMM2`, HDCF_CF performs worse, indicating that heterogeneous settings can be more sensitive to the calibration of β. A plausible reason is the scale variability of the KL-based similarity term: larger discrepancies between low-level action distributions can amplify the effective shaping magnitude and increase sensitivity to β, which is more likely to arise in heterogeneous settings.

2) Effect of removing the similarity-based shaping: HDCF_WB removes r_T^{cf} from the high-level return to isolate and examine the contribution of the high-level regulation mechanism. To quantify its effect on the learned macro-action set, we collect evaluation trajectories using the final checkpoint, and compute the normalised entropy of the macro-action usage distribution aggregated over all agents during evaluation. Specifically, we evaluate 32 episodes per seed and report the mean normalised entropy averaged over 5 seeds. We additionally report MaxR, defined as the maximum macro-action usage ratio among all macro-actions, where a higher MaxR indicates stronger mode collapse.

Compared with HDCF, HDCF_WB attains consistently lower win rates, suggesting that diversity also needs to be explicitly coordinated at the macro-action level. As shown in Table II, HDCF yields higher normalised macro-action usage entropy than HDCF_WB across the evaluated maps, indicating that removing the high-level shaping term tends to induce a

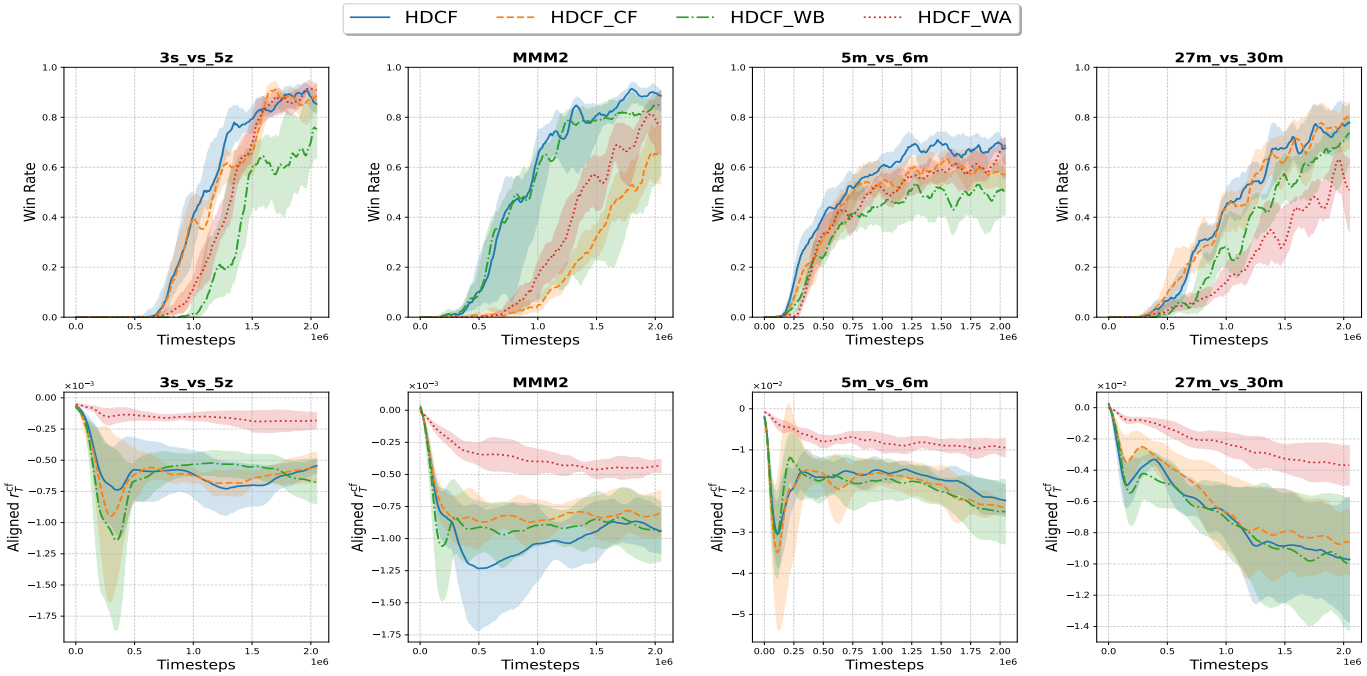


Fig. 4: Ablation study across four SMAC maps.

TABLE II: Macro-action usage diversity on SMAC.

Map	Ent (HDCF)	Ent (HDCF_WB)	ΔEnt	ΔEnt / HDCF_WB	MaxR (HDCF)	MaxR (HDCF_WB)
3s_vs_5z	0.932779	0.888508	0.044271	+4.98%	0.243669	0.296411
5m_vs_6m	0.964238	0.939514	0.024724	+2.63%	0.213560	0.242710
MMM2	0.947705	0.947293	0.000412	+0.04%	0.229874	0.223483
27m_vs_30m	0.928838	0.916743	0.012096	+1.32%	0.245940	0.268786

more skewed macro-action usage distribution. Such skew can reduce the visitation frequency of some macro-actions and is consistent with the larger performance drop of HDCF_WB in small-scale homogeneous scenarios.

In contrast, the high-level shaping term in HDCF regularises the macro-action policy and mitigates excessive skew, improving macro-action coverage and enabling more macro-actions to be effectively utilised, thereby delivering more evident performance improvements. While macro-action usage skew is generally milder in large-scale scenarios, the shaping term still provides observable gains, indicating that its regularisation remains effective under more complex coordination conditions.

3) Effect of removing the advantage incentive: HDCF_WA removes the advantage-based inter-level incentive and leads to clear performance degradation. Moreover, the r_T^{cf} curves in Fig. 4 are substantially lower than those of other variants, indicating weaker behavioural differentiation among macro-actions at the macro-step start. This highlights that without inter-level alignment, macro-actions become less controllable at the low level, making the intended behaviors difficult to realize reliably.

VI. Conclusion

This paper proposes HDCF, a hierarchical cooperative multi-agent reinforcement learning framework that formulates diversity learning as a bi-level optimisation problem and coordinates high-level macro-action selection with low-level execution. HDCF introduces counterfactual behavioural-similarity shaping as an explicit high-level objective to regulate macro-action learning and mitigate behaviour-level degeneracy, and adopts an advantage-based inter-level incentive to strengthen hierarchical alignment. Experimental results on SMAC and GRF indicate that combining behaviour-grounded high-level regulation with inter-level alignment can improve the effectiveness and robustness of hierarchical cooperative MARL. Future work will investigate heterogeneity-aware extensions to better adapt HDCF to heterogeneous scenarios, along with adaptive tuning of the shaping strength and more efficient similarity estimation to further improve scalability and robustness.

Acknowledgment

This work was supported by the National Natural Science Foundation of China Innovation Research Group Project under Grant No. 62421004.

REFERENCES

[1] A. H. Tan, F. P. Bejarano, Y. Zhu, R. Ren, and G. Nejat, "Deep reinforcement learning for decentralized multi-robot exploration with macro actions," *IEEE Robotics and Automation Letters*, vol. 8, no. 1, pp. 272–279, 2022.

[2] Y. Bie, Y. Ji, and D. Ma, "Multi-agent deep reinforcement learning collaborative traffic signal control method considering intersection heterogeneity," *Transportation Research Part C: Emerging Technologies*, vol. 164, p. 104663, 2024.

[3] 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*. PMLR, 2021, pp. 264–285.

[4] H. Chen, B. Zhang, and G. Fan, "Sgcd: Subgroup contribution decomposition for multi-agent reinforcement learning," in *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024, pp. 1–8.

[5] Z. Ning and L. Xie, "A survey on multi-agent reinforcement learning and its application," *Journal of Automation and Intelligence*, vol. 3, no. 2, pp. 73–91, 2024.

[6] P. K. Sharma, R. Fernandez, E. Zaroukian, M. Dorothy, A. Basak, and D. E. Asher, "Survey of recent multi-agent reinforcement learning algorithms utilizing centralized training," in *Artificial intelligence and machine learning for multi-domain operations applications III*, vol. 11746. SPIE, 2021, pp. 665–676.

[7] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls *et al.*, "Value-decomposition networks for cooperative multi-agent learning based on team reward," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 2018, pp. 2085–2087.

[8] T. Rashid, M. Samvelyan, C. S. De Witt, G. Farquhar, J. Foerster, and S. Whiteson, "Monotonic value function factorisation for deep multi-agent reinforcement learning," *Journal of Machine Learning Research*, vol. 21, no. 178, pp. 1–51, 2020.

[9] C. Li, T. Wang, C. Wu, Q. Zhao, J. Yang, and C. Zhang, "Celebrating diversity in shared multi-agent reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3991–4002, 2021.

[10] T. Li, C. Bai, K. Xu, C. Chu, P. Zhu, and Z. Wang, "Skill matters: Dynamic skill learning for multi-agent cooperative reinforcement learning," *Neural Networks*, vol. 181, p. 106852, 2025.

[11] L. Yuan, Z. Zhang, L. Li, C. Guan, and Y. Yu, "A survey of progress on cooperative multi-agent reinforcement learning in open environment," *arXiv preprint arXiv:2312.01058*, 2023.

[12] M. Ghavamzadeh, S. Mahadevan, and R. Makar, "Hierarchical multi-agent reinforcement learning," *Autonomous Agents and Multi-Agent Systems*, vol. 13, no. 2, pp. 197–229, 2006.

[13] M. Yang, Y. Yang, Z. Lu, W. Zhou, and H. Li, "Hierarchical multi-agent skill discovery," *Advances in Neural Information Processing Systems*, vol. 36, pp. 61 759–61 776, 2023.

[14] Z. Xu, Y. Bai, B. Zhang, D. Li, and G. Fan, "Haven: Hierarchical cooperative multi-agent reinforcement learning with dual coordination mechanism," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 10, pp. 11 735–11 743, Jun. 2023. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/26386

[15] P.-L. Bacon, J. Harb, and D. Precup, "The option-critic architecture," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.

[16] Y. Liu, Y. Li, X. Xu, Y. Dou, and D. Liu, "Heterogeneous skill learning for multi-agent tasks," *Advances in neural information processing systems*, vol. 35, pp. 37 011–37 023, 2022.

[17] J. Yang, I. Borovikov, and H. Zha, "Hierarchical cooperative multi-agent reinforcement learning with skill discovery," in *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, 2020, pp. 1566–1574.

[18] C. Li, S. Dong, S. Yang, Y. Hu, W. Li, and Y. Gao, "Coordinating multi-agent reinforcement learning via dual collaborative constraints," *Neural Networks*, vol. 182, p. 106858, 2025.

[19] T. Wang, T. Gupta, A. Mahajan, B. Peng, S. Whiteson, and C. Zhang, "Rode: Learning roles to decompose multi-agent tasks," *arXiv preprint arXiv:2010.01523*, 2020.

[20] L. Yang, S. Xie, and X. Luo, "Harft: Role and function-based cooperation approach for heterogeneous multi-agent," in *2025 International Joint Conference on Neural Networks (IJCNN)*, 2025, pp. 1–7.

[21] M. Yang, J. Zhao, X. Hu, W. Zhou, J. Zhu, and H. Li, "Ldsa: Learning dynamic subtask assignment in cooperative multi-agent reinforcement learning," *Advances in neural information processing systems*, vol. 35, pp. 1698–1710, 2022.

[22] A. Mahajan, T. Rashid, M. Samvelyan, and S. Whiteson, "Maven: Multi-agent variational exploration," *Advances in neural information processing systems*, vol. 32, 2019.

[23] W. Hu, H. Wang, M. He, and N. Wang, "Uncertainty-aware hierarchical reinforcement learning for long-horizon tasks," *Applied Intelligence*, vol. 53, no. 23, pp. 28 555–28 569, 2023.

[24] M. Samvelyan, T. Rashid, C. S. De Witt, G. Farquhar, N. Nardelli, T. G. Rudner, C.-M. Hung, P. H. Torr, J. Foerster, and S. Whiteson, "The starcraft multi-agent challenge," *arXiv preprint arXiv:1902.04043*, 2019.

[25] K. Kurach, A. Raichuk, P. Stanczyk, M. Zajac, O. Bachem, L. Espeholt, C. Riquelme, D. Vincent, M. Michalski, O. Bousquet *et al.*, "Google research football: A novel reinforcement learning environment," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 04, 2020, pp. 4501–4510.

[26] K. Son, D. Kim, W. J. Kang, D. E. Hostallero, and Y. Yi, "Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning," in *International conference on machine learning*. PMLR, 2019, pp. 5887–5896.

[27] J. Wang, Z. Ren, T. Liu, Y. Yu, and C. Zhang, "Qplex: Duplex dueling multi-agent q-learning," *arXiv preprint arXiv:2008.01062*, 2020.

[28] T. Rashid, G. Farquhar, B. Peng, and S. Whiteson, "Weighted qmix: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning," *Advances in neural information processing systems*, vol. 33, pp. 10 199–10 210, 2020.

[29] S. Pateria, B. Subagdja, A.-h. Tan, and C. Quek, "Hierarchical reinforcement learning: A comprehensive survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 5, pp. 1–35, 2021.

[30] H. Tang, J. Hao, T. Lv, Y. Chen, Z. Zhang, H. Jia, C. Ren, Y. Zheng, Z. Meng, C. Fan *et al.*, "Hierarchical deep multiagent reinforcement learning with temporal abstraction," *arXiv preprint arXiv:1809.09332*, 2018.

[31] T. Wang, H. Dong, V. Lesser, and C. Zhang, "Roma: Multi-agent reinforcement learning with emergent roles," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9876–9886.

[32] X. Zeng, H. Peng, and A. Li, "Effective and stable role-based multi-agent collaboration by structural information principles," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 10, 2023, pp. 11 772–11 780.

[33] Z. Tian, R. Chen, X. Hu, L. Li, R. Zhang, F. Wu, S. Peng, J. Guo, Z. Du, Q. Guo *et al.*, "Decompose a task into generalizable subtasks in multi-agent reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 78 514–78 532, 2023.

[34] J. Jeon, W. Kim, W. Jung, and Y. Sung, "Maser: Multi-agent reinforcement learning with subgoals generated from experience replay buffer," in *International conference on machine learning*. PMLR, 2022, pp. 10 041–10 052.

[35] M. Geng, S. Pateria, B. Subagdja, and A.-H. Tan, "Hisoma: A hierarchical multi-agent model integrating self-organizing neural networks with multi-agent deep reinforcement learning," *Expert Systems with Applications*, vol. 252, p. 124117, 2024.

[36] Z. W. Hong, P. Nagarajan, and G. Maeda, "Swarm-inspired reinforcement learning via collaborative inter-agent knowledge distillation," 2019.

[37] Z. W. Hong, T. Y. Shann, S. Y. Su *et al.*, "Diversity-driven exploration strategy for deep reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[38] Y. Loo, C. Gong, and M. Meghjani, "A hierarchical approach to population training for human-ai collaboration," 2023.

[39] N. Jaques, A. Lazaridou, E. Hughes *et al.*, "Social influence as intrinsic motivation for multi-agent deep reinforcement learning," in *International Conference on Machine Learning*, 2019, pp. 3040–3049.