BEAD: BLIND-SPOT ENTITY-AWARE ADVANTAGE DISTILLATION FOR COOPERATIVE MULTI-AGENT REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Centralized Training with Decentralized Execution (CTDE) is widely adopted in cooperative Multi-agent Reinforcement Learning. Building on CTDE, Centralized Teacher with Decentralized Student (CTDS) exploits global information to construct teachers, but indiscriminate use of such information may degrade teacher decision quality. We propose **Blind-spot Entity-aware Advantage Distillation (BEAD)**, a centralized-teacher decentralized-student framework that selectively exploits global information for effective knowledge transfer. BEAD represents the global state as entities and employs gated blind-spot compensation to selectively incorporate locally unobserved entities into the privileged teacher. In addition, BEAD introduces explicit advantage distillation anchored to the student's value estimates, transferring the teacher's relative action preferences while preserving the student's own value learning. Theoretical analysis shows that BEAD preserves action-selection preferences, while experiments on challenging multi-agent reinforcement learning benchmarks demonstrate consistent improvements over strong CTDE and teacher-student baselines.

1 INTRODUCTION

Cooperative Multi-Agent Reinforcement Learning (MARL) provides a general framework for sequential decision-making problems in which multiple agents must coordinate their behaviors to optimize a shared objective in dynamic environments. It has been applied to a wide range of domains, including autonomous driving (Zhou et al., 2020), coordinated robot control (Huang et al., 2020), logistics management (Ngu et al., 2022) and learning-based structure discovery (Xue et al., 2026). A central challenge in these applications is partial observability. Limited sensing ranges, communication bandwidth, and system latency often restrict agents to local observations, making it difficult for them to fully capture environmental dynamics and the behaviors of other agents (Canese et al., 2021; Ning & Xie, 2024).

Centralized Training with Decentralized Execution (CTDE) addresses this challenge by allowing agents to exploit global states or joint information during training while making independent decisions from local observations at execution time (Ning & Xie, 2024; Gronauer & Diepold, 2022). Nevertheless, global information is often confined to centralized training components rather than explicitly incorporated into decentralized decision-making(Lowe et al., 2017; Foerster et al., 2018; Rashid et al., 2020b; Gronauer & Diepold, 2022). For instance, value-decomposition methods use global states mainly for value mixing or credit assignment, while individual utilities remain conditioned on local observations(Rashid et al., 2020b;a; Wang et al., 2020a). Likewise, centralized critics in actor-critic methods use global information to estimate joint values or generate policy-update signals, but decentralized actors cannot access it during execution (Lowe et al., 2017; Foerster et al., 2018; Yu et al., 2022). Thus, global coordination cues affect decentralized policies only indirectly, limiting the ability of local policies to acquire decision-relevant coordination knowledge.

Centralized Teacher with Decentralized Student (CTDS) introduces knowledge distillation into CTDE, using globally informed teachers to guide decentralized students (Zhao et al., 2022). PTDE further constructs agent-personalized global representations to better capture individual information needs (Chen et al., 2022). Subsequent studies improve knowledge transfer by distilling action-value

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

distributions into lightweight decentralized agents (Pei et al., 2025), combining external and internal distillation (Zhou et al., 2026), or prioritizing informative teacher–student samples (Jo & Kim, 2025). More recently, MAGPO employs a centralized guider to coordinate exploration while aligning its guidance with decentralized learners (Xie et al., 2026). Despite these advances, privileged teachers still commonly rely on the complete global state or its learned representations, without explicitly distinguishing locally observed from unobserved information. Consequently, redundant or locally irrelevant information may degrade teacher decision quality and subsequent knowledge transfer.

To address these limitations, we propose Blind-spot Entity-aware Advantage Distillation (BEAD), a CTDS framework that selectively transfers globally available and decision-relevant coordination knowledge. BEAD represents the global state as entities and uses gated blind-spot compensation to augment each teacher's local context with useful entities that are unavailable in the corresponding local observation. A gating mechanism suppresses redundant information and constructs an agent-specific coordination context. Furthermore, student-value-anchored explicit advantage distillation transfers the teacher's relative action preferences through action advantages rather than raw action values, while anchoring the distillation process to the student's own value estimates. This design enables decentralized students to learn coordination-relevant action preferences while preserving a student-centered value representation compatible with decentralized execution.

Our main contributions are summarized as follows:

- We propose **gated blind-spot compensation**, which constructs agent-specific privileged contexts from entity-level global information unavailable in local observations. A gating mechanism adaptively integrates blind-spot information with local representations, reducing the influence of redundant or locally irrelevant information on teacher decision making.
- We introduce **student-value-anchored explicit advantage distillation**, which transfers the teacher's relative action preferences instead of directly matching raw action values. The objective anchors distillation to the student's value estimates, preserving student-centered TD learning while theoretically maintaining the teacher's action-selection preferences.
- Experiments on challenging SMAC and SMACv2 scenarios demonstrate the effectiveness of BEAD, which achieves the highest decentralized win rates on six of eight scenarios. In particular, BEAD attains 56.25% on 3s5z_vs_3s6z and 65.63% on 6h_vs_8z, compared with 21.88% and 56.25%, respectively, for the second strongest baseline.

2 PRELIMINARIES

2.1 DEC-POMDP SETTING

We formulate the fully cooperative multi-agent reinforcement learning task as a decentralized partially observable Markov decision process (Dec-POMDP) (Oliehoek & Amato, 2016), specified by:

$$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \Omega, \mathcal{O}, n, \gamma \rangle . \tag{1}$$

Here, $\mathcal{S}$ and $\mathcal{A} = \prod_{i=1}^{n} \mathcal{A}^i$ denote the global state space and joint action space, respectively. $\mathcal{P}$ is the transition function, $\mathcal{R}$ is the shared reward function, Ω^i is the observation space of agent i, and $\mathcal{O}^i$ is its observation function. At time t, agent i receives a local observation o_t^i, selects $a_t^i \in \mathcal{A}^i$, and all agents jointly execute $\mathbf{a}_t = (a_t^1, \ldots, a_t^n)$. The environment then returns $r_t = \mathcal{R}(s_t, \mathbf{a}_t)$ and transitions to s_{t+1}. Each agent acts according to its local action-observation history τ_t^i.

2.2 VALUE-BASED CTDE

In value-based CTDE methods, each agent estimates an individual action utility $Q^i(\tau_t^i, a)$ from its local action-observation history τ_t^i. A centralized mixing network combines individual utilities into a joint action-value function $Q_{\text{tot}}(s_t, \mathbf{a}_t)$ during training, while decentralized action selection is performed from the individual utilities at execution time (Rashid et al., 2020b).

For an available action set $\mathcal{A}_t^i$, the advantage of action a is defined as its utility relative to the mean utility of available actions:

$$A^i(\tau_t^i, a) = Q^i(\tau_t^i, a) - \frac{1}{|\mathcal{A}_t^i|} \sum_{a' \in \mathcal{A}_t^i} Q^i(\tau_t^i, a'), \quad a \in \mathcal{A}_t^i. \tag{2}$$

This definition preserves the relative preference among available actions and is invariant to an action-independent value translation.

Value-based methods optimize the joint action-value function using a TD loss:

$$\begin{aligned} y_t &= r_t + \gamma(1 - d_t) \max_{\mathbf{a}'} \bar{Q}_{\text{tot}}(s_{t+1}, \mathbf{a}'), \\ \mathcal{L}_{\text{TD}} &= \tfrac{1}{2}\mathbb{E}_{\mathcal{B}\sim\mathcal{D}} \left[\left(Q_{\text{tot}}(s_t, \mathbf{a}_t) - y_t \right)^2 \right]. \end{aligned} \tag{3}$$

where $\bar{Q}_{\text{tot}}$ denotes the target joint action-value function, d_t is the termination indicator, $\mathcal{D}$ is the replay buffer, and $\mathcal{B}$ is a minibatch sampled from $\mathcal{D}$.

3 RELATED WORK

Conventional CTDE Methods. Centralized training with decentralized execution (CTDE) is a widely adopted framework for cooperative MARL, allowing agents to exploit centralized information during training while acting from local observations at execution (Gronauer & Diepold, 2022; Ning & Xie, 2024). A representative family of CTDE methods is value decomposition, which factorizes a joint action-value function to enable decentralized action selection. VDN adopts an additive factorization (Sunehag et al., 2017), whereas QMIX improves expressiveness through state-conditioned monotonic value mixing (Rashid et al., 2020b). Subsequent work extends this framework through transformation constraints, weighted objectives, duplex dueling decomposition, and residual learning (Son et al., 2019; Rashid et al., 2020a; Wang et al., 2020a; Shen et al., 2022). Other studies examine the limitations of IGM and monotonic factorization (Hong et al., 2022; Dou et al., 2022; Xu et al., 2023; Baisero et al., 2025; Hu et al., 2026), and optimization challenges such as joint-action identification, suboptimal stable points, and gradient interference (Huang et al., 2026; Jo et al., 2026; Tao et al., 2026; Hou et al., 2026). Beyond value decomposition, centralized critics are widely used to improve policy optimization and credit assignment, as exemplified by MADDPG, COMA, DOP, FACMAC, and MAPPO (Lowe et al., 2017; Foerster et al., 2018; Wang et al., 2020b; Peng et al., 2021; Yu et al., 2022). Despite their success, conventional CTDE methods typically consume global information within value mixers or centralized critics. Its influence on decentralized policies is therefore indirect, rather than being explicitly transferred as decision-level supervision.

Privileged Teacher–Student Distillation. Privileged teacher–student methods establish an explicit pathway for transferring training-time global information to decentralized policies. CTDS introduces a centralized teacher that provides global-information-aware utility supervision to students acting from local observations (Zhao et al., 2022). Related work further explores stronger centralized training and decentralized policy factorization (Wang et al., 2022). PTDE constructs agent-specific global representations before distillation to mitigate redundancy in raw global states (Chen et al., 2022). Subsequent studies improve teacher representations, training–execution consistency, or knowledge-transfer mechanisms from different perspectives. For example, policy distillation methods combine centralized attention with decentralized execution (Pei et al., 2025); RGP uses agent-wise state reconstruction and state-consistency learning (Liang et al., 2025); and DDN combines external teacher–student transfer with intrinsic reward-based internal distillation (Zhou et al., 2026). Other work improves teacher guidance through evolutionary sampling of informative teacher–student discrepancies (Jo & Kim, 2025) and decentralization-aligned multi-agent policy guidance (Xie et al., 2026). Despite these advances, existing distillation-based methods may still introduce redundant or locally irrelevant global information into privileged teachers, while directly transferring policy outputs or action values can impose unnecessary constraints on decentralized students. To address these limitations, BEAD selectively incorporates locally unobserved entities through gated blind-spot compensation and transfers relative action preferences via student-value-anchored explicit advantage distillation.

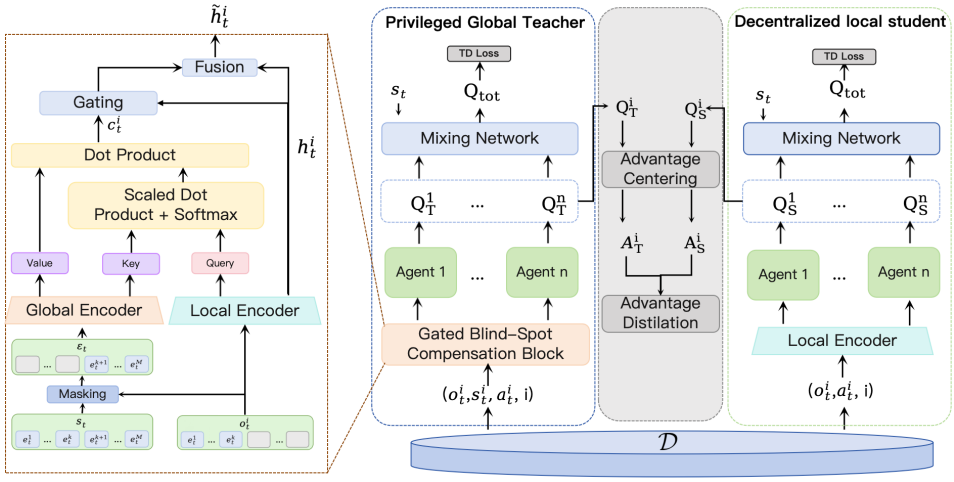


Figure 1: Overview of BEAD. The privileged teacher selectively incorporates locally unobserved global entities through gated blind-spot compensation, while the decentralized student learns from local information through TD learning and explicit advantage distillation. Only the student is retained for decentralized execution.

4 METHOD

4.1 ARCHITECTURE OVERVIEW

As illustrated in Figure 1, BEAD follows a centralized teacher with decentralized student architecture consisting of a privileged global teacher, a decentralized local student, and an explicit advantage distillation pathway between them. During centralized training, the teacher augments each agent's local representation with complementary information extracted from locally unobserved global entities through the Gated Blind-Spot Compensation block, while the student operates solely on local information. Both networks are trained with their respective TD objectives, and the teacher additionally supervises the student by transferring relative action preferences through advantage distillation. During decentralized execution, the teacher, global state, and distillation pathway are removed, leaving only the local student policies. BEAD is designed to be compatible with different value-based MARL backbones; unless otherwise specified, we instantiate the agent networks with HPN (Jianye et al., 2022), which provides entity-structured representations through permutation-invariant and permutation-equivariant network modules.

4.2 GATED BLIND-SPOT COMPENSATION

Directly incorporating the complete global state into teacher decision making may introduce decision-irrelevant information and degrade the teacher's decision quality. To address this issue, BEAD introduces **gated blind-spot compensation**, which identifies globally available but locally unobserved entities and selectively incorporates them into the teacher's local decision context.

Specifically, we represent the global state as an entity set $\mathcal{E}_t = \{e_t^1, \ldots, e_t^M\}$. For each agent i, we construct a binary visibility mask from its local observation and retain only active entities that are not locally observable, yielding the blind-spot entity set $\mathcal{E}_t^i \subseteq \mathcal{E}_t$. The agent's local information is encoded as h_t^i, while the blind-spot entities are mapped to entity representations. We then use the local representation as the query and the blind-spot entity representations as keys and values to obtain an agent-specific context:

$$c_t^i = \mathrm{Attn}\left(W_q h_t^i,\, W_k \mathcal{E}_t^i,\, W_v \mathcal{E}_t^i\right), \tag{4}$$

where the attention mechanism selectively aggregates blind-spot information according to its relevance to the current local decision context.

Not all blind-spot information is necessarily useful for every decision. Therefore, we introduce a learned gate to adaptively regulate the contribution of the extracted context:

$$g_t^i = \sigma\big(W_g[h_t^i; c_t^i] + b_g\big), \qquad \hat{c}_t^i = g_t^i \odot W_c c_t^i, \tag{5}$$

where W_g and b_g parameterize the gate, W_c projects the blind-spot context, and $\odot$ denotes element-wise multiplication. We initialize $W_g = 0$ and $b_g = -3$, yielding an initial gate value of $\sigma(-3) \approx 0.047$. This initialization suppresses blind-spot compensation at the beginning of training and allows its contribution to be learned gradually as the gate is optimized.

The gated context is then integrated with the local representation through a backbone-dependent fusion operation:

$$\tilde{h}_t^i = \mathrm{Fusion}\big(h_t^i, \hat{c}_t^i\big). \tag{6}$$

In our primary HPN-based implementation (Jianye et al., 2022), the fusion is instantiated as element-wise addition, $\tilde{h}_t^i = h_t^i + \hat{c}_t^i$; for backbones with different representation structures, the fusion operation can be adapted accordingly. The resulting representation $\tilde{h}_t^i$ is used by the teacher in the CTDS framework to estimate its individual action values Q_{T}^i.

4.3 Student-Value-Anchored Explicit Advantage Distillation

Directly distilling action values may transfer action-invariant value components that are unnecessary for action selection and constrain the student's own value estimation. To address this issue, BEAD introduces **student-value-anchored explicit advantage distillation**, which transfers the teacher's relative action preferences while retaining the student's TD-based value learning.

Following equation 2, we compute the teacher and student advantages over their respective available actions and minimize their action-wise discrepancy:

$$\mathcal{L}_{\mathrm{EAD}} = \frac{1}{|\mathcal{V}|} \sum_{(t,i,a)\in\mathcal{V}} \Big(A_{\mathrm{S},t}^i(a) - \mathrm{sg}\left[A_{\mathrm{T},t}^i(a)\right]\Big)^2, \tag{7}$$

where $\mathcal{V}$ contains valid time-step, agent, and available-action tuples, and $\mathrm{sg}[\cdot]$ denotes stop-gradient. Unlike action-value matching, this objective explicitly transfers the relative preference structure over feasible actions without requiring the student to match the teacher's absolute value scale.

We further characterize the decision-preservation property of explicit advantage distillation. For agent i at time t, define the action-averaged utility and advantage matching errors as:

$$\mathcal{L}_Q^{i,t} = \frac{1}{|\mathcal{A}_t^i|} \sum_{a\in\mathcal{A}_t^i} \big(Q_{\mathrm{T},t}^i(a) - Q_{\mathrm{S},t}^i(a)\big)^2, \qquad \mathcal{L}_A^{i,t} = \frac{1}{|\mathcal{A}_t^i|} \sum_{a\in\mathcal{A}_t^i} \big(A_{\mathrm{T},t}^i(a) - A_{\mathrm{S},t}^i(a)\big)^2, \tag{8}$$

where $\mathcal{A}_t^i$ denotes the available action set. Let V_{T}^i and V_{S}^i denote the corresponding mean action values over $\mathcal{A}_t^i$.

Theorem 1 (Decision Preservation and Error Decomposition). *Mean-centering over the same available action set preserves all pairwise action-value differences and greedy action sets. Furthermore:*

$$\mathcal{L}_Q^{i,t} = \mathcal{L}_A^{i,t} + \big(V_{\mathrm{T}}^i - V_{\mathrm{S}}^i\big)^2. \tag{9}$$

Theorem 1 shows that advantage matching isolates disagreement in relative action preferences from differences in mean action values while preserving the action-selection structure relevant to decentralized execution. The complete proof is provided in Appendix A.1.

To retain return-grounded value learning, the student is jointly optimized with the TD objective defined in equation 3 and explicit advantage distillation:

$$\mathcal{L}_{\mathrm{S}} = \lambda_{\mathrm{TD}}\mathcal{L}_{\mathrm{TD}}^{\mathrm{S}} + \lambda_{\mathrm{EAD}}(t)\mathcal{L}_{\mathrm{EAD}}, \tag{10}$$

where $\mathcal{L}_{\mathrm{TD}}^{\mathrm{S}}$ denotes the student TD loss following equation 3, which anchors the student's value estimation to environmental returns, while $\mathcal{L}_{\mathrm{EAD}}$ transfers the teacher's relative action preferences. We gradually increase $\lambda_{\mathrm{EAD}}(t)$ after initial TD learning, allowing distillation to act on an increasingly stable student value representation. After training, the teacher and distillation components are discarded, and the student acts independently using only local information. The complete training procedure is provided in Algorithm A.2.

5 EXPERIMENT

We evaluate BEAD on the StarCraft Multi-Agent Challenge (SMAC) (Samvelyan et al., 2019) and SMACv2 (Ellis et al., 2023). Our experiments address the following questions: (i) whether BEAD improves decentralized execution performance under partial observability; (ii) whether the proposed approach generalizes to procedurally generated scenarios; and (iii) how gated blind-spot compensation and explicit advantage distillation contribute to performance.

5.1 EXPERIMENTAL SETUP

Environments. We evaluate BEAD on two challenging cooperative MARL benchmarks, SMAC and SMACv2. For SMAC, we consider six scenarios: `3s_vs_5z`, `5m_vs_6m`, `MMM2`, `3s5z_vs_3s6z`, `6h_vs_8z`, and `corridor`, covering heterogeneous unit compositions, asymmetric forces, and challenging coordination requirements. We further evaluate on `10gen_terran` and `10gen_zerg` from SMACv2, where unit configurations vary across episodes to provide greater stochasticity and diversity.

Baselines. We compare BEAD with representative cooperative MARL methods from three categories. First, we consider value factorization methods, including QMIX (Rashid et al., 2020b), CW-QMIX (Rashid et al., 2020a), S2Q (Jo et al., 2026), and POW (Huang et al., 2026). Second, we include HPN (Jianye et al., 2022), which enhances individual utility networks with entity-aware permutation-invariant and permutation-equivariant representations. Third, we compare against teacher–student distillation methods, including CTDS (Zhao et al., 2022) and PTDE (Chen et al., 2022), which exploit centralized information during training to guide decentralized policies. These baselines provide comparisons with both conventional value-based MARL and methods that explicitly transfer centralized knowledge to decentralized agents.

Training and Evaluation. All methods are trained for two million environment steps. For BEAD, the teacher collects trajectories during centralized training, while the student is trained from the same replay data with TD learning and explicit advantage distillation. Only the student is retained for decentralized evaluation. For SMAC, we report the final test win rate. For SMACv2, where test performance can exhibit noticeable late-stage fluctuations, we report the mean test win rate over the final five evaluation points.

5.2 MAIN RESULTS

Table 1 summarizes the performance of BEAD and the baselines on SMAC and SMACv2. On SMAC, the BEAD student achieves the best decentralized performance on four of the six scenarios and matches the best result on `MMM2`. The improvements are particularly pronounced on the challenging asymmetric scenarios. Compared with HPN, BEAD improves the win rate from 21.88% to 56.25% on `3s5z_vs_3s6z`, from 56.25% to 65.63% on `6h_vs_8z`, and from 71.88% to 78.13% on `corridor`. These results suggest that transferring complementary blind-spot information through the teacher can improve decentralized decision making in scenarios requiring stronger coordination.

On SMACv2, the BEAD student achieves the best decentralized performance on both evaluated scenarios, obtaining 73.75% on `10gen_terran` and 80.00% on `10gen_zerg`. Compared with HPN, this corresponds to improvements of 1.87 and 2.50 percentage points, respectively. The gains under procedurally varying unit configurations further demonstrate the effectiveness of BEAD beyond fixed SMAC scenarios.

Performance on Challenging Coordination Tasks. To evaluate whether BEAD can effectively improve decentralized policies in tasks requiring complex coordination under partial observability, we examine its learning behavior on the six SMAC scenarios. As shown in Figure 2, BEAD achieves substantial improvements on the more challenging scenarios, with particularly clear advantages on `3s5z_vs_3s6z`, `6h_vs_8z`, and `corridor`. Notably, on `3s5z_vs_3s6z`, where most baselines fail to obtain non-zero win rates, BEAD exhibits a substantially stronger learning trajectory and reaches a final win rate of 56.25%. Together with the improvements on `6h_vs_8z` and `corridor`, these results demonstrate that BEAD is particularly effective in scenarios where decentralized agents require strong coordination from incomplete local observations.

Table 1: Performance comparison on SMAC and SMACv2. SMAC entries report the final test win rates (%) at or before 2M environment steps, while SMACv2 entries report the mean test win rate (%) over the last five evaluation points. † denotes the teacher used only during training and is excluded from decentralized performance ranking. Bold and underlined values denote the best and second-best decentralized performance, respectively.

	SMAC						SMACv2	
Method	3s5z	5m_vs_6m	MMM2	3s5z_vs_3s6z	6h_vs_8z	corridor	10gen_terran	10gen_zerg
QMIX	84.38	59.38	84.38	0.00	0.00	0.00	61.25	56.25
CW-QMIX	71.88	43.75	40.63	0.00	0.00	0.00	55.63	53.13
S2Q	<u>96.88</u>	59.38	46.88	0.00	0.00	0.00	63.75	55.00
POW	90.63	59.38	59.38	0.00	0.00	0.00	64.38	67.50
HPN	**100.00**	<u>93.75</u>	**100.00**	<u>21.88</u>	<u>56.25</u>	<u>71.88</u>	<u>71.88</u>	<u>77.50</u>
CTDS	81.25	53.13	6.25	0.00	0.00	50.00	54.38	52.50
PTDE	87.50	62.50	<u>96.88</u>	0.00	0.00	0.00	68.13	69.38
BEAD Teacher†	100.00	100.00	100.00	37.50	56.25	90.63	74.38	71.25
BEAD Student	<u>96.88</u>	**96.88**	**100.00**	**56.25**	**65.63**	**78.13**	**73.75**	**80.00**

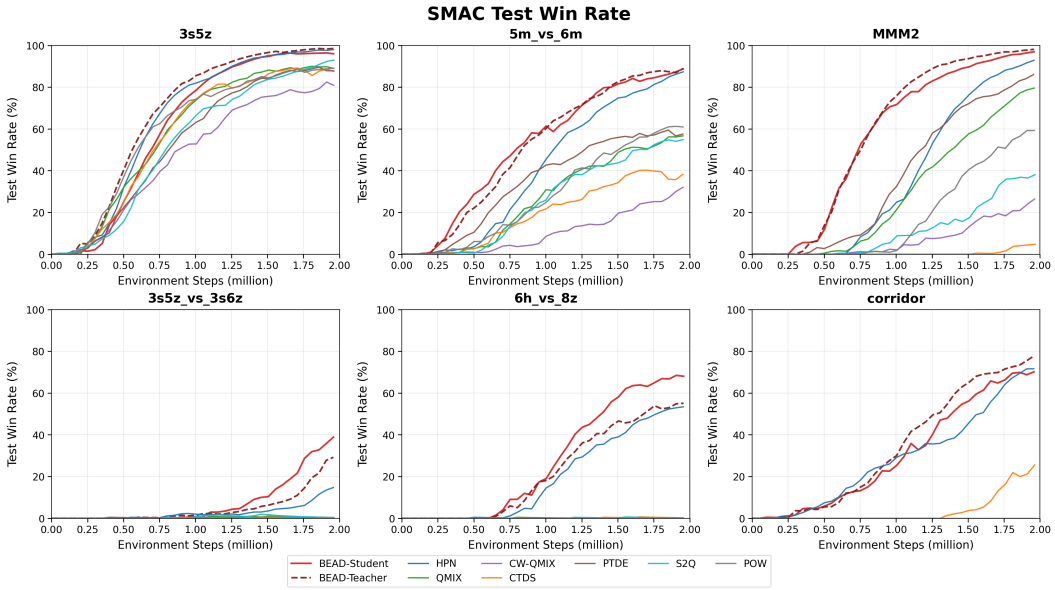


Figure 2: Test win rates on six SMAC scenarios over two million environment steps. Solid and dashed red curves correspond to the BEAD student and teacher, respectively.

Performance under Procedurally Generated Configurations. To further evaluate whether the transferred coordination knowledge remains effective under varying environment configurations, we conduct experiments on the procedurally generated `10gen_terran` and `10gen_zerg` scenarios in SMACv2. Figure 3 shows that BEAD maintains stable learning performance despite episode-wise variations in unit configurations. The BEAD student reaches late-stage win rates of 73.75% and 80.00% on the two scenarios, respectively, outperforming all decentralized baselines. These results indicate that the benefit of blind-spot-aware teacher guidance is not restricted to fixed unit configurations, but remains effective when the underlying entity composition varies across episodes.

5.3 Ablation Studies

We conduct ablation studies to examine the contributions of the two key components of BEAD and its compatibility with different value-based MARL backbones. Specifically, we investigate the effectiveness of explicit advantage distillation, the contribution of gated blind-spot compensation, and the performance of BEAD when instantiated with a standard QMIX backbone.

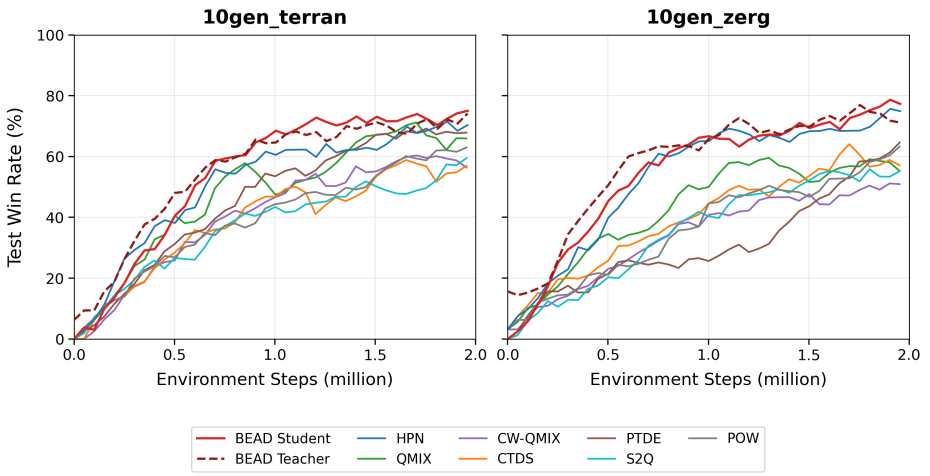


Figure 3: Test win rates on the procedurally generated SMACv2 scenarios `10gen_terran` and `10gen_zerg`. Solid and dashed orange curves denote the BEAD student and teacher.

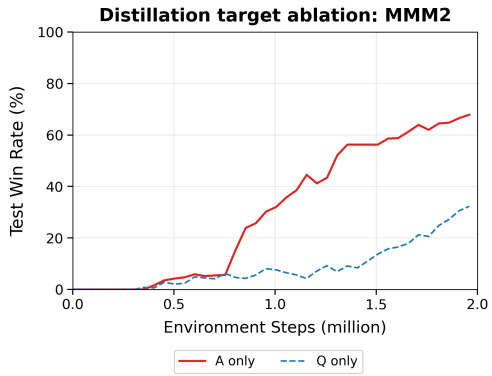


Figure 4: Comparison of explicit advantage and direct action-value distillation on `MMM2`.

Effect of Explicit Advantage Distillation. We compare explicit advantage distillation with direct action-value distillation on `MMM2`. To isolate the distillation target, both variants remove the student TD loss and differ only in matching the teacher's centered advantages or raw action values.

Explicit advantage distillation reaches approximately 68% test win rate, compared with less than 35% for action-value distillation (Figure 4). This gap supports the benefit of transferring relative action preferences and is consistent with Theorem 1, which separates advantage disagreement from mean-value discrepancy in action-value matching.

Effect of Gated Blind-Spot Compensation. A central motivation of BEAD is that additional global information is not necessarily beneficial when incorporated into the teacher indiscriminately. We examine this hypothesis on `5m_vs_6m` by removing the gating mechanism while retaining the same blind-spot entity extraction and context aggregation. In this variant, the aggregated blind-spot context is directly fused with the teacher's local representation without adaptive filtering.

The ungated teacher converges to approximately 77% test win rate, below HPN at approximately 83%, whereas gating raises teacher performance above 90% (Figure 5). Thus, additional global information can degrade teacher decision making when injected indiscriminately, while adaptive gating enables the teacher to exploit it effectively.

Compatibility with QMIX. To examine whether BEAD depends on the HPN-based implementation, we replace HPN with standard QMIX agent networks and evaluate BEAD-QMIX on

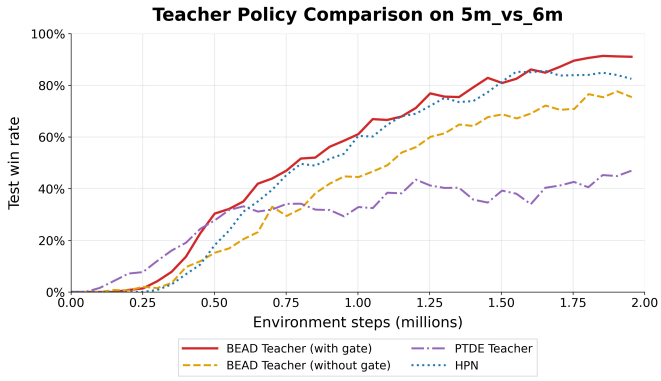


Figure 5: Teacher-policy ablation on `5m_vs_6m`. Directly injecting blind-spot information without gating degrades teacher performance, whereas gated compensation enables the teacher to effectively exploit complementary global information.

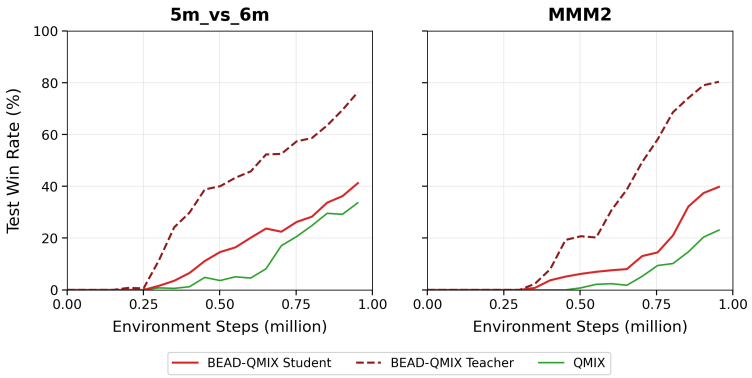


Figure 6: Performance of BEAD instantiated with standard QMIX agent networks on `5m_vs_6m` and `MMM2`, with vanilla QMIX as the corresponding baseline.

`5m_vs_6m` and `MMM2`. BEAD-QMIX consistently outperforms vanilla QMIX on both scenarios (Figure 6), showing that BEAD is not restricted to HPN-based architectures.

6 Conclusion

In this paper, we propose BEAD, a CTDS framework for selectively exploiting and effectively transferring privileged information in cooperative MARL. BEAD introduces gated blind-spot compensation to identify globally available but locally unobserved entities and adaptively incorporate their information into the teacher's local decision context. To transfer the resulting privileged knowledge to decentralized policies, we further develop student-value-anchored explicit advantage distillation, which transfers the teacher's relative action preferences while retaining the student's own value learning. Theoretical analysis establishes the action-preference preservation property of the proposed distillation objective. Experiments on challenging SMAC and SMACv2 scenarios demonstrate that BEAD improves decentralized decision making over strong CTDE and teacher–student baselines, supporting the effectiveness of selectively exploiting and transferring decision-relevant global information.

The current formulation of BEAD assumes an entity-structured global state to identify locally unobserved information, which may require additional representation mechanisms in environments with less structured observations. Future work will investigate more general blind-spot identification and representation approaches to extend BEAD to broader cooperative MARL settings.

REFERENCES

Andrea Baisero, Rupali Bhati, Shuo Liu, Aathira Pillai, and Christopher Amato. Fixing incomplete value function decomposition for multi-agent reinforcement learning. *arXiv preprint arXiv:2505.10484*, 2025.

Lorenzo Canese, Gian Carlo Cardarilli, Luca Di Nunzio, Rocco Fazzolari, Daniele Giardino, Marco Re, and Sergio Spanò. Multi-agent reinforcement learning: A review of challenges and applications. *Applied Sciences*, 11(11):4948, 2021.

Yiqun Chen, Hangyu Mao, Jiaxin Mao, Shiguang Wu, Tianle Zhang, Bin Zhang, Wei Yang, and Hongxing Chang. Ptde: Personalized training with distilled execution for multi-agent reinforcement learning. *arXiv preprint arXiv:2210.08872*, 2022.

Zehao Dou, Jakub Grudzien Kuba, and Yaodong Yang. Understanding value decomposition algorithms in deep cooperative multi-agent reinforcement learning. *arXiv preprint arXiv:2202.04868*, 2022.

Benjamin Ellis, Jonathan Cook, Skander Moalla, Mikayel Samvelyan, Mingfei Sun, Anuj Mahajan, Jakob Foerster, and Shimon Whiteson. Smacv2: An improved benchmark for cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 36:37567–37593, 2023.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Sven Gronauer and Klaus Diepold. Multi-agent deep reinforcement learning: a survey: S. gronauer, k. diepold. *Artificial Intelligence Review*, 55(2):895–943, 2022.

Yitian Hong, Yaochu Jin, and Yang Tang. Rethinking individual global max in cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 35:32438–32449, 2022.

Jie Hou, Haowen Dou, Lujuan Dang, Liangjun Chen, and Chenyang Ge. Gradient-protected value decomposition for cooperative multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21779–21787, 2026.

Tianmeng Hu, Yongzheng Cui, Rui Tang, Biao Luo, and Ke Li. Beyond monotonicity: Revisiting factorization principles in multi-agent q-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21876–21884, 2026.

Chang Huang, Shatong Zhu, Junqiao Zhao, Hongtu Zhou, Hai Zhang, Di Zhang, Chen Ye, Ziqiao Wang, and Guang Chen. Potentially optimal joint actions recognition for cooperative multi-agent reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 47566–47589, 2026.

Yixin Huang, Shufan Wu, Zhongcheng Mu, Xiangyu Long, Sunhao Chu, and Guohong Zhao. A multi-agent reinforcement learning method for swarm robots in space collaborative exploration. In *2020 6th International Conference on Control, Automation and Robotics (ICCAR)*, pp. 139–144. IEEE, 2020.

HAO Jianye, Xiaotian Hao, Hangyu Mao, Weixun Wang, Yaodong Yang, Dong Li, Yan Zheng, and Zhen Wang. Boosting multiagent reinforcement learning via permutation invariant and permutation equivariant networks. In *The Eleventh International Conference on Learning Representations*, 2022.

Ha Young Jo and Man-Je Kim. Evolutionary sampling for knowledge distillation in multi-agent reinforcement learning. *Mathematics*, 13(17):2734, 2025.

Yonghyeon Jo, Sunwoo Lee, and Seungyul Han. Retaining suboptimal actions to follow shifting optima in multi-agent reinforcement learning. *arXiv preprint arXiv:2602.17062*, 2026.

Qifan Liang, Yixiang Shan, Haipeng Liu, Zhengbang Zhu, Ting Long, Weinan Zhang, and Yuan Tian. Reconstruction-guided policy: Enhancing decision-making through agent-wise state consistency. In *International Conference on Learning Representations*, volume 2025, pp. 57967–57984, 2025.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in Neural Information Processing Systems*, 30, 2017.

Elvin Ngu, Leandro Parada, Jose Javier Escribano Macias, and Panagiotis Angeloudis. Decentralised multi-agent reinforcement learning approach for the same-day delivery problem. *Transportation Research Record*, 2676(11):385–395, 2022.

Zepeng Ning and Lihua Xie. A survey on multi-agent reinforcement learning and its application. *Journal of Automation and Intelligence*, 3(2):73–91, 2024.

Frans A Oliehoek and Christopher Amato. *A Concise Introduction to Decentralized POMDPs*. SpringerBriefs in Intelligent Systems. Springer, 2016.

Yuhang Pei, Tao Ren, Yuxiang Zhang, Zhipeng Sun, and Matys Champeyrol. Policy distillation for efficient decentralized execution in multi-agent reinforcement learning. *Neurocomputing*, 626: 129617, 2025.

Bei Peng, Tabish Rashid, Christian Schroeder de Witt, Pierre-Alexandre Kamienny, Philip Torr, Wendelin Böhmer, and Shimon Whiteson. Facmac: Factored multi-agent centralised policy gradients. *Advances in Neural Information Processing Systems*, 34:12208–12221, 2021.

Tabish Rashid, Gregory Farquhar, Bei Peng, and Shimon Whiteson. Weighted qmix: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 33:10199–10210, 2020a.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51, 2020b.

Mikayel Samvelyan, Tabish Rashid, Christian Schroeder De Witt, Gregory Farquhar, Nantas Nardelli, Tim GJ Rudner, Chia-Man Hung, Philip HS Torr, Jakob Foerster, and Shimon Whiteson. The starcraft multi-agent challenge. *arXiv preprint arXiv:1902.04043*, 2019.

Siqi Shen, Mengwei Qiu, Jun Liu, Weiquan Liu, Yongquan Fu, Xinwang Liu, and Cheng Wang. Resq: A residual q function-based approach for multi-agent reinforcement learning value factorization. *Advances in Neural Information Processing Systems*, 35:5471–5483, 2022.

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

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*, 2017.

Lesong Tao, Yifei Wang, Haodong Jing, Jingwen Fu, Miao Kang, Shitao Chen, and Nanning Zheng. Breakthrough the suboptimal stable point in value-factorization-based multi-agent reinforcement learning. *arXiv preprint arXiv:2604.05297*, 2026.

Jiangxing Wang, Deheng Ye, and Zongqing Lu. More centralized training, still decentralized execution: Multi-agent conditional policy factorization. *arXiv preprint arXiv:2209.12681*, 2022.

Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. Qplex: Duplex dueling multi-agent q-learning. *arXiv preprint arXiv:2008.01062*, 2020a.

Yihan Wang, Beining Han, Tonghan Wang, Heng Dong, and Chongjie Zhang. Dop: Off-policy multi-agent decomposed policy gradients. In *International Conference on Learning Representations*, 2020b.

Guangming Xie, Zongqing Lu, et al. Multi-agent guided policy optimization. In *International Conference on Learning Representations*, volume 2026, pp. 100398–100418, 2026.

Zhiwei Xu, Bin Zhang, Guangchong Zhou, Zeren Zhang, Guoliang Fan, et al. Dual self-awareness value decomposition framework without individual global max for cooperative marl. *Advances in Neural Information Processing Systems*, 36:73898–73918, 2023.

Xuqian Xue, Jun Zhang, Qi Cai, Zhizhong Huang, Hongming Shan, and Junping Zhang. Grpo-based cluster decision agent for unknown-k multi-view clustering. In *Forty-third International Conference on Machine Learning*, 2026.

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

Jian Zhao, Xunhan Hu, Mingyu Yang, Wengang Zhou, Jiangcheng Zhu, and Houqiang Li. Ctds: Centralized teacher with decentralized student for multiagent reinforcement learning. *IEEE Transactions on Games*, 16(1):140–150, 2022.

Ming Zhou, Jun Luo, Julian Villella, Yaodong Yang, David Rusu, Jiayu Miao, Weinan Zhang, Montgomery Alban, Iman Fadakar, Zheng Chen, et al. Smarts: Scalable multi-agent reinforcement learning training school for autonomous driving. *arXiv preprint arXiv:2010.09776*, 2020.

Yang Zhou, Siying Wang, Wenyu Chen, Ruoning Zhang, Zhitong Zhao, Zixuan Zhang, and Zijun Ma. Double distillation network for multi-agent reinforcement learning. *Neurocomputing*, pp. 133438, 2026.

A Appendix

A.1 Proof of Theorem

Proof. Fix an agent i and a time step t, and let $\mathcal{A} = \mathcal{A}_t^i$ with $|\mathcal{A}| = m$. Define the mean action values

$$V_{\mathrm{T}}^i(\tau_t^i, \mathcal{E}_t) = \frac{1}{m} \sum_{a \in \mathcal{A}} Q_{\mathrm{T}}^i(\tau_t^i, \mathcal{E}_t, a), \tag{A1}$$

and

$$V_{\mathrm{S}}^i(\tau_t^i) = \frac{1}{m} \sum_{a \in \mathcal{A}} Q_{\mathrm{S}}^i(\tau_t^i, a). \tag{A2}$$

For any $a, b \in \mathcal{A}$, the teacher advantages satisfy

$$\begin{aligned} &A_{\mathrm{T}}^i(\tau_t^i, \mathcal{E}_t, a) - A_{\mathrm{T}}^i(\tau_t^i, \mathcal{E}_t, b) \\ &\quad = Q_{\mathrm{T}}^i(\tau_t^i, \mathcal{E}_t, a) - Q_{\mathrm{T}}^i(\tau_t^i, \mathcal{E}_t, b). \end{aligned} \tag{A3}$$

The same result holds for the student advantages. Hence, mean-centering preserves all pairwise action-value differences and the greedy action sets.

Let

$$\Delta A(a) = A_{\mathrm{T}}^i(\tau_t^i, \mathcal{E}_t, a) - A_{\mathrm{S}}^i(\tau_t^i, a), \tag{A4}$$

and

$$\Delta V = V_{\mathrm{T}}^i(\tau_t^i, \mathcal{E}_t) - V_{\mathrm{S}}^i(\tau_t^i). \tag{A5}$$

Since $Q_{\mathrm{T}}^i = A_{\mathrm{T}}^i + V_{\mathrm{T}}^i$ and $Q_{\mathrm{S}}^i = A_{\mathrm{S}}^i + V_{\mathrm{S}}^i$, we obtain

$$\begin{aligned} \mathcal{L}_Q &= \frac{1}{m} \sum_{a \in \mathcal{A}} \left(\Delta A(a) + \Delta V \right)^2 \\ &= \mathcal{L}_A + \frac{2\Delta V}{m} \sum_{a \in \mathcal{A}} \Delta A(a) + (\Delta V)^2. \end{aligned} \tag{A6}$$

By the definitions of $A_{\mathrm{T}}^i(\tau_t^i, \mathcal{E}_t, a)$ and $A_{\mathrm{S}}^i(\tau_t^i, a)$,

$$\sum_{a \in \mathcal{A}} \Delta A(a) = 0. \tag{A7}$$

Therefore, the cross term vanishes and

$$\mathcal{L}_Q = \mathcal{L}_A + \left(V_{\mathrm{T}}^i(\tau_t^i, \mathcal{E}_t) - V_{\mathrm{S}}^i(\tau_t^i) \right)^2. \tag{A8}$$

□

A.2 Training Procedure

Algorithm 1 summarizes the complete training procedure of BEAD. During centralized training, the teacher learns with gated blind-spot compensation, while the student is jointly optimized by TD learning and explicit advantage distillation.

A.3 Experimental Details

A.3.1 SMAC Benchmark Details

The StarCraft Multi-Agent Challenge (SMAC) is a cooperative micromanagement benchmark built on StarCraft II. Each allied unit is controlled by one agent that receives local observations within its field of view. Agents select from movement, attack, and other available actions while sharing a team-level reward and the objective of defeating the enemy team.

The considered maps include scenarios with asymmetric unit numbers, heterogeneous unit compositions, and demanding coordination requirements. Table 2 summarizes their configurations, where

Algorithm 1 Training Procedure of BEAD

Require: Teacher parameters θ_{T}, student parameters θ_{S}, replay buffer $\mathcal{D}$
Ensure: Decentralized student Q_{S}
1: Initialize teacher and student networks
2: Initialize replay buffer $\mathcal{D}$
3: **repeat**
4: Collect trajectories using the teacher policy and store them in $\mathcal{D}$
5: Sample a mini-batch $\mathcal{B}$ from $\mathcal{D}$
6: Construct agent-specific teacher representations using gated blind-spot compensation according to equation 4–equation 6
7: Update the teacher by minimizing $\mathcal{L}_{\mathrm{TD}}^{\mathrm{T}}$ in equation 3
8: Compute teacher and student advantages according to equation 2
9: Compute the explicit advantage distillation loss $\mathcal{L}_{\mathrm{EAD}}$ in equation 7
10: Update the student using $\mathcal{L}_{\mathrm{S}}$ in equation 10
11: Periodically update the target networks
12: **until** reaching the maximum number of environment steps
13: **return** Q_{S} for decentralized execution

Table 2: Characteristics of the considered SMAC scenarios.

Scenario	Difficulty	Ally Units	Enemy Units	Type
5m_vs_6m	HARD	5 Marines	6 Marines	asymmetric & homogeneous
		3 Stalkers	3 Stalkers	
3s5z	HARD	5 Zealots	5 Zealots	heterogeneous & symmetric
		1 Medivac	1 Medivac	
		2 Marauders	3 Marauders	
MMM2	SUPER HARD	7 Marines	8 Marines	heterogeneous & asymmetric
		3 Stalkers	3 Stalkers	
3s5z_vs_3s6z	SUPER HARD	5 Zealots	6 Zealots	heterogeneous & asymmetric
6h_vs_8z	SUPER HARD	6 Hydralisks	8 Zealots	micro-trick: focus fire
corridor	SUPER HARD	6 Zealots	24 Zerglings	micro-trick: wall off

n_a and n_e denote the numbers of allied agents and enemy units, respectively, and the episode limit is measured in environment steps.

In the map names, m, s, z, and h denote Marine, Stalker, Zealot, and Hydralisk, respectively. The 5m_vs_6m and 3s_vs_5z maps test coordination under a numerical disadvantage. The 3s5z_vs_3s6z and 6h_vs_8z maps further increase difficulty through asymmetric force compositions. The heterogeneous MMM2 scenario requires coordination among damage-dealing units and the Medivac support unit.

A.3.2 Hyperparameters

The hyperparameters for the baseline algorithms are listed in Table 3. The QMIX-family baselines include QMIX, CW-QMIX, S2Q, CTDS, POW-QMIX, and PTDE. HPN-QMIX uses the same value-learning settings, with additional entity-structured network hyperparameters. The hyperparameters for G-EAD are listed in Table 4.

For 3s5z_vs_3s6z, the number of parallel environments is reduced to 4. For 6h_vs_8z, TD(λ) is set to 0.3 and epsilon is annealed over 500K environment steps. In HPN-based variants, the number of heads is set to 2 for 3s_vs_5z and 6h_vs_8z.

Large Language Model Use Disclosure

Large language models were used only as writing assistants for English language polishing, grammar checking, and improving the clarity of presentation. They were not used to design the proposed method, derive the theoretical results, generate or modify experimental data, conduct the experiments, select the reported results, or determine the scientific claims of this work. The authors are responsible for the final content, correctness, and originality of the manuscript.

Table 3: Hyperparameter settings for the baseline algorithms.

Hyperparameter	QMIX-family	HPN-QMIX
Optimizer	Adam	Adam
Training steps	2M	2M
Learning rate	0.001	0.001
Discount factor	0.99	0.99
Batch size	128	128
Replay buffer size	5000	5000
Target update interval	200	200
TD(λ)	0.6	0.6
Epsilon anneal steps	100K	100K
RNN hidden dimension	64	64
Mixing embedding dimension	32	32
Hypernetwork embedding	64	64
Entity hypernetwork dimension	–	64
Number of heads	–	1

Table 4: Hyperparameter settings for G-EAD.

Hyperparameter	Value
Teacher optimizer	Adam
Student optimizer	Adam
Teacher learning rate	0.001
Student learning rate	0.001
Training steps	2M
Discount factor	0.99
Batch size	128
Replay buffer size	5000
Target update interval	200
TD(λ)	0.6
Global context dimension	32
Entity hypernetwork dimension	64
Global gate initialization	-3.0
Distillation mode	Full advantage
Student TD coefficient	0.2
Maximum distillation weight	0.3
Distillation start step	100K
Distillation ramp end	200K
KRD filtering	Disabled