A FRAMEWORK FOR ENHANCING CTDE COOPERATION THROUGH COMMUNICATION POLICY TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

By enabling agents to exchange complementary information, communication-based multi-agent reinforcement learning (MARL) substantially enhances coordination, spurring growing interest in transferring its cooperative knowledge to decentralized agents under centralized training with decentralized execution (CTDE). However, communication policies may exploit information unavailable during decentralized execution, causing policy discrepancy and negative transfer. We theoretically analyze this gap and derive a Q-smoothing condition requiring the communication Q-function to remain consistent across feasible observations of other agents under the same local observation. Based on this analysis, we propose Communication Smoothing Distillation (CSD). CSD learns cross-agent conditional observation distributions and constructs latent communication through distribution sampling. Gradient ascent is further employed to search for communication variations that maximize Q-value discrepancy, while a smoothing perturbation loss constrains the resulting Q-value variation. The communication-smoothed agents provide Q-value targets for CTDE students, enabling the transfer of cooperative knowledge while reducing policy discrepancy. Extensive evaluations on partially observable cooperative tasks demonstrate that CSD mitigates negative transfer, outperforms advanced communication policy transfer and CTDE methods, and can be integrated into existing value-decomposition algorithms.

1 INTRODUCTION

Coordination under partial observability remains a persistent challenge in cooperative decision-making Phan et al. (2023); Kontogiannis et al. (2025). Recent work Dreyer et al. (2025) on transporting T-shaped objects through constrained passages shows that restricted communication can degrade coordination even in human groups. Cooperative multi-agent reinforcement learning (MARL) is similarly constrained by partial observability, which impairs agents' inference of the global state and other agents' behaviors and hinders cooperative policy learning. Communication-based MARL methods Yuan et al. (2022); Du et al. (2023) alleviate this information bottleneck by enabling agents to exchange complementary information through differentiable messages, attention routing, or graph-based interactions. These methods have achieved strong performance in complex cooperative tasks. However, communication deviates from the widely adopted centralized training with decentralized execution (CTDE) paradigm in MARL, limiting deployment in communication-denied tasks.

To exploit the benefits of communication while preserving decentralized execution, recent studies have explored transferring knowledge from communication agents to non-communicating agents. CTDS first formulates such a teacher-student framework, where a teacher with richer information guides CTDE agents through distillation Zhao et al. (2024). PTDE further introduces a staged training pipeline that combines offline data and feature transfer to improve policy migration Chen et al. (2024). CADP extends this line by incorporating communication pruning, where auxiliary losses constrain attention weights used for communication and gradually shift agents from full communication toward self-communication (Zhou et al., 2025). These methods demonstrate that learning guided by communication can substantially enhance CTDE policies.

Despite these advances, existing methods assume that knowledge acquired with communication can be directly transferred to non-communicating agents. This assumption overlooks a fundamental difference in how information is obtained. In a predator-prey task, once a communicating predator

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.

observes the prey, it can immediately inform its teammates and guide them toward coordinated encirclement. In contrast, under CTDE execution, other predators cannot receive this message and must still search for the prey through their own local observations. Therefore, the communication-based agent may learn its own policy via information that is unavailable to the CTDE student. Directly distilling such communication-dependent values can provide the student with targets that cannot be realized under non-communicating execution, thereby weakening knowledge transfer and even causing negative transfer.

This observation raises a more specific issue: what kind of communication-based teacher is suitable for CTDE distillation? A natural intuition is that, under the same local observation of an agent, a communicating agent should preserve policy consistency across all possible observations of the other agents. For example, in the above predator-prey task, even if one predator has found the prey, the remaining predators should still follow their own search policies when they only observe the same local information. We formalize this intuition theoretically and further extend it into a smoothing constraint suitable for MARL learning. The analysis shows that the key to reducing the gap between a communication-based teacher and a CTDE policy is to preserve the teacher's policy consistency, given the same local observation, across feasible observations of the other agents.

To realize this constraint, we propose **C**ommunication **S**moothing **D**istillation (CSD), a communication policy transfer framework that distills cooperative knowledge from communication agents to CTDE agents while mitigating negative transfer caused by their policy discrepancy. CSD employs a Gaussian mixture model (GMM) to estimate the conditional distribution of other agents' observations from each agent's local history representation, and constructs communication by sampling latent messages from the learned cross-agent distributions. It further searches within the learned observation distribution for perturbations that induce the largest Q-value discrepancy, and uses the induced variation as a smoothing constraint. The constrained communication agents are finally used as teachers to transfer their cooperative advantages to CTDE agents through teacher-student distillation.

We evaluate CSD on a diagnostic task and two challenging cooperative MARL benchmarks, StarCraft Multi-Agent Challenge (SMAC) Samvelyan et al. (2019) and Google Research Football (GRF) Kurach et al. (2020) under visibility occlusion. In the diagnostic task with severe policy discrepancy, CSD nearly recovers the performance of the original QMIX baseline, whereas existing communication policy transfer methods exhibit pronounced negative transfer. On SMAC and GRF, CSD is compared with both advanced CTDE Liu et al. (2024) and prior communication policy transfer methods Zhao et al. (2024); Chen et al. (2024); Zhou et al. (2025). The results show that CSD alleviates negative transfer in communication policy distillation while substantially improving cooperative performance. CSD outperforms the SOTA communication policy transfer method by 36.0% on SMAC and 12.5% on GRF, and exceeds the CTDE baseline by 58.0% on SMAC.

Our contributions are summarized as follows.

- We theoretically characterize communication to CTDE transfer and show that effective distillation requires preserving policy consistency under the same local observation, leading to a Q-smoothing condition.
- We propose CSD, which smooths communication policies by identifying perturbations within learned cross-agent distributions and constraining the resulting Q-value discrepancy.
- We evaluate CSD on predator-prey, SMAC, and GRF, showing reduced negative transfer, improved performance over CTDE and communication policy transfer baselines, and plug-and-play applicability to existing value decomposition methods.

2 Related Work

Value decomposition: In MARL, value-decomposition methods address credit assignment under partially observable CTDE by factorizing the joint action-value function into agent-wise utilities. Following the additive decomposition introduced by VDN Sunehag et al. (2018), subsequent methods have developed state-conditioned monotonic mixing Rashid et al. (2020b), constraint-relaxation and weighted optimization Son et al. (2019); Rashid et al. (2020a), and attention-based contribution modeling Yang et al. (2020) to capture more complex agent interactions and alleviate the representational limitations of value factorization.

Communication-based MARL: Communication alleviates partial observability in cooperative MARL by enabling agents to augment local observations with complementary information from teammates, thereby improving coordinated decision-making. Existing methods instantiate this mechanism through attention-based routing and dynamic grouping Das et al. (2019); Jiang & Lu (2018), as well as graph-structured interactions and adaptive message passing Jiang et al. (2020); Niu et al. (2021). However, their reliance on message exchange during execution violates the decentralized-execution setting of CTDE and limits their deployment in communication-constrained tasks.

Communication policy transfer: The cooperative gains achieved by communication agents have motivated recent studies to transfer communication-acquired knowledge into CTDE policies. These methods distill cooperative knowledge from global-information, centralized, or communication-enabled teachers into local policies through individual value guidance Zhao et al. (2024), agent-specific representation transfer Chen et al. (2024), or progressive communication pruning Zhou et al. (2025), thereby enabling communication-free decentralized execution. Despite these advances, existing methods overlook policy discrepancy by assuming that communication-derived knowledge can be directly transferred to non-communicating agents.

3 PRELIMINARIES

We consider a cooperative multi-agent system with n agents, modeled as a decentralized partially observable Markov decision process (Dec-POMDP). At time step t, agent i receives a local observation o_i^t and selects an action $a_i^t \in \mathcal{A}_i$, while all agents share the reward r_t. The joint observation is denoted by $\mathbf{o}^t = (o_1^t, \ldots, o_n^t) \in \Omega$, where Ω is the set of joint observations that may occur during execution. The joint policy $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$ is optimized to maximize

$$J(\boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t r_t \right], \tag{1}$$

where $\gamma \in [0, 1)$ is the discount factor. To mitigate the learning challenges induced by partial observability, cooperative MARL commonly adopts CTDE, under which global information supports value estimation and credit assignment during training, whereas each agent executes a decentralized policy $\pi_i^{\mathrm{ctde}}(a_i \mid o_i)$ conditioned solely on its local observation.

4 METHOD

4.1 THEORETICAL ANALYSIS

As discussed above, the challenge of distilling communication policy into CTDE agents lies in the policy discrepancy induced by communication. A communication policy may condition its decision on information received from other agents, whereas a CTDE policy can only use the local observation o_i during execution. Consequently, variations induced by other agents' information may produce decision targets that cannot be represented by $\pi_i^{\mathrm{ctde}}(a_i \mid o_i)$. We first characterize the exact representability of communication policies under CTDE and then relax this condition through communication smoothing.

Definition 1. *Agent i is called a fully communicating agent if its execution-time decision can depend on the entire joint observation $\mathbf{o}$. The fully communicating policy is denoted by $\pi_i^{\mathrm{com}}(a_i \mid o)$, where $a_i \in \mathcal{A}_i$.*

Based on this definition, we now characterize the condition under which a fully communicating policy remains exactly representable under the CTDE execution constraint.

Proposition 1. *A fully communicating policy π_i^{com} is exactly representable on Ω by a CTDE policy $\pi_i^{\mathrm{ctde}}(a_i \mid o_i)$, i.e.,*

$$\pi_i^{\mathrm{com}}(a_i \mid o) = \pi_i^{\mathrm{ctde}}(a_i \mid o_i), \quad \forall o \in \Omega,\ \forall a_i \in \mathcal{A}_i, \tag{2}$$

if and only if π_i^{com} assigns the same action distribution to any two joint observations in Ω that share the same local observation o_i for agent i.

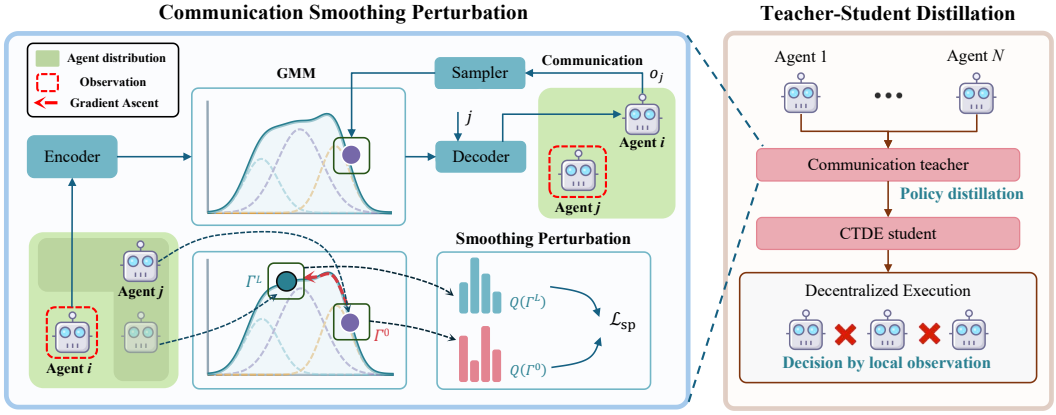


Figure 1: CSD constructs latent communication from cross-agent observation distributions, constrains the perturbation that maximizes Q-value discrepancy, and distills the communication-smoothed teacher into CTDE students.

Proposition 1 clarifies the source of the policy discrepancy between communicating and non-communicating agents. Specifically, a communication policy can be exactly transferred to a CTDE agent only if it assigns the same action distribution to all joint observations that share the same local observation for that agent. Since exact invariance is a hard structural requirement and is difficult to enforce in high-dimensional RL tasks, we relax it into a bounded-variation constraint, leading to the following communication-smoothed policy.

Definition 2. *Let $D(\cdot,\cdot)$ be a discrepancy measure between action distributions and let $\epsilon \geq 0$. A communication policy π^{cs} is called an ϵ-smooth communication policy on Ω if, for every agent i and any two joint observations $o, o' \in \Omega$ with the same local observation $o_i = o'_i$, it satisfies*

$$D\left(\pi_i^{\mathrm{cs}}(\cdot \mid o), \pi_i^{\mathrm{cs}}(\cdot \mid o')\right) \leq \epsilon. \tag{3}$$

When $\epsilon = 0$, Definition 2 reduces to the exact invariance condition in Proposition 1. We further present the following proposition to characterize the relationship between communication-smoothed policies and the CTDE policy class, showing that every communication-smoothed policy admits a CTDE policy with bounded representation error.

Proposition 2. *Suppose that the communication-smoothed policy π^{cs} satisfies Definition 2 on Ω with constant ϵ. Then, for each agent i, there exists a CTDE policy $\pi_i^{\mathrm{ctde}}(a_i \mid o_i)$ such that*

$$D\left(\pi_i^{\mathrm{cs}}(\cdot \mid o), \pi_i^{\mathrm{ctde}}(\cdot \mid o_i)\right) \leq \epsilon, \quad \forall o \in \Omega. \tag{4}$$

Proposition 2 shows that an ϵ-smooth communication policy admits a CTDE policy whose discrepancy is upper bounded by the smoothing constant. Hence, reducing the smoothing error leads to a tighter upper bound on the discrepancy between the communication-based policy and its CTDE counterpart. This result provides a principled objective for distilling communication-based policies into CTDE policies. Since many multi-agent RL methods define decentralized decisions through action-value functions, we further extend the analysis to Q-functions.

Lemma 1. *Suppose that Q_i^{cs} satisfies the ϵ_Q-smoothing condition on Ω, i.e., for any $o, o' \in \Omega$ with $o_i = o'_i$,*

$$|Q_i^{\mathrm{cs}}(o, a_i) - Q_i^{\mathrm{cs}}(o', a_i)| \leq \epsilon_Q, \quad \forall a_i \in \mathcal{A}_i. \tag{5}$$

Then there exists a CTDE action-value function $Q_i^{\mathrm{ctde}}(o_i, a_i)$ such that

$$\left|Q_i^{\mathrm{cs}}(\mathbf{o}, a_i) - Q_i^{\mathrm{ctde}}(o_i, a_i)\right| \leq \epsilon_Q, \quad \forall \mathbf{o} \in \Omega,\ \forall a_i \in \mathcal{A}_i. \tag{6}$$

4.2 Communication Smoothing Learning

4.2.1 Cross-Agent Conditional Distribution Model

Lemma 1 indicates that distilling a communication-based policy into a CTDE policy requires Q_i^{cs} to remain consistent across possible observations of other agents under the same local observation o_i.

However, each replay-buffer sample contains only one realized joint observation $o = (o_i, o_{-i})$. CSD therefore learns a cross-agent conditional distribution to model other-agent observations from agent i's local representation. From agent i's perspective, we model the joint observation distribution of the other agents under a conditional-independence assumption and introduce a target-specific latent variable z_{ij} for each $j \neq i$, yielding

$$p_{\theta_d}\left(o_{-i} \mid h_i\right) = \prod_{j \neq i} \int p_{\theta_d}\left(o_j \mid z_{ij}, h_i, e_j\right) p_0\left(z_{ij}\right) dz_{ij}, \tag{7}$$

where h_i is the GRU-encoded local history representation of agent i, e_j is the identity embedding of target agent j, p_0 is the latent prior, and z_{ij} captures the latent information required to reconstruct o_j. The likelihood $p_{\theta_d}(o_j \mid z_{ij}, h_i, e_j)$ maps the latent representation to the observation space.

Since the marginalization in Eq. (7) is generally intractable, we introduce $q_\theta(z_{ij} \mid h_i, e_j)$ as a condition-dependent variational distribution that can be generated solely from agent i's local information. Rewriting each marginal likelihood using q_θ and applying Jensen's inequality yields a tractable variational lower bound. Minimizing its negative value over all ordered agent pairs gives

$$\mathcal{L}_{\text{recon}} = \sum_{i=1}^{n} \sum_{j \neq i} \Big(- \mathbb{E}_{q_\theta(z_{ij} \mid h_i, e_j)} \left[\log p_{\theta_d}\left(o_j \mid z_{ij}, h_i, e_j\right)\right] + D_{\mathrm{KL}}\left(q_\theta\left(z_{ij} \mid h_i, e_j\right) \| p_0\left(z_{ij}\right)\right) \Big). \tag{8}$$

The observation distribution of other agents is inherently complex due to partial observability and inter-agent dependencies. To capture its multimodal structure, we parameterize $q_\theta(z_{ij} \mid h_i, e_j)$ with a GMM:

$$q_\theta\left(z_{ij} \mid h_i, e_j\right) = \sum_{k=1}^{K} w_{ij}^k \mathcal{N}\left(z_{ij}; \mu_{ij}^k, \operatorname{diag}\left((\sigma_{ij}^k)^2\right)\right), \tag{9}$$

where K denotes the number of mixture components, and w_{ij}^k, μ_{ij}^k, and σ_{ij}^k are the mixture weight, mean, and standard deviation of the k-th component, respectively. The mixture weights are non-negative and normalized over all components.

4.2.2 Latent Space Communication

After estimating the conditional observation distributions of other agents, communication is implemented through distribution-guided latent sampling. For each ordered agent pair (i, j), agent i constructs the latent distribution $q_\theta(z_{ij} \mid h_i, e_j)$, which represents the possible observation information of agent j given the local history of agent i. To encode its current observation for transmission to agent i, agent j provides o_j, together with the GMM parameters generated by agent i, to the sampler S_ψ. The sampler produces the mixture assignment $\hat{w}_{ij} = (\hat{w}_{ij}^1, \ldots, \hat{w}_{ij}^K)$ and the sampled variable $\hat{\epsilon}_{ij}$, which jointly determine a sampling position in the latent space defined by the GMM in Eq. (9). Accordingly, the latent message transmitted from agent j to agent i is given by

$$\tilde{z}_{ij} = \begin{cases} \sum_{k=1}^{K} \hat{w}_{ij}^k \left(\mu_{ij}^k + \sigma_{ij}^k \odot \hat{\epsilon}_{ij}\right), j \neq i, \\ f_i(o_i), j = i, \end{cases} \tag{10}$$

where f_i denotes the observation encoder of agent i. The latent messages received by each agent are aggregated into a unified communication feature through soft attention, which assigns larger weights to messages that are more relevant to the current decision. For agent i, the aggregation is formulated as

$$c_i = \sum_{j=1}^{n} \left[\mathcal{A}_\eta\left(\left[\phi_\eta(\tilde{z}_{i1}), \ldots, \phi_\eta(\tilde{z}_{in})\right]\right)\right]_j \phi_\eta(\tilde{z}_{ij}), \tag{11}$$

where ϕ_η is an MLP that embeds the latent messages, $\mathcal{A}_\eta$ is a soft attention module implemented by an MLP followed by a softmax operation, and $[\cdot]_j$ denotes the j-th attention weight. The aggregated feature c_i is combined with the local decision representation of agent i for value estimation.

To ensure that $\tilde{z}_{ij}$ expresses the observation information of agent j, the sampler S_ψ is trained by observation reconstruction. The latent message is decoded by the likelihood $p_{\theta_d}(o_j \mid \tilde{z}_{ij}, h_i, e_j)$, whose mean prediction $\hat{o}_{j|i}$ is required to reconstruct o_j. The reconstruction loss is defined as

$$\mathcal{L}_{\text{sample}} = \mathbb{E}_{\mathcal{D}} \left[\sum_{i=1}^{n} \sum_{j \neq i} \left\| \hat{o}_{j|i} - o_j \right\|_2^2 \right], \tag{12}$$

where $\mathcal{D}$ denotes the replay buffer. Minimizing $\mathcal{L}_{\text{sample}}$ trains the sampler to generate mixture assignments and residual codes through which the sampled latent messages retain the observation information of the sending agents.

4.2.3 Communication Smoothing Perturbation

Lemma 1 requires Q_i^{cs} to remain consistent across possible observations of other agents when the local observation of agent i is fixed. Accordingly, CSD searches the learned GMM distribution for the communication variation that maximizes the Q-value discrepancy. A smoothing perturbation loss then constrains this discrepancy to reduce the smoothing error ϵ_Q.

We define the perturbation variables for all communication samples as $\Gamma = \{(\hat{w}_{ij}, \hat{\epsilon}_{ij})\}_{i=1,j\neq i}^{n}$, with Γ^0 denoting their initial values produced by the sampler S_ψ. For any Γ, Eqs. (10) and (11) generate the aggregated latent messages used to evaluate $Q_i^{\text{cs}}(h_i^q, \Gamma, \cdot)$. The worst-case perturbation is obtained by

$$\Gamma^* = \arg\max_{\Gamma \in \mathcal{S}} \sum_{i=1}^{n} \left\| Q_i^{\text{cs}}\left(h_i^q, \Gamma, \cdot\right) - Q_i^{\text{cs}}\left(h_i^q, \Gamma^0, \cdot\right) \right\|_2^2. \tag{13}$$

where h_i^q denotes the GRU-encoded decision representation of agent i, and $\mathcal{S}$ denotes the perturbation region induced by the learned GMM, in which the perturbation associated with each Gaussian component is restricted to its corresponding 3σ range. To identify the worst-case perturbation, we approximate the inner maximization through gradient ascent over Γ. Starting from Γ^0, the perturbation can be updated as

$$\Gamma^{l+1} = \Pi_{\mathcal{S}}\left(\Gamma^l + \alpha \nabla_\Gamma D_Q\left(\Gamma^0, \Gamma^l\right)\right), \quad l = 0, \ldots, L-1, \tag{14}$$

where D_Q denotes the Q-value discrepancy in Eq. (13), α is the ascent step size, and $\Pi_{\mathcal{S}}$ projects the variables back to the feasible GMM region. After L iterations, Γ^L is used as an approximation of Γ^*. The communication agent is then optimized to reduce the Q-value variation induced by this approximate worst-case perturbation. The smoothing perturbation loss is computed as

$$\mathcal{L}_{\text{csp}} = \mathbb{E}_{\mathcal{D}} \left[\sum_{i=1}^{n} \left\| Q_i^{\text{cs}}\left(h_i^q, \Gamma^L, \cdot\right) - \left[Q_i^{\text{cs}}\left(h_i^q, \Gamma^0, \cdot\right)\right] \right\|_2^2 \right]. \tag{15}$$

By constraining the maximal perturbation, $\mathcal{L}_{\text{csp}}$ promotes consistency of the communication Q-function across feasible observations represented by the learned latent communication distribution. The smoothing perturbation loss is further incorporated into the RL training loss, yielding the overall objective:

$$\mathcal{L}_{\text{cs}} = \mathcal{L}_{\text{RL}} + \lambda_{\text{sample}} \mathcal{L}_{\text{sample}} + \lambda_{\text{csp}} \mathcal{L}_{\text{csp}}. \tag{16}$$

4.3 Teacher-Student Distillation

Minimizing the communication smoothing loss progressively reduces the Q-value approximation error between the communication agent and its corresponding CTDE agent. We further employ teacher-student knowledge distillation to learn the associated CTDE agent, using the smoothed individual Q-values as distillation targets.

$$\mathcal{L}_{\text{KD}} = \mathbb{E}_{\mathcal{D}} \left[\sum_{i=1}^{n} \left\| Q_i^{\text{ctde}}\left(h_i^s, \cdot\right) - \left[Q_i^{\text{cs}}\left(h_i^q, \Gamma^0, \cdot\right)\right] \right\|_2^2 \right], \tag{17}$$

where h_i^s denotes the local history representation of the CTDE student. Alongside distillation, we add an auxiliary RL objective for task learning, which can be instantiated by any value-decomposition method. The complete student objective is

$$\mathcal{L}_{\text{ctde}} = \mathcal{L}_{\text{RL}}^{\text{ctde}} + \lambda_{\text{KD}} \mathcal{L}_{\text{KD}}. \tag{18}$$

4.4 Overall Framework

As illustrated in Figure 1, the training process of the communication-based teacher consists of two parts. The first optimizes the cross-agent conditional distribution model using the objective in Eq. (8). The second learns latent communication through the reconstruction loss in Eq. (12), while the RL loss and communication smoothing perturbation loss jointly optimize the communication-smoothed policy. The CTDE student is trained concurrently with the teacher through Q-value distillation, together with the RL objective supplied by the underlying value-decomposition method. The complete training procedure and pseudocode are provided in Appendix B.

5 Experiments

Our experiments are organized around the following three questions:

1) Can CSD reduce teacher-student policy discrepancy and mitigate negative transfer?
2) Can CSD improve cooperative performance under CTDE in partially observable environments?
3) Can CSD be integrated with different value-decomposition methods?

To address these questions, we first conduct diagnostic experiments to examine negative transfer arising from the policy discrepancy between communication-based teachers and CTDE students, and assess the effectiveness of CSD in mitigating this discrepancy. We then evaluate CSD extensively on the cooperative SMAC and GRF benchmarks against SOTA communication policy transfer methods Zhao et al. (2024); Chen et al. (2024); Zhou et al. (2025) and advanced CTDE methods Liu et al. (2024). Finally, we integrate CSD with multiple value-decomposition methods to verify its plug-and-play capability.

5.1 Diagnostic Experiment

We conduct a diagnostic experiment in the Predator-Prey environment to examine how the discrepancy between communication-based teachers and non-communicating students affects knowledge transfer and whether CSD can mitigate it. Multiple predators must locate and jointly capture the same prey based on local observations. We compare CSD with three advanced communication policy transfer methods, CTDS Zhao et al. (2024), PTDE Chen et al. (2024), and CADP Zhou et al. (2025), together with CSD-S, which removes the smoothing perturbation loss, and standard QMIX Rashid et al. (2020b). All transfer methods use QMIX as the value-decomposition method, while standard QMIX is trained solely with the CTDE TD objective. All methods are trained for 1.5M environment steps and evaluated every $10{,}000$ steps over five random seeds. Further details are provided in Appendix C.1.

Figure 2 shows that existing communication policy transfer methods generally underperform QMIX, with negative transfer worsening as the number of agents increases. CSD instead remains comparable to QMIX. Without the smoothing perturbation loss, CSD-S student performance decreases by 6.8%,

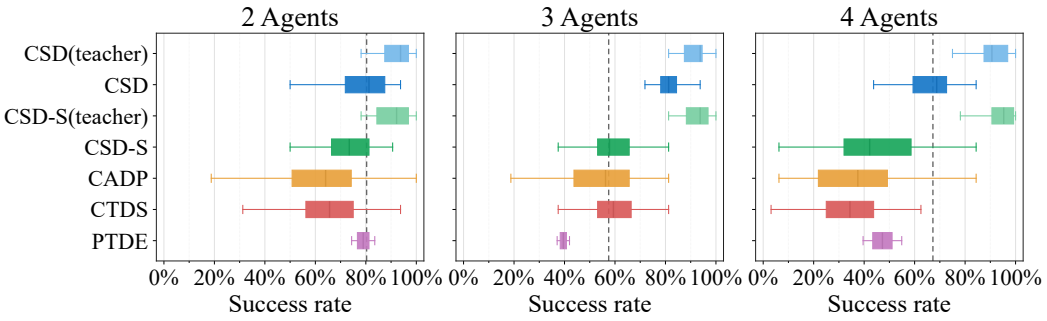


Figure 2: Box plots of success rates for different communication policy transfer methods in the diagnostic experiment. The dashed lines indicate the performance of QMIX.

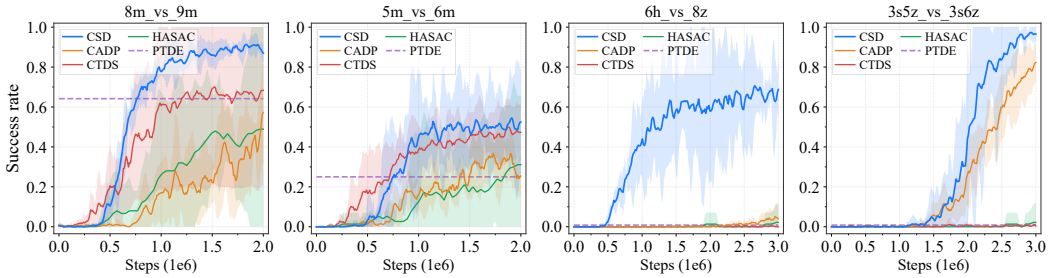


Figure 3: Mean success rates on SMAC tasks, with shaded regions denoting the 95% confidence intervals.

24.0%, and 22.6% in the two-, three-, and four-agent settings despite nearly full teacher success rates, confirming the importance of communication smoothing for effective transfer.

5.2 Cooperative MARL Experiment

To evaluate cooperative performance under CTDE, we instantiate CSD with QMIX on the SMAC and GRF benchmarks, comparing it with advanced communication policy transfer methods and the CTDE baseline Liu et al. (2024).

5.2.1 StarCraft Multi-Agent Challenge

We reduce the sight range in SMAC from 9 to 3, making policy discrepancy more likely to affect distillation under severe partial observability. We evaluate two difficult tasks, `8m_vs_9m` and `5m_vs_6m`, and two super-hard tasks, `6h_vs_8z` and `3s5z_vs_3s6z`. All methods are trained for 2M environment steps on the difficult tasks and 3M steps on the super-hard tasks, with evaluations every 10,000 steps across five random seeds.

As shown in Figure 3, CSD outperforms the other communication policy transfer methods, exceeding the second-best method CADP by 36.0% on average. It also surpasses HASAC by 58.0% on average, demonstrating that mitigating the policy discrepancy improves cooperative performance under partial observability.

Figure A2 compares teacher and student performance, with each box aggregating the final ten evaluations from five random seeds. Existing methods exhibit pronounced gaps under partial observability, especially in `8m_vs_9m`, where the teacher of the second-best method exceeds 90% success while its student performs substantially worse. In contrast, the average gap between the CSD teacher and student is only 2.6% across the four tasks, demonstrating more effective knowledge transfer to CTDE students.

In `6h_vs_8z`, the poor performance of the communication teacher also negatively affects the student. Compared with the additional communication constraint in CADP and the feature-level transfer adopted by PTDE, the smoothing perturbation loss causes less degradation to the communication teacher while directly providing individual Q-value targets for the CTDE student, thereby offering more effective policy guidance.

5.2.2 Google Research Football

We further evaluate CSD on two GRF tasks, `academy_3_vs_1_with_keeper` and `academy_run_pass_and_shoot_with_keeper`, using restricted fields of view to introduce partial observability. The visibility configurations are provided in Appendix C.3.1. Figure 4 shows that CSD achieves the best performance on both tasks and outperforms the second-best method by 12.5% on average.

5.3 Plug-and-play experiment

To evaluate the plug-and-play capability of CSD, we integrate it with three value-decomposition methods, VDN Sunehag et al. (2018), QMIX Rashid et al. (2020b), and Qatten Yang et al. (2020), and

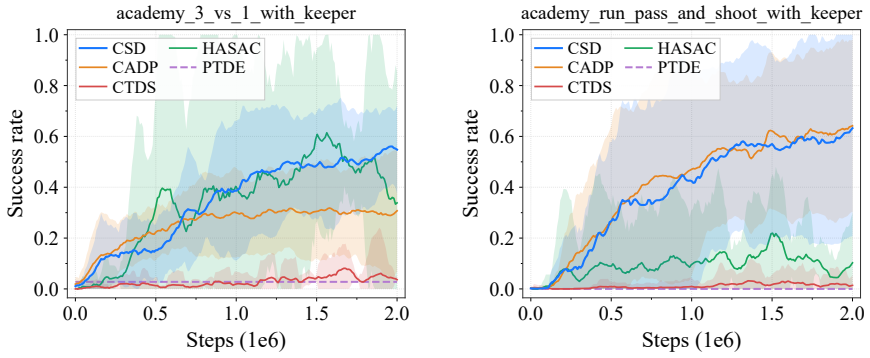


Figure 4: Success rates on two partially observable GRF tasks.

conduct experiments on two representative SMAC tasks, `8m_vs_9m` and `6h_vs_8z`. Figure A3 presents box plots based on the final ten evaluations from five random seeds. Incorporating CSD consistently improves all three methods on both tasks. The improvement is particularly pronounced on `6h_vs_8z`, where the original methods achieve success rates close to zero and almost never complete the task successfully, whereas their CSD-enhanced variants attain substantially higher success rates. These results demonstrate that CSD can be integrated with different value-decomposition methods to improve their cooperative performance under decentralized execution.

5.4 Ablation Experiment

To assess the contribution of each component, we conduct ablation experiments on two representative SMAC tasks using CADP as a reference. CSD-V replaces the GMM with a VAE, while CSD-R replaces gradient-based worst-case perturbation search with random sampling within the 3σ region. Figure A5 reports the average performance over the final ten evaluations across five random seeds. CSD consistently achieves the best or comparable performance across both tasks. Replacing the GMM with a VAE leads to a moderate performance reduction, indicating that multimodal distribution modeling provides a more suitable representation of cross-agent observations. CSD-R remains competitive on `6h_vs_8z`, but shows lower performance and substantially higher variance on `8m_vs_9m`. These results indicate that random perturbations provide partial smoothing, whereas gradient-based search more effectively identifies communication variations that induce large Q-value discrepancies and improves training stability. Both variants still substantially outperform CADP, further validating the effectiveness of the overall communication smoothing framework.

6 Conclusion

This paper proposes CSD, a plug-and-play cooperative learning framework for MARL. CSD learns communication-smoothed teacher policies by enforcing Q-value consistency across possible observations of other agents, and transfers the resulting cooperative knowledge to value-decomposition-based CTDE students through Q-value distillation. Evaluations under partially observable settings of SMAC and GRF demonstrate that CSD mitigates negative transfer and outperforms the second-best communication policy transfer method by 36.0% and 12.5%, respectively. The plug-and-play experiments further demonstrate that CSD can be integrated with different value-decomposition methods and consistently improve cooperative policy learning. Future work will extend CSD to broader non-cooperative multi-agent settings and policy-gradient MARL methods.

AI use statement

In this work, we did not use any generative AI tools to design the research, derive theoretical results, conduct experiments, analyze data, or draw scientific conclusions. All these aspects of the work were completed and verified by the authors, who take full responsibility for the manuscript.

Reproducibility statement

To support reproducibility, Appendix B describes the implementation and training procedures of CSD, and Appendix C reports the experimental settings and evaluation protocols. We also submit the main source code for CSD as supplementary material.

References

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. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 31–39. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/4. Main Track.

Caroline Claus and Craig Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pp. 746–752, 1998.

Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Mike Rabbat, and Joelle Pineau. TarMAC: Targeted multi-agent communication. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1538–1546. PMLR, 09–15 Jun 2019.

Gang Ding, Zeyuan Liu, Zhirui Fang, Kefan Su, Liwen Zhu, and Zongqing Lu. Multi-agent coordination via multi-level communication. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/d6be51e667e0b263e89a23294b57f8cf-Abstract-Conference.html.

Ziluo Ding, Tiejun Huang, and Zongqing Lu. Learning individually inferred communication for multi-agent cooperation. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Tabea Dreyer, Amir Haluts, Amos Korman, Nir Gov, Ehud Fonio, and Ofer Feinerman. Comparing cooperative geometric puzzle solving in ants versus humans. *Proceedings of the National Academy of Sciences*, 122(1):e2414274121, 2025.

Wei Du, Shifei Ding, Chenglong Zhang, and Zhongzhi Shi. Multiagent reinforcement learning with heterogeneous graph attention network. *IEEE Transactions on Neural Networks and Learning Systems*, 34(10):6851–6860, 2023. doi: 10.1109/TNNLS.2022.3215774.

Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. *CoRR*, abs/1705.08926, 2017. URL http://arxiv.org/abs/1705.08926.

Sven Gronauer and Klaus Diepold. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review*, 55:895–943, 2022. doi: 10.1007/s10462-021-09996-w.

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.

Jian Hu, Siyang Jiang, Seth Austin Harding, Haibin Wu, and Shih-wei Liao. Rethinking the implementation tricks and monotonicity constraint in cooperative multi-agent reinforcement learning. *arXiv preprint arXiv:2102.03479*, 2021.

Jiechuan Jiang and Zongqing Lu. Learning attentional communication for multi-agent cooperation. *Advances in neural information processing systems*, 31, 2018.

Jiechuan Jiang, Chen Dun, Tiejun Huang, and Zongqing Lu. Graph convolutional reinforcement learning. In *International Conference on Learning Representations*, 2020.

Andreas Kontogiannis, Konstantinos Papathanasiou, Yi Shen, Giorgos Stamou, Michael M Zavlanos, and George Vouros. Enhancing cooperative multi-agent reinforcement learning with state modelling and adversarial exploration. *arXiv preprint arXiv:2505.05262*, 2025.

Karol Kurach, Anton Raichuk, Piotr Stańczyk, Michał Zając, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, et al. Google research football: A novel reinforcement learning environment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 4501–4510, 2020.

Jiarong Liu, Yifan Zhong, Siyi Hu, Haobo Fu, QIANG FU, Xiaojun Chang, and Yaodong Yang. Maximum entropy heterogeneous-agent reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Yaru Niu, Rohan Paleja, and Matthew Gombolay. Multi-agent graph-attention communication and teaming. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, pp. 964–973, Richland, SC, 2021. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450383073.

Thomy Phan, Fabian Ritz, Philipp Altmann, Maximilian Zorn, Jonas Nüßlein, Michael Kölle, Thomas Gabor, and Claudia Linnhoff-Popien. Attention-based recurrence for multi-agent reinforcement learning under stochastic partial observability. In *International Conference on Machine Learning*, pp. 27840–27853. PMLR, 2023.

Tabish Rashid, Gregory Farquhar, Bei Peng, and Shimon Whiteson. Weighted qmix: expanding monotonic value function factorisation for deep multi-agent reinforcement learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020a. Curran Associates Inc. ISBN 9781713829546.

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 G. J. Rudner, Chia-Man Hung, Philiph H. S. Torr, Jakob Foerster, and Shimon Whiteson. The StarCraft Multi-Agent Challenge. *CoRR*, abs/1902.04043, 2019.

Hassam Ullah Sheikh and Ladislau Bölöni. Multi-agent reinforcement learning for problems with combined individual and team reward. *CoRR*, abs/2003.10598, 2020.

Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Hostallero, and Yung Yi. QTRAN: learning to factorize with transformation for cooperative multi-agent reinforcement learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5887–5896. PMLR, 2019.

Chuxiong Sun, Zehua Zang, Jiabao Li, Jiangmeng Li, Xiao Xu, Rui Wang, and Changwen Zheng. T2MAC: targeted and trusted multi-agent communication through selective engagement and evidence-driven integration. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pp. 15154–15163. AAAI Press, 2024. doi: 10.1609/AAAI.V38I13.29438. URL https://doi.org/10.1609/aaai.v38i13.29438.

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 based on team reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, pp. 2085–2087, Richland, SC, 2018. International Foundation for Autonomous Agents and Multiagent Systems.

Ming Tan. Multi-agent reinforcement learning: Independent versus cooperative agents. In *Proceedings of the Tenth International Conference on Machine Learning*, pp. 330–337, 1993.

Tonghan Wang, Jianhao Wang, Chongyi Zheng, and Chongjie Zhang. Learning nearly decomposable value functions via communication minimization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJx-3grYDB.

Yaodong Yang, Jianye Hao, Ben Liao, Kun Shao, Guangyong Chen, Wulong Liu, and Hongyao Tang. Qatten: A general framework for cooperative multiagent reinforcement learning. *arXiv preprint arXiv:2002.03939*, 2020.

Dayong Ye, Tianqing Zhu, Wanlei Zhou, and Philip S. Yu. Differentially private malicious agent avoidance in multiagent advising learning. *IEEE Transactions on Cybernetics*, 50(10):4214–4227, 2020. doi: 10.1109/TCYB.2019.2906574.

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. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=YVXaxB6L2Pl.

Lei Yuan, Jianhao Wang, Fuxiang Zhang, Chenghe Wang, ZongZhang Zhang, Yang Yu, and Chongjie Zhang. Multi-agent incentive communication via decentralized teammate modeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(9):9466–9474, Jun. 2022. doi: 10.1609/aaai.v36i9.21179.

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, 2024. doi: 10.1109/TG.2022.3232390.

Yihe Zhou, Shunyu Liu, Yunpeng Qing, Tongya Zheng, Kaixuan Chen, Jie Song, and Mingli Song. Cadp: Towards better centralized learning for decentralized execution in marl. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pp. 2838–2840, 2025.

APPENDIX

A PROOF

A.1 PROOF OF PROPOSITION 1

Proof. We prove necessity and sufficiency separately.

First, we prove necessity. Suppose that the fully communicating policy π_i^{com} is exactly representable on Ω by a CTDE policy $\pi_i^{\mathrm{ctde}}(a_i \mid o_i)$. Then, by Eq. (2) in the main paper, for every $o = (o_1, \ldots, o_n) \in \Omega$ and every $a_i \in \mathcal{A}_i$,

$$\pi_i^{\mathrm{com}}(a_i \mid o) = \pi_i^{\mathrm{ctde}}(a_i \mid o_i). \tag{19}$$

Consider any two joint observations $o = (o_1, \ldots, o_n) \in \Omega$ and $o' = (o'_1, \ldots, o'_n) \in \Omega$ that share the same local observation for agent i, i.e., $o_i = o'_i$. Since the CTDE policy takes only the local observation of agent i as input, the two joint observations induce the same input to π_i^{ctde}. Hence, for every $a_i \in \mathcal{A}_i$,

$$\pi_i^{\mathrm{ctde}}(a_i \mid o_i) = \pi_i^{\mathrm{ctde}}(a_i \mid o'_i). \tag{20}$$

Combining Eq. (19) for o and o' with Eq. (20), we obtain

$$\pi_i^{\mathrm{com}}(a_i \mid o) = \pi_i^{\mathrm{com}}(a_i \mid o'), \quad \forall a_i \in \mathcal{A}_i. \tag{21}$$

Therefore, the two action distributions are identical:

$$\pi_i^{\mathrm{com}}(\cdot \mid o) = \pi_i^{\mathrm{com}}(\cdot \mid o'). \tag{22}$$

Thus, any exactly representable fully communicating policy must assign the same action distribution to joint observations in Ω that share the same local observation for agent i.

We next prove sufficiency. Suppose that π_i^{com} assigns the same action distribution to any two joint observations in Ω that share the same local observation for agent i. We construct a CTDE policy that exactly represents π_i^{com} on Ω.

Let

$$\Omega_i = \{o_i : o = (o_1, \ldots, o_n) \in \Omega\} \tag{23}$$

be the set of local observations of agent i induced by joint observations in Ω. For each $z \in \Omega_i$, choose one representative joint observation $r(z) \in \Omega$ whose i-th component is z. Define the CTDE policy on Ω_i as

$$\pi_i^{\mathrm{ctde}}(\cdot \mid z) = \pi_i^{\mathrm{com}}(\cdot \mid r(z)), \quad z \in \Omega_i. \tag{24}$$

For local observations not contained in Ω_i, the value of π_i^{ctde} can be assigned arbitrarily, since such observations do not appear in Ω and therefore do not affect exact representation on Ω.

We first show that the construction in Eq. (24) is well-defined. Suppose that $o_1, o_2 \in \Omega$ are two possible representatives for the same local observation z, meaning that their i-th local observations are both equal to z. By the assumed invariance of π_i^{com} over joint observations with the same local observation, we have

$$\pi_i^{\mathrm{com}}(\cdot \mid o_1) = \pi_i^{\mathrm{com}}(\cdot \mid o_2). \tag{25}$$

Hence, the right-hand side of Eq. (24) is independent of the representative selected. Therefore, $\pi_i^{\mathrm{ctde}}(\cdot \mid z)$ is well-defined for every $z \in \Omega_i$.

It remains to verify exact representation. Take any joint observation $o = (o_1, \ldots, o_n) \in \Omega$. By construction,

$$\pi_i^{\mathrm{ctde}}(\cdot \mid o_i) = \pi_i^{\mathrm{com}}(\cdot \mid r(o_i)). \tag{26}$$

The representative $r(o_i)$ and the joint observation o share the same local observation for agent i. Therefore, by the assumed invariance condition,

$$\pi_i^{\mathrm{com}}(\cdot \mid r(o_i)) = \pi_i^{\mathrm{com}}(\cdot \mid o). \tag{27}$$

Combining Eq. (26) and Eq. (27) gives

$$\pi_i^{\mathrm{ctde}}(\cdot \mid o_i) = \pi_i^{\mathrm{com}}(\cdot \mid o), \quad \forall o \in \Omega. \tag{28}$$

Equivalently, for every $o \in \Omega$ and every $a_i \in \mathcal{A}_i$,

$$\pi_i^{\mathrm{com}}(a_i \mid o) = \pi_i^{\mathrm{ctde}}(a_i \mid o_i). \tag{29}$$

This is exactly Eq. (2) in the main paper. Therefore, π_i^{com} is exactly representable on Ω by the constructed CTDE policy. Combining necessity and sufficiency completes the proof. □

A.2 Proof of Proposition 2

Proof. Fix an arbitrary agent i. Since π^{cs} satisfies Definition 2, for any two joint observations $o, o' \in \Omega$ with the same local observation of agent i, i.e., $o_i = o'_i$, we have

$$D\left(\pi_i^{\mathrm{cs}}(\cdot \mid o), \pi_i^{\mathrm{cs}}(\cdot \mid o')\right) \leq \epsilon. \tag{30}$$

Let Ω_i denote the set of local observations of agent i induced by Ω:

$$\Omega_i = \{o_i : o = (o_1, \ldots, o_n) \in \Omega\}. \tag{31}$$

For each $z \in \Omega_i$, choose one representative joint observation $r(z) \in \Omega$ whose i-th local observation is z. Define a CTDE policy π_i^{ctde} on Ω_i by

$$\pi_i^{\mathrm{ctde}}(\cdot \mid z) = \pi_i^{\mathrm{cs}}(\cdot \mid r(z)), \quad z \in \Omega_i. \tag{32}$$

For local observations outside Ω_i, the value of π_i^{ctde} can be assigned arbitrarily, since they do not appear in Ω and therefore do not affect the approximation bound on Ω.

Now take any joint observation $o = (o_1, \ldots, o_n) \in \Omega$. By construction,

$$\pi_i^{\mathrm{ctde}}(\cdot \mid o_i) = \pi_i^{\mathrm{cs}}(\cdot \mid r(o_i)). \tag{33}$$

The representative $r(o_i)$ and the joint observation o have the same local observation for agent i. Therefore, applying the communication smoothing condition in Eq. (30) gives

$$D\left(\pi_i^{\mathrm{cs}}(\cdot \mid o), \pi_i^{\mathrm{cs}}(\cdot \mid r(o_i))\right) \leq \epsilon. \tag{34}$$

Combining Eq. (33) and Eq. (34), we obtain

$$D\left(\pi_i^{\mathrm{cs}}(\cdot \mid o), \pi_i^{\mathrm{ctde}}(\cdot \mid o_i)\right) \leq \epsilon, \quad \forall o \in \Omega. \tag{35}$$

Thus, there exists a CTDE policy whose approximation error to π_i^{cs} is bounded by ϵ on Ω. Since agent i was chosen arbitrarily, the conclusion holds for each agent. □

A.3 Proof of Lemma 1

Proof. Let $\Omega_i = \{o_i : o = (o_1, \ldots, o_n) \in \Omega\}$. For each $z \in \Omega_i$, choose one representative joint observation $r(z) \in \Omega$ whose i-th local observation is z. Define

$$Q_i^{\mathrm{ctde}}(z, a_i) = Q_i^{\mathrm{cs}}(r(z), a_i), \quad z \in \Omega_i,\ a_i \in \mathcal{A}_i. \tag{36}$$

For local observations outside Ω_i, Q_i^{ctde} can be assigned arbitrarily.

For any $o = (o_1, \ldots, o_n) \in \Omega$, the representative $r(o_i)$ and o have the same local observation for agent i. Therefore, by Eq. (5) in the main paper,

$$|Q_i^{\mathrm{cs}}(o, a_i) - Q_i^{\mathrm{cs}}(r(o_i), a_i)| \le \epsilon_Q, \quad \forall a_i \in \mathcal{A}_i. \tag{37}$$

Using the construction in Eq. (36), we obtain

$$\left|Q_i^{\mathrm{cs}}(o, a_i) - Q_i^{\mathrm{ctde}}(o_i, a_i)\right| \le \epsilon_Q, \quad \forall o \in \Omega,\ \forall a_i \in \mathcal{A}_i. \tag{38}$$

This proves the lemma. □

B Implementation Details of CSD

B.1 Cross-Agent Distribution Model

For each agent i, its input is projected by a linear layer into a 64-dimensional feature space, followed by a ReLU activation and a GRUCell with a 64-dimensional hidden state to obtain the local history representation h_i. From the perspective of agent i, the conditional joint distribution over the observations of the remaining agents is factorized under a conditional-independence assumption. A target-specific latent variable z_{ij} is associated with each agent $j \neq i$, and the corresponding factorization is formulated as

$$\begin{aligned} p_{\theta_d}(o_{-i} \mid h_i) = \prod_{j \neq i} \int & p_{\theta_d}(o_j \mid z_{ij}, h_i, e_j) \\ & p_0(z_{ij})\, dz_{ij}, \end{aligned} \tag{39}$$

where e_j is the identity embedding of target agent j, $p_0(z_{ij})$ denotes the latent prior, and $p_{\theta_d}(o_j \mid z_{ij}, h_i, e_j)$ reconstructs the observation of agent j from the latent variable z_{ij}. The marginal likelihood in Eq. (39) is generally intractable. We therefore introduce a condition-dependent variational distribution $q_\theta(z_{ij} \mid h_i, e_j)$. For each ordered pair (i, j), multiplying and dividing the integrand by $q_\theta(z_{ij} \mid h_i, e_j)$ rewrites the pairwise marginal likelihood as

$$\begin{aligned} p_{\theta_d}(o_j \mid h_i, e_j) = \int & q_\theta(z_{ij} \mid h_i, e_j) \\ & \cdot \frac{p_{\theta_d}(o_j \mid z_{ij}, h_i, e_j)\, p_0(z_{ij})}{q_\theta(z_{ij} \mid h_i, e_j)} dz_{ij}. \end{aligned} \tag{40}$$

Expressing the integral in Eq. (40) as an expectation under $q_\theta(z_{ij} \mid h_i, e_j)$ and applying Jensen's inequality gives

$$\begin{aligned} \log p_{\theta_d}(o_j \mid h_i, e_j) &= \log \mathbb{E}_{q_\theta(z_{ij} \mid h_i, e_j)} \left[\frac{p_{\theta_d}(o_j \mid z_{ij}, h_i, e_j) p_0(z_{ij})}{q_\theta(z_{ij} \mid h_i, e_j)} \right] \\ &\ge \mathbb{E}_{q_\theta(z_{ij} \mid h_i, e_j)} \left[\log p_{\theta_d}(o_j \mid z_{ij}, h_i, e_j) \right. \\ &\qquad \left. + \log p_0(z_{ij}) - \log q_\theta(z_{ij} \mid h_i, e_j) \right]. \end{aligned} \tag{41}$$

The last two terms inside the expectation form the negative KL divergence between the variational distribution and the latent prior. Accordingly, the pairwise evidence lower bound can be written as

$$\begin{aligned} \log p_{\theta_d}(o_j \mid h_i, e_j) \ge & \mathbb{E}_{q_\theta(z_{ij} \mid h_i, e_j)} \left[\log p_{\theta_d}(o_j \mid z_{ij}, h_i, e_j) \right] \\ & - D_{\mathrm{KL}}\left(q_\theta(z_{ij} \mid h_i, e_j) \| p_0(z_{ij}) \right). \end{aligned} \tag{42}$$

Under the pairwise factorization in Eq. (39), the joint log-likelihood is the sum of the corresponding pairwise log-likelihoods. Taking the negative of the lower bound in Eq. (42) and summing it over all ordered agent pairs therefore gives

$$\mathcal{L}_{\text{recon}} = \sum_{i=1}^{n} \sum_{j \neq i} (- \mathbb{E}_{q_\theta(z_{ij} | h_i, e_j)} \left[\log p_{\theta_d} \left(o_j \mid z_{ij}, h_i, e_j \right) \right] + D_{\text{KL}} \left(q_\theta \left(z_{ij} \mid h_i, e_j \right) \| p_0(z_{ij}) \right)). \tag{43}$$

CSD sets the latent prior to the standard Gaussian, $p_0(z_{ij}) = \mathcal{N}(0, I)$, and implements $q_\theta(z_{ij} \mid h_i, e_j)$ as an eight-component diagonal GMM with an eight-dimensional latent space. The encoder is a two-layer MLP with a 64-dimensional hidden layer and ReLU activation, which takes h_i and e_j as inputs and outputs the GMM parameters $\{w_{ij}^k, \mu_{ij}^k, \sigma_{ij}^k\}_{k=1}^K$. The likelihood $p_{\theta_d}(o_j \mid z_{ij}, h_i, e_j)$ is parameterized by a two-layer MLP decoder with a 64-dimensional hidden layer and ReLU activation, which reconstructs o_j from z_{ij}, h_i, and e_j.

B.2 Communication-Smoothed Teacher

For each ordered pair (i, j), the sampler S_ψ receives the current observation o_j and the GMM parameters associated with $q_\theta(z_{ij} \mid h_i, e_j)$. The sampler is implemented as a two-layer MLP with a 64-dimensional hidden layer and ReLU activation. Its two output heads produce K mixture logits and an eight-dimensional residual code. The mixture logits are normalized by softmax to obtain $\hat{w}_{ij}$, while the residual code $\hat{\epsilon}_{ij}$ is produced through a hyperbolic tangent activation. The pairwise latent message is computed as

$$\tilde{z}_{ij} = \begin{cases} \sum_{k=1}^{K} \hat{w}_{ij}^k \left(\mu_{ij}^k + \sigma_{ij}^k \odot \hat{\epsilon}_{ij} \right), & j \neq i, \\ f_i(o_i), & j = i, \end{cases} \tag{44}$$

where f_i is a linear encoder that maps o_i to the latent-message space.

Each latent message is subsequently mapped to a 32-dimensional representation by the two-layer MLP ϕ_η. The attention module $\mathcal{A}_\eta$ uses an MLP to assign a scalar logit to each message and normalizes these logits through softmax. The communication feature of agent i is computed as

$$c_i = \sum_{j=1}^{n} \left[\mathcal{A}_\eta \left([\phi_\eta(\tilde{z}_{i1}), \ldots, \phi_\eta(\tilde{z}_{in})] \right) \right]_j \phi_\eta(\tilde{z}_{ij}). \tag{45}$$

In parallel, the local decision input of agent i is projected by a linear layer to 64 dimensions and processed by a ReLU activation and a GRUCell with a 64-dimensional hidden state to obtain the agent's recurrent representation. The communication feature c_i is concatenated with the recurrent representation, and a two-layer MLP with a 64-dimensional hidden layer and ReLU activation maps the resulting representation to the individual teacher Q-values. The resulting teacher Q-values are then used in Eqs. (13)–(15) to perform communication smoothing, and the teacher parameters are updated according to the overall objective in Eq. (16). The pseudocode for training the communication-smoothed teacher is shown below.

B.3 Student Distillation

The communication-smoothed teacher provides individual Q-value targets for the student. The student is optimized using the overall objective in Eq. (18), which combines its RL loss with the distillation loss in Eq. (17). The pseudocode is presented in Algorithm 2.

Algorithm 1 Training Process of the Communication-Smoothed Teacher

Require: Conditional distribution model q_θ and p_{θ_d}; sampler S_ψ; teacher network Q^{cs}

for each training iteration **do**
  Construct the cross-agent conditional distributions and update the distribution model using Eq. (8)
  Compute the latent messages and communication features using Eqs. (10) and (11) and obtain Q-values.
  Initialize Γ^0 as the communication perturbation
  for $\ell = 0, \ldots, L-1$ **do**
    Update Γ^ℓ by gradient ascent and projection according to Eqs. (13) and (14)
  end for
  Compute the sampling and smoothing losses using Eqs. (12) and (15)
  Update the communication-smoothed teacher by Eq. (16)
end for

Algorithm 2 Training Process of the CTDE Student

Require: Maximum environment steps T; replay buffer $\mathcal{D}$; CTDE student; communication-smoothed teacher

for $t = 1$ to T **do**
  Interact with the environment by teacher and student, then store the resulting transition in $\mathcal{D}$
  Update the communication-smoothed teacher according to Algorithm 1
  Sample a trajectory batch $\mathcal{B}_S$ from $\mathcal{D}$
  Compute the student Q-values and the original RL loss $\mathcal{L}_{\mathrm{RL}}^{\mathrm{ctde}}$
  Evaluate the communication-smoothed teacher on $\mathcal{B}_S$ to obtain the teacher Q-values
  Compute the distillation loss $\mathcal{L}_{\mathrm{KD}}$ according to Eq. (17)
  Compute the complete student loss $\mathcal{L}_{\mathrm{ctde}}$ according to Eq. (18)
  Update the student by minimizing $\mathcal{L}_{\mathrm{ctde}}$
end for

C Experimental Details

C.1 Diagnostic experiment

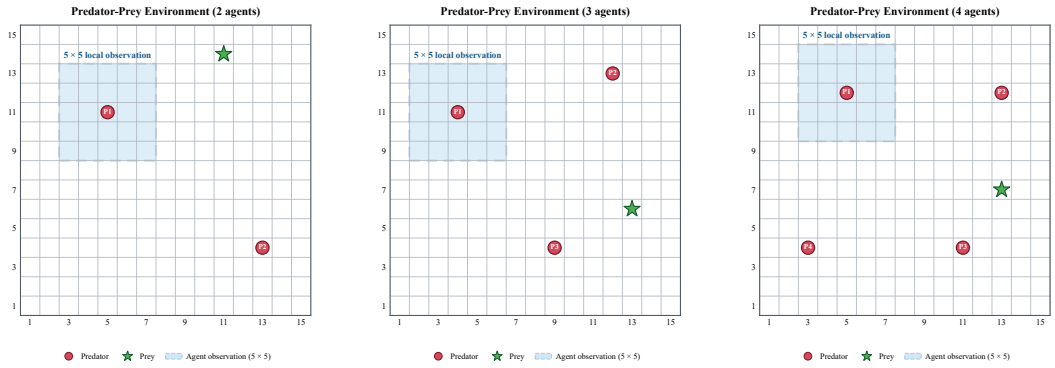
C.1.1 Environment settings

We conduct the diagnostic experiment using the Predator-Prey task. Two, three, or four predators and one prey are initialized on a 15×15 grid. Each grid cell is encoded by a single integer that jointly indicates the number of predators occupying that cell and whether the prey is present. This encoding allows multiple predators to occupy the same cell and preserves the prey indicator when predators and the prey coincide. The detailed experimental settings are described below.

Observation Each predator observes only a 5×5 local region centered on its current position, corresponding to a sight range of two grid cells in each direction. The 25 cells are flattened in row-major order, with each dimension representing the integer-encoded content of one cell. A value of 0 denotes a position outside the map boundary, whereas 1 denotes a valid empty cell. A cell containing k predators but no prey is encoded as $1 + k$. If the prey is present, an additional offset of $N + 1$ is added, where N is the total number of predators. The flattened local grid is further concatenated with the predator's two-dimensional absolute coordinates, resulting in a 27-dimensional observation. When a predator is close to the boundary, cells outside the grid are padded with zeros, allowing them to be distinguished from valid empty cells.

Action Each predator has a discrete action space consisting of five actions: moving one grid cell upward, downward, left, or right, or remaining stationary. At each time step, every predator selects one of these five actions.

Reward Each predator receives an individual reward of $+1$ when it reaches the prey and a step penalty of -0.05 otherwise. An episode is considered successful only when all predators reach the prey simultaneously. The team reward is the sum of the individual rewards.



(a) Two agents (b) Three agents (c) Four agents

Figure A1: Illustration of the Predator-Prey environments with different numbers of agents.

Under the restricted observation range, a predator that has not observed the prey cannot infer its location from its own input, whereas the communication teacher can obtain this information from another predator. The resulting information discrepancy provides a direct diagnostic setting for evaluating negative transfer from communication teachers to decentralized students. The detailed environment settings are summarized in Table 1. The hyperparameter settings of all comparison methods follow their respective original papers, while the QMIX implementation is adopted from PyMARL2 Hu et al. (2021).

Table 1: Hyperparameters for the diagnostic experiment

Parameter	Value
Optimizer	Adam
Learning rate	5×10^{-4}
Discount factor (γ)	0.99
Batch size	32
Replay buffer size	5,000
Number of GMM components	8
Number of perturbation iterations L	5
Perturbation iteration step size	0.1
Smoothing loss weight λ_{csp}	10
λ_{KD}	10
λ_{sample}	1
Sight radius	2

C.1.2 Experimental Results

Table 2 supplements the results in the main text by reporting the detailed student success rates in the diagnostic experiment. Each entry presents the mean and standard deviation over the available seeds, computed from the final ten evaluations within 1.5 million environment steps, while PTDE uses its available final-evaluation records.

Table 2: Detailed success rates in the diagnostic experiment.

Algorithm	2 Agents	3 Agents	4 Agents	Average
CSD	$78.3\% \pm 9.5\%$	$\mathbf{80.1\% \pm 1.5\%}$	$\mathbf{67.3\% \pm 3.3\%}$	$\mathbf{75.2\%}$
CSD-S	$71.5\% \pm 8.8\%$	$56.1\% \pm 7.3\%$	$44.7\% \pm 19.2\%$	57.4%
CADP	$62.3\% \pm 13.5\%$	$53.8\% \pm 9.6\%$	$36.3\% \pm 16.1\%$	50.8%
CTDS	$65.3\% \pm 12.0\%$	$58.9\% \pm 8.1\%$	$35.4\% \pm 12.1\%$	53.2%
QMIX	$\mathbf{80.3\% \pm 6.8\%}$	$57.5\% \pm 24.9\%$	$\mathbf{67.3\% \pm 5.1\%}$	68.4%
PTDE	$79.0\% \pm 6.5\%$	$39.5\% \pm 3.5\%$	$47.2\% \pm 10.8\%$	55.2%

C.2 StarCraft Multi-Agent Challenge

C.2.1 Environment settings

We evaluate CSD on two hard StarCraft Multi-Agent Challenge (SMAC) tasks, `8m_vs_9m` and `5m_vs_6m`, and two super-hard tasks, `6h_vs_8z` and `3s5z_vs_3s6z`.

To strengthen partial observability, we reduce each agent's sight range from 9 to 3. Allies and enemies outside this range are excluded from the agent's local observation by setting their corresponding features to zero. For units within sight, their relative positions and distances are normalized by the sight range. The global state representation, action-availability mechanism, environment dynamics, and reward function remain unchanged. The remaining settings are given in Table 3, and the same hyperparameter configuration is adopted for GRF.

Table 3: Hyperparameters for SMAC

Parameter	Value
Optimizer	Adam
Learning rate	5×10^{-4}
Discount factor (γ)	0.99
Batch size	32
Replay buffer size	5,000
Target update interval	200 episodes
Number of GMM components	8
Number of perturbation iterations L	5
Perturbation iteration step size	0.1
Smoothing loss weight λ_{csp}	10
λ_{KD}	10
λ_{sample}	1
SC2 difficulty	7

C.2.2 Experimental Results

Table 4 reports detailed student success rates on the four SMAC tasks. Each entry reports the mean and standard deviation over five seeds based on the final ten evaluations, with training limited to 2M steps for difficult tasks and 3M steps for super-hard tasks. Teacher–Student Gap denotes the task-averaged student performance change relative to the corresponding teacher.

Table 4: Detailed success rates in the SMAC experiment.

Algorithm	`8m_vs_9m`	`5m_vs_6m`	`6h_vs_8z`	`3s5z_vs_3s6z`	Average	Teacher Gap
CSD	$\mathbf{89.8\% \pm 1.6\%}$	$51.1\% \pm 11.0\%$	$\mathbf{65.3\% \pm 8.6\%}$	$\mathbf{97.0\% \pm 1.7\%}$	$\mathbf{75.8\%}$	$\uparrow 2.6\%$
CADP	$47.1\% \pm 10.2\%$	$28.4\% \pm 7.8\%$	$3.9\% \pm 4.0\%$	$79.8\% \pm 8.8\%$	39.8%	$\downarrow 18.0\%$
CTDS	$65.1\% \pm 36.7\%$	$46.8\% \pm 10.9\%$	$0.1\% \pm 0.2\%$	$0.6\% \pm 0.6\%$	28.1%	$\downarrow 6.0\%$
QMIX	$57.1\% \pm 13.0\%$	$\mathbf{55.3\% \pm 16.0\%}$	$0.6\% \pm 0.7\%$	$0.1\% \pm 0.2\%$	28.3%	–
HASAC	$45.3\% \pm 13.3\%$	$24.0\% \pm 10.0\%$	$0.7\% \pm 1.2\%$	$1.3\% \pm 2.3\%$	17.8%	–
PTDE	$64.2\% \pm 13.7\%$	$25.0\% \pm 14.1\%$	$0.8\% \pm 1.9\%$	$0.8\% \pm 1.9\%$	22.7%	$\downarrow 6.9\%$

Teacher–student comparison. Figure A2 complements Table 4 by directly comparing the success-rate distributions of each communication teacher and its CTDE student. Existing transfer methods exhibit pronounced teacher–student gaps on several tasks, indicating that strong communication policies do not necessarily yield equally strong decentralized policies after transfer. CSD maintains a substantially smaller gap overall, with an average student improvement of 2.6% relative to its teacher across the four tasks. The result on `6h_vs_8z` also shows that student performance remains constrained when the communication teacher itself learns a weak policy, highlighting the importance of preserving teacher performance while reducing the discrepancy between communication and decentralized execution.

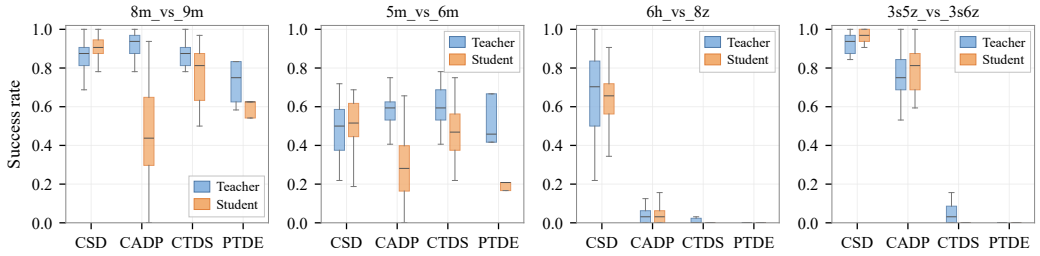


Figure A2: Distributions of teacher and student success rates on four SMAC tasks.

Plug-and-play comparison. Figures A3 and A4 provide complementary views of the plug-and-play experiments. Figure A3 compares the final performance distributions of VDN, QMIX, and Qatten with their CSD-enhanced counterparts. Additionally, we present the evaluation performance throughout training in the plug-and-play experiments in Figure A4. Incorporating CSD consistently improves all three value-decomposition methods on both tasks. The improvement is especially clear on `6h_vs_8z`: the original methods rarely solve the task, whereas the corresponding CSD variants achieve substantial success rates. These results show that the communication-smoothing and distillation procedure is not tied to a particular mixing architecture and can provide useful learning targets for different value-decomposition methods.

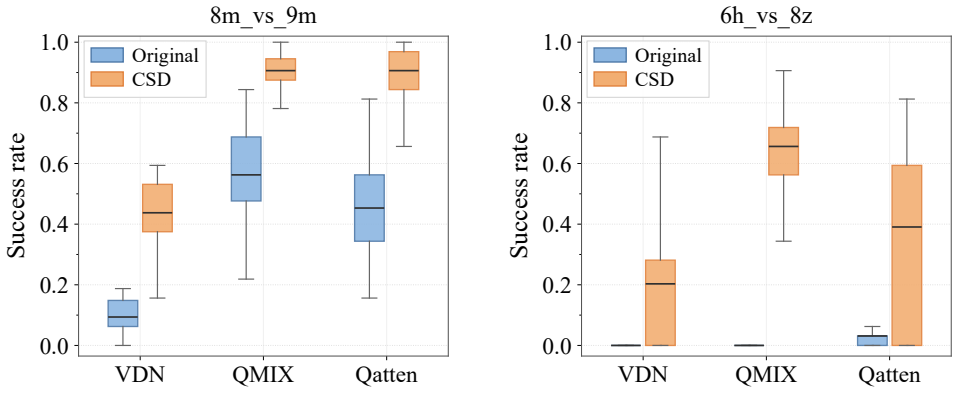


Figure A3: Evaluation of integrating CSD with different value-decomposition methods.

Ablation. Figure A5 examines the effects of the conditional distribution model and the perturbation strategy. CSD-V replaces the GMM with a VAE, whereas CSD-R replaces gradient-based worst-case perturbation search with random sampling in the 3σ region. Replacing the GMM produces a moderate performance reduction on both tasks, suggesting that explicitly modeling multimodal cross-agent observation distributions is beneficial. Random perturbations remain effective on `6h_vs_8z`, but lead to lower performance on `8m_vs_9m`.

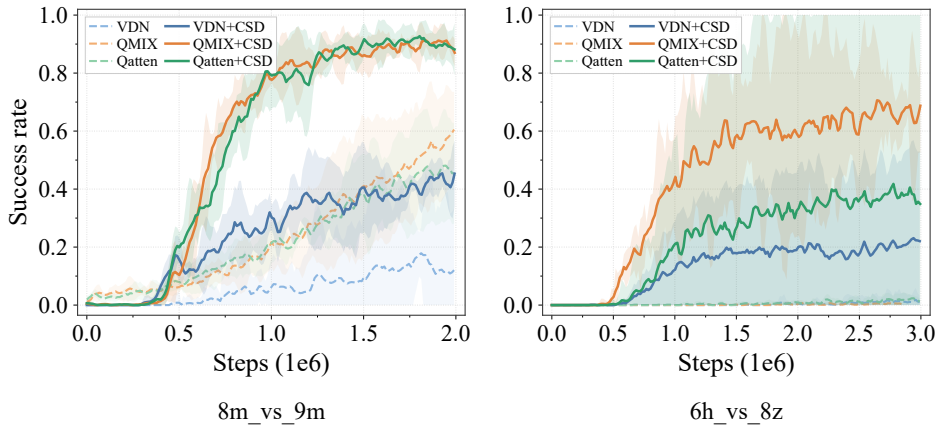


Figure A4: Plug-and-play evaluation on SMAC.

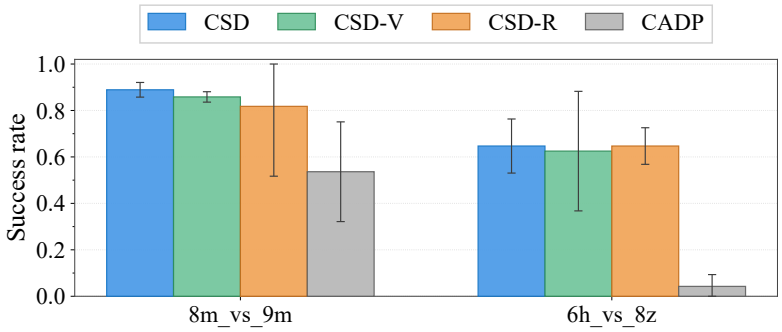


Figure A5: Ablation results on SMAC tasks.

C.3 Google Research Football

C.3.1 Environment settings

We evaluate CSD on two GRF Football Academy scenarios, `academy_3_vs_1_with_keeper` and `academy_run_pass_and_shoot_with_keeper`. In the former, three agents cooperate in a 3-vs-1 attacking task; in the latter, two agents cooperate to run, pass, and shoot. Both scenarios include a goalkeeper.

The default observations in GRF provide each controlled player with global information about the positions and movement directions of all entities. To align GRF with the partially observable setting considered in this work, we modify the observation construction in the underlying GRF wrapper by masking entities outside a predefined visibility range. Specifically, the visibility threshold defines the maximum relative distance within which other players can be observed in the normalized pitch coordinate space. It is set to 0.05 for both GRF tasks. Teammates and opponents beyond this range are masked, while the ball remains observable.

C.3.2 Experimental Results

Detailed results for the two GRF tasks are summarized in Table 5, complementing the performance curves presented in the main text. Each entry reports the mean and standard deviation over the available seeds based on the final ten evaluations within 2M environment steps, while PTDE is evaluated using its available final records.

Table 5: Detailed student success rates on both GRF tasks.

Algorithm	academy_3_vs_1_with_keeper	academy_run_pass_and_shoot_with_keeper	Average
CSD	$\mathbf{55.8\% \pm 12.6\%}$	$61.3\% \pm 33.7\%$	$\mathbf{58.5\%}$
CADP	$28.8\% \pm 20.0\%$	$\mathbf{63.1\% \pm 28.1\%}$	46.0%
CTDS	$4.6\% \pm 2.0\%$	$1.8\% \pm 2.3\%$	3.2%
HASAC	$27.3\% \pm 20.0\%$	$10.0\% \pm 6.9\%$	18.7%
PTDE	$2.8\% \pm 2.4\%$	$0.0\% \pm 0.0\%$	1.4%

D Supplementary Related Work

D.1 CTDE

Early cooperative MARL frameworks are typically divided into centralized, decentralized, and CTDE paradigms Gronauer & Diepold (2022). Centralized learning, reliant on a joint policy, faces challenges under partial observability and communication constraints Claus & Boutilier (1998); Liu et al. (2024). Decentralized learning permits independent optimization but introduces non-stationarity Tan (1993). The CTDE framework combines their strengths, enabling agents to use global information during training while executing in a decentralized manner based on local observations.

Within the CTDE framework, value decomposition methods address the multi-agent credit assignment problem by factorizing the joint action-value function into a set of individual utilities, such that decentralized action selection can remain consistent with centralized value estimation. Following the additive decomposition introduced by VDN Sunehag et al. (2018), subsequent methods improve the expressiveness and optimization of value factorization from different perspectives. QMIX Rashid et al. (2020b) introduces a state-conditioned monotonic mixing network to model nonlinear relationships between individual utilities and the joint action value. Qatten Yang et al. (2020) further employs attention mechanisms to dynamically estimate the contributions of different agents, while Weighted QMIX Rashid et al. (2020a) reweights the training objective to improve the approximation of optimal joint actions.

In parallel with value decomposition, policy gradient methods within the CTDE framework directly optimize decentralized policies using centralized value estimation. MADDPG Sheikh & Bölöni (2020) extends deterministic policy gradients to multi-agent settings by equipping each agent with a centralized critic conditioned on the observations and actions of all agents. COMA Foerster et al. (2017) further addresses multi-agent credit assignment through a counterfactual baseline that evaluates an agent's contribution while marginalizing over its own actions and holding the actions of the other agents fixed. More recently, MAPPO Yu et al. (2022) adapts proximal policy optimization to cooperative MARL by combining decentralized policies with centralized value functions and clipped policy updates. HASAC Liu et al. (2024) incorporates maximum-entropy reinforcement learning into the heterogeneous-agent framework, jointly exploiting stochastic policies, entropy regularization, and sequential policy updates to promote exploration, stabilize training, and enhance robustness.

D.2 Communication in MARL

In MARL, communication facilitates information sharing among agents, mitigating partial observability and promoting cooperation. A major research line leverages attention mechanisms to structure message passing, where cross-attention serves as the communication channel, as exemplified by DGN Jiang et al. (2020), TarMAC Das et al. (2019), and MAGIC Niu et al. (2021). Another line of work models agent interactions via coordination graphs. Unlike communication-based methods that allow information exchange during both training and execution, the CTDE paradigm relies on centralized training while disallowing communication at execution.

Existing communication-based MARL studies mainly improve communication efficiency through message compression, pruning, selective transmission, single-round accelerated optimization, or privacy-budgeted advising Wang et al. (2020); Ding et al. (2020); Sun et al. (2024); Ding et al. (2024); Ye et al. (2020). Although these methods support execution under extremely sparse communication,

they still rely on inter-agent message exchange and therefore cannot operate when communication is completely denied.

D.3 COMMUNICATION POLICY TRANSFER IN MARL

Conventional CTDE methods may suffer from inefficient joint policy exploration and limited coordination because global information available during centralized training is not fully exploited. To address this limitation, recent studies have explored communication policy transfer, using communication-enabled policies to guide the learning of decentralized CTDE policies. Representative approaches include IGM-DA Hong et al. (2022), which distills global knowledge from a centralized teacher to decentralized students; CTDS Zhao et al. (2024), which employs a teacher with an infinite field of view; PTDE Chen et al. (2024), which integrates local and global representations to facilitate knowledge transfer; and, more recently, CADP Zhou et al. (2025), which incorporates a communication-pruning mechanism to progressively reduce the student's dependence on communication.

Despite their competitive performance on benchmarks like SMAC, they generally overlook the inherent policy discrepancies between communication-based and decentralized agents, which can hinder knowledge transfer and limit cooperative exploration in certain scenarios.