MULTI-AGENT INFORMATION-GUIDED POLICY OPTIMIZATION: AN INFORMATION-THEORETIC PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

In multi-agent reinforcement learning (MARL), partial observability limits the ability of decentralized agents to leverage global information, motivating widely used frameworks such as Centralized Training with Decentralized Execution (CTDE) and more recently, Multi-Agent Guided Policy Optimization (MAGPO). However, MAGPO relies solely on policy alignment and fails to explicitly align decision-relevant information between the guider and learner, limiting decentralized policy learning. In this work, we revisit the *imitation gap* between guider and learner from an information-theoretic perspective, and show that it fundamentally arises from an *information gap*—a mismatch in decision-relevant information between centralized and decentralized policies. To mitigate this gap, we propose *Multi-Agent Information-Guided Policy Optimization (MAIGPO)*, which employs the information bottleneck principle to extract decision-relevant representations and mutual information maximization to align them between the guider and learners through bidirectional information flow. By encouraging learners to capture the guider's decision rationale and constraining the guider to learner-accessible information, MAIGPO significantly improves decentralized policy learning and outperforms state-of-the-art methods on standard benchmarks.

1 INTRODUCTION

Multi-Agent Reinforcement Learning (MARL) has attracted growing attention for complex cooperative decision-making problems, including autonomous driving (Zhou et al., 2020; Schmidt et al., 2022), energy management (Wang et al., 2021), and robotic coordination (Hüttenrauch et al., 2019). Fully cooperative MARL, where agents coordinate toward a shared objective, is the most widely studied setting (Lowe et al., 2017). Directly treating such problems as single-agent decision-making leads to exponentially growing joint state and action spaces as the number of agents increases, resulting in prohibitive computational cost and poor scalability (Rizk et al., 2018; Nguyen et al., 2020). To address this issue, fully cooperative MARL is commonly formulated as a *Decentralized Partially Observable Markov Decision Process* (Dec-POMDP) (Oliehoek et al., 2016), where the joint policy is factorized into per-agent policies. This factorization improves scalability and enables decentralized execution based on local observations (Sunehag et al., 2017; Yu et al., 2022), but inevitably introduces partial observability, causing agents to lose decision-relevant information available to centralized policies (Chen et al., 2021; Wen et al., 2022; Mahjoub et al., 2024).

To fully exploit global information, Centralized Training with Centralized Execution (CTCE) learns centralized policies conditioned on the global state or joint observations and agents' previous actions. Representative methods, such as MAT and Sable (Wen et al., 2022; Mahjoub et al., 2024), leverage this centralized information structure to support advantage decomposition and sequential policy updates. However, fully centralized execution is often impractical in real-world multi-agent systems due to limited scalability and its reliance on global information at execution time (Li et al., 2025b). To address these limitations, Centralized Training with Decentralized Execution (CTDE) trains agents using centralized information while allowing each agent to act solely based on its local observations during execution. These approachs preserve the benefits of centralized training, such as access to global states and joint actions, while enabling scalable decentralized deployment. CTDE forms the basis of many widely adopted MARL algorithms, including QMIX (Rashid et al., 2020),

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.

MAPPO (Yu et al., 2022), and HAPPO (Zhong et al., 2024), providing a trade-off between coordination performance and practical scalability. Nevertheless, in many CTDE methods, centralized information is primarily incorporated through the critic, which may underutilize global information in decentralized policy learning and provides limited guarantees of optimal coordination.

More recently, the Centralized Teacher with Decentralized Students (CTDS) framework has been explored, extending the teacher-student paradigm from single-agent to multi-agent systems (Zhao et al., 2022; Zhou et al., 2023). In CTDS, a centralized teacher with global information guides decentralized agents, improving coordination and exploration (Chen et al., 2022). Although CTDS provides richer supervision than conventional CTDE, decentralized students often fail to fully reproduce the teacher's policy due to observation asymmetry and policy-space mismatch, resulting in the *imitation gap* (Weihs et al., 2021; Li et al., 2025b). Building upon these insights, MAGPO introduces a guider-learner structure, including a Guider Backtracking mechanism that projects centralized policies back toward the decentralized policies, thereby alleviating the imitation gap (Li et al., 2025b). This alignment allows the guider to leverage centralized information while remaining close to decentralized learner policies, improving performance across diverse environments.

Despite their effectiveness, existing CTDS methods, including MAGPO, fail to enable sufficient bidirectional exchange of task-relevant information between centralized policies and decentralized agents, as illustrated in Fig. 1. First, guider–learner interaction is largely restricted to action-level alignment, without explicitly transferring the decision-relevant information that supports coordinated behavior. As a result, learners may imitate actions without acquiring the essential decision-making structure encoded in the centralized policy. Second, MAGPO relies on Guider Backtracking to keep the guider close to learner policies, making the quality of guidance dependent on learner adaptability. When learners improve slowly, this constraint can anchor the guider to suboptimal policies, limiting exploration and ultimately restricting performance.

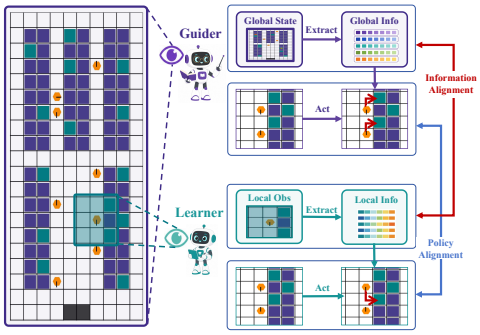

Figure 1: Motivation of MAIGPO. The guider–learner imitation gap stems from an information gap that policy alignment alone may not fully mitigate. MAIGPO further mitigates this gap through bidirectional information alignment.

Motivated by these observations, we revisit the guider–learner *imitation gap* from an information-theoretic perspective (Shannon, 1948; MacKay, 2003). We show that it fundamentally stems from an *information gap*, namely, a mismatch in decision-relevant information between centralized and decentralized policies. From this perspective, asymmetries in observation and policy spaces (Li et al., 2025b) are manifestations of the same underlying gap. Bridging this gap therefore requires extracting and aligning decision-relevant information between the guider and learners. To this end, we propose *Multi-Agent Information-Guided Policy Optimization (MAIGPO)*, which extracts decision-relevant representations through an information bottleneck and aligns them via mutual information maximization, enabling knowledge transfer beyond action-level imitation. MAIGPO encourages the guider to rely on learner-accessible information while enabling learners to acquire richer decision representations, thereby improving learner adaptability and promoting more effective joint optimization of the guider and learners. We evaluate MAIGPO across 23 tasks spanning 2 diverse environments. Experimental results demonstrate that MAIGPO consistently outperforms strong baseline methods, achieving superior coordination performance and learning efficiency.

2 Related Works

Centralized Training and Teacher-Guided MARL. Centralized training is widely used in cooperative MARL to exploit global information for coordinated decision-making. Centralized Training with Centralized Execution (CTCE) learns globally conditioned policies, as exemplified by MAT (Wen et al., 2022) and Sable (Mahjoub et al., 2024), but requires centralized information during execution. Centralized Training with Decentralized Execution (CTDE) instead restricts ex-

ecution to local policies while using global information during training. Representative methods include QMIX (Rashid et al., 2020), MAPPO (Yu et al., 2022), and HAPPO (Zhong et al., 2024). More recently, Centralized Teacher with Decentralized Students (CTDS) introduces a centralized policy to directly guide decentralized agents (Zhao et al., 2022; Zhou et al., 2023). MAGPO (Li et al., 2025a;b) further develops this paradigm through Guider Backtracking, which regulates the discrepancy between centralized and decentralized policies. Existing CTDS studies primarily focus on policy-level guidance, whereas information-level alignment between centralized and decentralized policies remains less explored. More details are provided in Appendix A.

Information-Theoretic Representation Learning. Information-theoretic principles provide a general framework for learning compact and informative representations (MacKay, 2003). In particular, the information bottleneck principle preserves task-relevant information while suppressing irrelevant input variations (Tishby et al., 2000; Alemi et al., 2016). Mutual information quantifies the shared information between representations and has been widely used for representation learning and alignment (Belghazi et al., 2018; MacKay, 2003; Wang et al., 2023). InfoNCE (Oord et al., 2018) provides a tractable lower bound on mutual information and underpins contrastive representation learning (Chen et al., 2020; He et al., 2020; Gao et al., 2021). In reinforcement learning, information-theoretic objectives have been extensively studied for intrinsic exploration (Li et al., 2021; Liu et al., 2023; Wang et al., 2019; Li et al., 2024b; Jo et al., 2024; Sukhija et al., 2024). Contrastive learning has also been used to learn structured representations (Yang et al., 2023; Patil et al., 2024) and promote behavioral diversity (Li et al., 2024a). In contrast, our work views the guider–learner imitation gap as arising from an underlying decision-relevant information gap and mitigates it by extracting and bidirectionally aligning decision-relevant representations between the centralized guider and decentralized learners, moving beyond action-level imitation.

3 PRELIMINARIES

Decentralized POMDP. We consider fully cooperative multi-agent reinforcement learning problems modeled as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP) Oliehoek et al. (2016), defined by $M = \langle \mathcal{N}, \mathcal{S}, \boldsymbol{\mathcal{A}}, \mathcal{P}, \mathcal{R}, \boldsymbol{\mathcal{O}}, \gamma \rangle$. Here, $\mathcal{N} = \{1, \dots, n\}$ is the set of agents, $\mathcal{S}$ is the global state space, and $\boldsymbol{\mathcal{A}} = \mathcal{A}^1 \times \cdots \times \mathcal{A}^n$ is the joint action space. $\gamma \in [0, 1)$ denotes the discount factor. At time step t, each agent i receives a local observation $o_t^i \in \mathcal{O}^i$, and selects an action $a_t^i \in \mathcal{A}^i$. The joint action $\boldsymbol{a}_t = (a_t^1, \dots, a_t^n)$ induces a transition $\mathcal{P}(s_{t+1} \mid s_t, \boldsymbol{a}_t)$, and all agents receive a shared reward $r_t = \mathcal{R}(s_t, \boldsymbol{a}_t)$.

Due to partial observability, each agent follows a decentralized policy $\pi^i(a_t^i \mid o_t^i)$, yielding the factorized joint policy $\boldsymbol{\pi}(\boldsymbol{a}_t \mid \boldsymbol{o}_t) = \prod_{i=1}^n \pi^i(a_t^i \mid o_t^i)$, where $\boldsymbol{o}_t = (o_t^1, \dots, o_t^n)$. The objective is to maximize $J(\boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\pi}, \mathcal{P}}[\sum_{t=0}^{\infty} \gamma^t r_t]$. In practice, decentralized policies are typically implemented with an RNN or GRU that encodes the local observation history into a recurrent hidden state, from which the action distribution is generated. For simplicity, we denote the policy as $\pi^i(a_t^i \mid o_t^i)$, where o_t^i includes the history-dependent local representation.

Centralized Guider and Decentralized Learners. Following the CTDS framework (Zhao et al., 2022; Zhou et al., 2023; Li et al., 2025a;b), we consider a centralized guider available only during training and decentralized learners retained for execution. Conditioned on the global state, the guider autoregressively factorizes the joint policy as $\pi_g(\boldsymbol{a}_t \mid s_t) = \prod_{i=1}^n \pi_g^i(a_t^i \mid s_t, \boldsymbol{a}_t^{1:i-1})$, where $\boldsymbol{a}_t^{1:i-1}$ denotes the actions generated for preceding agents. In contrast, learner i follows a local policy $\pi_l^i(a_t^i \mid o_t^i)$, yielding the decentralized joint policy $\pi_l(\boldsymbol{a}_t \mid \boldsymbol{o}_t) = \prod_{i=1}^n \pi_l^i(a_t^i \mid o_t^i)$.

Information-Theoretic Foundations. For random variables X and Y, mutual information is defined as $I(X;Y) = \mathbb{E}_{p(x,y)}\left[\log \frac{p(x,y)}{p(x)p(y)}\right]$, which quantifies their statistical dependence (Shannon, 1948; MacKay, 2003). Given an input X, representation Z, and target Y, the information bottleneck (IB) principle learns compact yet task-relevant representations by minimizing $\mathcal{L}_{\mathrm{IB}} = \beta I(X;Z) - I(Z;Y)$ (Tishby et al., 2000). The variational information bottleneck (VIB) makes this objective tractable by approximating the compression term with $\mathbb{E}_{p(x)} D_{\mathrm{KL}}(q_\phi(z \mid x) \| p(z))$ and the predictive term with a variational likelihood (Alemi et al., 2016). To estimate mutual information tractably, InfoNCE provides the lower bound $I(X;Y) \geq \log N - \mathcal{L}_{\mathrm{InfoNCE}}$, where N denotes the number of samples in the contrastive objective (Oord et al., 2018).

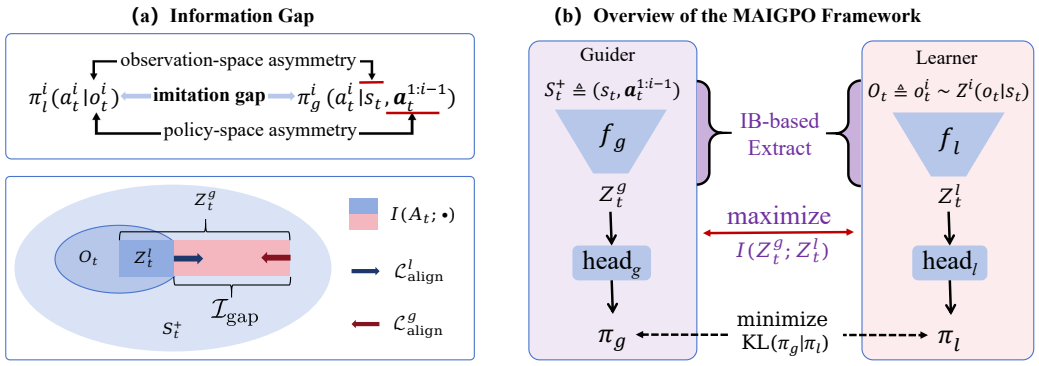

Figure 2: **Information gap and overview of MAIGPO.** (a) The imitation gap is fundamentally an information gap, with observation and policy asymmetries as its manifestations. (b) MAIGPO extracts decision-relevant representations and aligns guider–learner information by maximizing $I(Z_t^g; Z_t^l)$, enabling bidirectional knowledge transfer beyond policy alignment.

4 METHOD

In this work, we revisit guider–learner frameworks from an information-theoretic perspective and show that asymmetries in observation and policy spaces can be viewed as manifestations of a decision-relevant information gap. To mitigate this gap, MAIGPO extracts and aligns decision-relevant representations of the guider and learner. The overall framework is illustrated in Fig. 2.

4.1 DECISION-RELEVANT INFORMATION GAP

Problem Description and Notation. We consider a cooperative multi-agent system with N agents, modeled as a Dec-POMDP. To characterize the decision-relevant information gap, we focus on an individual agent i and compare its centralized guider policy and decentralized learner policy. Specifically, agent i is associated with a centralized guider policy $\pi_g^i(a_t^i \mid s_t, \boldsymbol{a}_t^{1:i-1})$ and a decentralized learner policy $\pi_l^i(a_t^i \mid o_t^i)$ For notational convenience, we define the *augmented global state* and *local observation* as $S_t^+ \triangleq (s_t, \boldsymbol{a}_t^{1:i-1})$, $O_t \triangleq o_t^i \sim \mathcal{O}^i(o_t \mid s_t)$, and let A_t denote the optimal action. Both the guider and the learner aim to learn mappings to the target action A_t:

$$\text{Guider:} \quad Z_t^g = f_g(S_t^+), \quad A_t \sim \pi_g(\cdot \mid Z_t^g),$$
$$\text{Learner:} \quad Z_t^l = f_l(O_t), \quad A_t \sim \pi_l(\cdot \mid Z_t^l),$$

where f_g and f_l denote the feature extractors for the guider and learner, respectively, and Z_t^g and Z_t^l are the corresponding latent representations. The guider is trained via a standard reinforcement learning loss, while the learner is trained through a imitation loss, for example by minimizing the KL divergence between the guider and learner policies.

Markov Structure and Information Flow. We characterize the guider and learner through the following information-theoretic Markov structures (Shannon, 1948; MacKay, 2003; Tishby et al., 2000):

$$\begin{array}{lccccc} \text{Guider:} & A_t & \longrightarrow & S_t^+ & \longrightarrow Z_t^g & \rightsquigarrow \hat{A}_t \\ & & & \downarrow & & \\ \text{Learner:} & A_t & \longrightarrow & O_t & \longrightarrow Z_t^l & \rightsquigarrow \hat{A}_t \end{array}$$

The horizontal paths $A_t \to S_t^+ \to Z_t^g$ and $A_t \to O_t \to Z_t^l$ represent information-theoretic Markov relations rather than causal generation. Specifically, they imply $A_t \perp Z_t^g \mid S_t^+$ and $A_t \perp Z_t^l \mid O_t$. The $\rightsquigarrow$ arrows denote approximate mappings from the learned representations to the target action distribution, while the vertical arrow indicates that O_t is generated from S_t^+.

We first characterize the information asymmetry at the input level. Since $A_t \to S_t^+ \to O_t$ forms an information-theoretic Markov chain, the data processing inequality gives

$$I(A_t; O_t) \le I(A_t; S_t^+). \tag{1}$$

Thus, the learner's local observation contains no more information about the target action than the augmented global state available to the guider. The latent representations are extracted from their corresponding inputs, yielding the Markov relations $A_t \to S_t^+ \to Z_t^g$ and $A_t \to O_t \to Z_t^l$. Accordingly,

$$I(A_t; Z_t^g) \le I(A_t; S_t^+), \qquad I(A_t; Z_t^l) \le I(A_t; O_t). \tag{2}$$

To formalize both the decision-relevant information asymmetry and the corresponding information gap, we introduce the following idealized encoding assumption.

Assumption 1 (Ideal Decision-Relevant Encoding)**.** *The representations Z_t^g and Z_t^l preserve all and only decision-relevant information from their respective inputs:*

$$\begin{aligned} I(A_t; S_t^+ \mid Z_t^g) = 0, \quad I(A_t; O_t \mid Z_t^l) = 0, \\ I(S_t^+; Z_t^g \mid A_t) = 0, \quad I(O_t; Z_t^l \mid A_t) = 0. \end{aligned} \tag{3}$$

Proposition 1 (Decision-Relevant Information Asymmetry)**.** *Under Assumption 1 and the Markov relations established above, the following information ordering holds:*

$$I(A_t; Z_t^l) = I(A_t; O_t) \le I(A_t; S_t^+) = I(A_t; Z_t^g). \tag{4}$$

Proposition 2 establishes the inherent *decision-relevant information asymmetry* between the centralized guider and decentralized learner. Even under ideal encoding, the learner remains limited by the decision-relevant information available in its local observation. We therefore define the corresponding information gap as follows.

Definition 1 (Decision-Relevant Information Gap)**.** *Under Assumption 1 and the established Markov relations, the guider-to-learner decision-relevant information gap is defined as*

$$\mathcal{I}_{\text{gap}} \triangleq I(A_t; Z_t^g) - I(A_t; Z_t^l) = I(A_t; Z_t^g \mid Z_t^l). \tag{5}$$

The gap measures the additional decision-relevant information contained in the guider representation but unavailable to the learner. It motivates information alignment: transferring decision-relevant knowledge to the learner while encouraging the guider to rely on learner-accessible information.

Information-Gap Decomposition. Direct optimization of $\mathcal{I}_{\text{gap}}$ is intractable, as the target optimal action A_t is not directly accessible. Applying the chain rule of mutual information yields

$$\begin{aligned} \mathcal{I}_{\text{gap}} &= I(A_t; Z_t^g \mid Z_t^l) \\ &= \underbrace{I(Z_t^g; A_t)}_{\text{decision information}} - \underbrace{I(Z_t^g; Z_t^l)}_{\text{shared information}} + \underbrace{\cancel{I(Z_t^g; Z_t^l \mid A_t)}}_{\text{residual dependence}}. \end{aligned} \tag{6}$$

Under the dependency structure $Z_t^g \leftarrow S_t^+ \to O_t \to Z_t^l$ and Assumption 1 , the conditional data processing inequality yields $I(Z_t^g; Z_t^l \mid A_t) = 0$. The term $I(Z_t^g; A_t)$ measures the decision-relevant information encoded in the guider representation and is promoted by optimizing the guider task objective $\mathcal{L}_{\text{task}}^g$. Under ideal encoding, minimizing the information gap therefore reduces to maximizing the decision-relevant information shared between the guider and learner representations. Even when the ideal assumption is only approximately satisfied, the guider typically retains more decision-relevant information than the learner, while both representations become increasingly action-relevant during training, making $I(Z_t^g; Z_t^l \mid A_t)$ small. Thus, maximizing $I(Z_t^g; Z_t^l)$ remains effective for mitigating the information gap. Further analysis is provided in Appendix B.

4.2 Information-Guided Policy Optimization

To mitigate the decision-relevant information gap, we propose Multi-Agent Information-Guided Policy Optimization (MAIGPO), which explicitly extracts and aligns the latent representations of centralized guiders and decentralized learners. Alongside the original task objectives $\mathcal{L}_{\text{task}}^g$ and $\mathcal{L}_{\text{task}}^l$, MAIGPO imposes two constraints on their intermediate representations: extracting compact decision-relevant representations, and maximizing the mutual information between the guider and learner representations.

Decision-Relevant Representation Extraction. We employ the variational information bottleneck (VIB) (Alemi et al., 2016) to extract compact decision-relevant representations for both the guider and learner, thereby approximating the ideal representations characterized in Assumption 1.

$$\mathcal{L}_{\mathrm{VIB}} = \mathbb{E}\left[-\log q_\phi(Y \mid Z) + \beta D_{\mathrm{KL}}\left(p_\theta(Z \mid X) \,\|\, r(Z)\right)\right]. \tag{7}$$

Here, $p_\theta(Z \mid X)$ denotes the stochastic encoder, $q_\phi(Y \mid Z)$ the task predictor, $r(Z)$ the prior, and β the compression coefficient. By balancing task prediction and information compression, VIB preserves decision-relevant information while suppressing task-irrelevant redundancy (Tishby et al., 2000; Alemi et al., 2016; Igl et al., 2019).

In our setting, we instantiate Gaussian encoders centered at the feature representations $Z_t^g = f_g(S_t^+)$ and $Z_t^l = f_l(O_t)$, with $p_{\theta_g}(\widetilde{Z}_t^g \mid S_t^+) = \mathcal{N}\big(Z_t^g, \mathrm{diag}((\sigma_t^g)^2)\big)$ and $p_{\theta_l}(\widetilde{Z}_t^l \mid O_t) = \mathcal{N}\big(Z_t^l, \mathrm{diag}((\sigma_t^l)^2)\big)$, where σ_t^g and σ_t^l denote the learned standard deviations. Thus, (X, Z, Y) in Eq. equation 7 corresponds to $(S_t^+, \widetilde{Z}_t^g, A_t)$ for the guider and $(O_t, \widetilde{Z}_t^l, A_t)$ for the learner, while $q_\phi(Y \mid Z)$ corresponds to the action policies $\pi_g(A_t \mid \widetilde{Z}_t^g)$ and $\pi_l(A_t \mid \widetilde{Z}_t^l)$, respectively. The predictive term is instantiated by the corresponding decision objective: a reinforcement learning loss for the guider and an imitation loss for the learner; see Appendix A.2 for details. Accordingly,

$$\begin{aligned} \mathcal{L}_{\mathrm{VIB}}^g &= \mathcal{L}_{\mathrm{task}}^g + \beta_g D_{\mathrm{KL}}\left(p_{\theta_g}(\cdot \mid S_t^+) \,\|\, r_g\right), \\ \mathcal{L}_{\mathrm{VIB}}^l &= \mathcal{L}_{\mathrm{task}}^l + \beta_l D_{\mathrm{KL}}\left(p_{\theta_l}(\cdot \mid O_t) \,\|\, r_l\right), \end{aligned} \tag{8}$$

where $\mathcal{L}_{\mathrm{task}}^g$ and $\mathcal{L}_{\mathrm{task}}^l$ are the original guider and learner objectives in MAGPO, β_g and β_l control the compression strength, and r_g and r_l are standard Gaussian priors (Kingma & Welling, 2013).

Bidirectional Information Alignment. To mitigate the decision-relevant information gap, we maximize the mutual information between the guider and learner representations:

$$\mathcal{L}_{\mathrm{align}} = -I(Z_t^g; Z_t^l), \qquad \min_{\theta_g, \theta_l} \mathcal{L}_{\mathrm{align}}, \tag{9}$$

where θ_g and θ_l denote the parameters of the guider and learner encoders, respectively. Since mutual information is difficult to estimate directly, we employ InfoNCE as a tractable contrastive lower bound (Hjelm et al., 2018; Chen et al., 2020). InfoNCE has been widely used to align representations across modalities and temporal contexts (Radford et al., 2021; Lan et al., 2025). We therefore use it to perform bidirectional alignment between the centralized guider and decentralized learner representations. Since Z_t^g and Z_t^l lie in different representation spaces, we introduce learnable projection heads h^g and h^l:

$$H_t^g = h^g(Z_t^g), \qquad H_t^l = h^l(Z_t^l), \tag{10}$$

where h^g projects the guider representation into the learner representation space, while h^l projects the learner representation into the guider representation space. This symmetric design enables bidirectional contrastive alignment and mutual representation learning between the guider and learner.

Guider-side information alignment. We define a compact InfoNCE loss:

$$\mathcal{L}_{\mathrm{align}}^g = -\mathbb{E}\left[\log \frac{e^{f(H_t^g, \bar{Z}_t^l)}}{e^{f(H_t^g, \bar{Z}_t^l)} + \sum_k e^{f(H_t^g, \bar{Z}_{t,k}^{l-})}}\right], \tag{11}$$

where $f(\cdot, \cdot)$ denotes cosine similarity scaled by a temperature $\tau > 0$, and $\{H_t^g, \bar{Z}_{t,k}^{l-}\}_{k=1}^K$ are negative pairs. The bars indicate gradient truncation. By contrasting the pair $(H_t^g, \bar{Z}_t^l)$ with mismatched pairs $(H_t^g, \bar{Z}_{t,k}^{l-})$, the guider is encouraged to emphasize decision-relevant features accessible to the learner. It acts as a *learner-aware teacher* and adapts its representation to the learner's capacity.

Learner-side information alignment. The loss is

$$\mathcal{L}_{\mathrm{align}}^l = -\mathbb{E}\left[\log \frac{e^{f(H_t^l, \bar{Z}_t^g)}}{e^{f(H_t^l, \bar{Z}_t^g)} + \sum_k e^{f(H_t^l, \bar{Z}_{t,k}^{g-})}}\right], \tag{12}$$

where $\{H_t^l, \bar{Z}_{t,k}^{g-}\}_{k=1}^K$ are negative pairs. By treating the guider representation as a fixed alignment target, this loss updates the learner to capture not only which actions to take but also the decision-relevant rationale, fostering a *more intelligent student* that learns beyond action-level imitation.

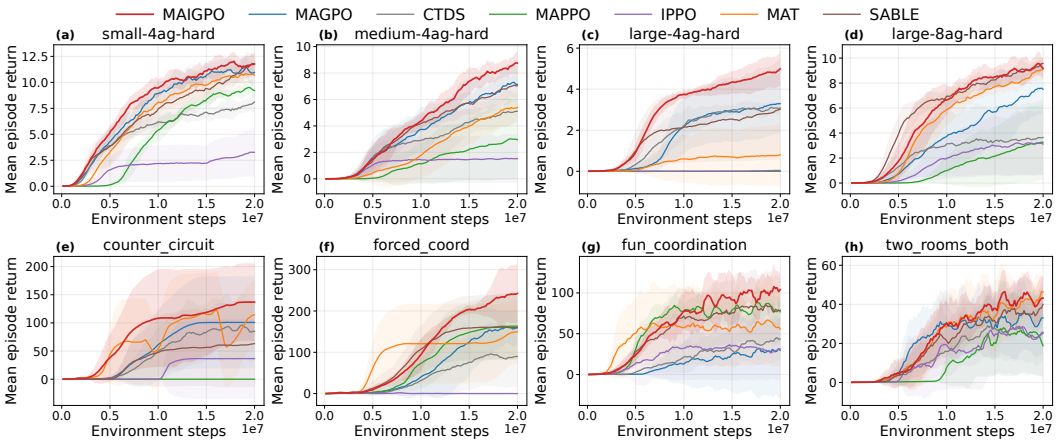

Figure 3: Learning curves on representative tasks. MAT and SABLE are CTCE methods. (a–d) RobotWarehouse. (e–h) OvercookedV2. Full task-wise results are provided in Appendix E.2.

Together, the two alignment losses provide a tractable bidirectional surrogate for mutual information maximization: the guider adapts toward learner-accessible information, while the learner captures decision-relevant structures from the guider. This alignment mitigates the information gap and promotes coordinated learning between the guider and learner. See Appendix C for details.

Joint Policy Optimization. MAIGPO jointly optimizes the original MAGPO objectives and representation-level information alignment:

$$\begin{aligned} \mathcal{L}_g &= \mathcal{L}^g_{\text{VIB}} + \lambda_g \mathcal{L}^g_{\text{align}}, \\ \mathcal{L}_l &= \mathcal{L}^l_{\text{VIB}} + \lambda_l \mathcal{L}^l_{\text{align}}. \end{aligned} \tag{13}$$

Here, $\mathcal{L}^g_{\text{VIB}}$ and $\mathcal{L}^l_{\text{VIB}}$ augment the corresponding MAGPO objectives, including reinforcement learning and action-level alignment, with information-bottleneck regularization. The terms $\mathcal{L}^g_{\text{align}}$ and $\mathcal{L}^l_{\text{align}}$ perform bidirectional representation alignment, with λ_g and λ_l controlling their strengths.

5 Experiments

We compare MAIGPO with SOTA MARL methods under CTDE, CTCE, and CTDS settings, and include IPPO (Tan, 1993) as a fully decentralized baseline. See Appendix E.1 for hardware details.

5.1 Experimental Setup

Baselines. We consider two representative CTCE methods, SABLE (Mahjoub et al., 2024) and MAT (Wen et al., 2022), which leverage centralized information during both training and execution. For CTDE baselines, we consider MAPPO (Yu et al., 2022), which employ centralized critics while executing decentralized policies. We include a basic CTDS baseline and MAGPO as a direct baseline, which shares the same guider–learner framework and training pipeline as MAIGPO but performs only action-level alignment via KL regularization and backtracking. MAIGPO, MAGPO, and CTDS adopt SABLE as the default joint-policy backbone. All algorithms are implemented using the JAX-based MARL library MAVA (de Kock et al., 2021).

Evaluation Protocol and Hyperparameters. We follow the evaluation protocol of (Mahjoub et al., 2024). Each algorithm is trained with 1–5 seeds per task for 20M environment steps and evaluated at 122 checkpoints, with 32 episodes per checkpoint. We report task-level means with 95% confidence intervals and suite-level min–max normalized IQM with 95% stratified bootstrap confidence intervals. For probability-of-improvement analysis, we use the absolute metrics evaluated over 320 episodes; Baseline results are taken from MAVA and MAGPO or reproduced using

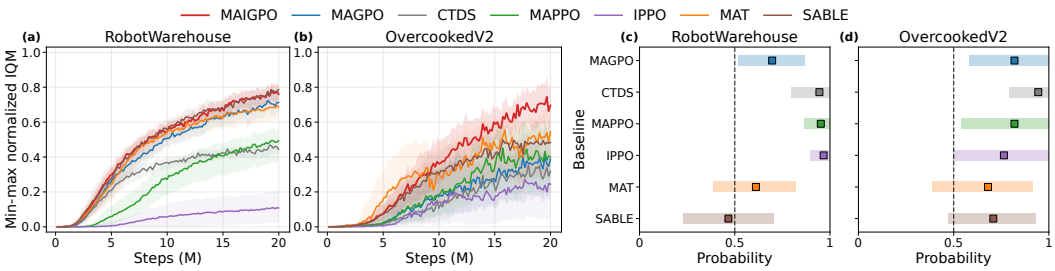

Figure 4: Aggregate performance over 15 RobotWarehouse and 8 OvercookedV2 tasks. (a–b) Min–max normalized IQM with 95% stratified bootstrap confidence intervals. (c–d) Probability of improvement (PoI) of MAIGPO over each baseline; values above 0.5 favor MAIGPO.

their best-performing hyperparameter configurations. MAIGPO retains all backbone hyperparameters and tunes only its own coefficients; see Appendix F.1.

Environments. We evaluate MAIGPO on two JAX-based multi-agent benchmark suites: Robotic Warehouse (RobotWarehouse) (Papoudakis et al., 2020) with 15 tasks, and OvercookedV2 (Gessler et al., 2025) with 8 tasks. Further environment details are provided in Appendix D.

5.2 Results

Comparative Evaluation. Figure 3 shows the learning curves on representative RobotWarehouse and OvercookedV2 tasks, where MAIGPO consistently outperforms all baselines, including the privileged CTCE methods MAT and SABLE. See Appendices E.2 and F for additional results. Figure 4 further summarizes aggregate performance. On RobotWarehouse, MAIGPO substantially closes the performance gap between MAGPO and its privileged SABLE teacher, bringing aggregate performance close to the teacher level. On OvercookedV2, MAIGPO further surpasses all compared methods. The probability-of-improvement results further support these findings. MAIGPO achieves a PoI above 0.5 against most baselines on both RobotWarehouse and OvercookedV2, demonstrating robust improvements across tasks. Together, these results highlight the effectiveness of MAIGPO for decentralized policy learning.

Effectiveness under Increasing Information Gaps. Figure 5 presents the task-wise normalized improvement of MAIGPO over MAGPO on RobotWarehouse. Tasks are arranged clockwise by increasing local–global information gap, from `tiny-2ag` (1×3, 2 agents) to `xlarge-4ag` (4×5, 4 agents). Across all tasks, the local sensor range is fixed at 1, while larger layouts and more agents can increase the amount of global information unavailable from local observations. The Hard variants further reduce the request queue size, increasing the dependence on global task allocation and inter-agent coordination. These factors therefore tend to widen the decision-relevant information gap. The plotted values denote the per-task min–max normalized differences in absolute return between MAIGPO and MAGPO. As the information gap increases, the relative advantage of MAIGPO generally becomes more pronounced, especially on the larger layouts and Hard variants. This trend indicates that explicit information extraction and representation alignment become increasingly valuable as the guider–learner information gap widens.

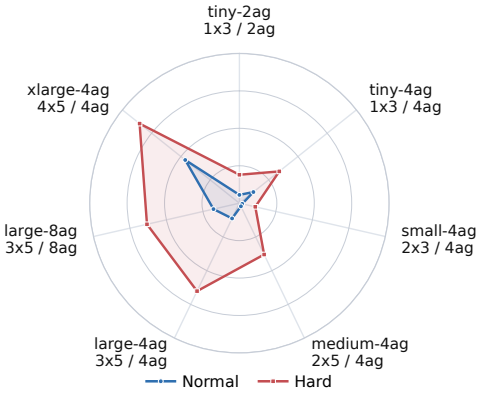

Figure 5: Normalized MAIGPO–MAGPO improvement across RobotWarehouse tasks, ordered clockwise by an increasing proxy for the local–global information gap. Labels show the layout and agent count; Hard variants use smaller request queues.

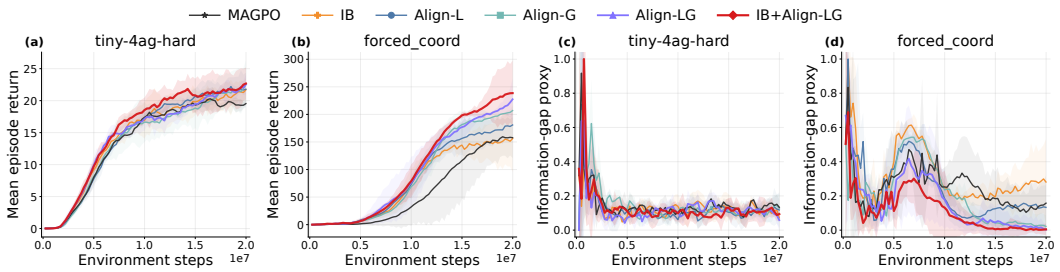

Figure 6: Ablation study on task performance and information-gap proxy. (a–b) Mean episode return on RobotWarehouse `tiny-4ag-hard` and OvercookedV2 `forced_coord`. (c–d) Corresponding information-gap proxy, normalized within each task.

Ablation Study and Hyperparameter Analysis. We evaluate the contributions of information extraction and bidirectional information alignment on `tiny-4ag-hard` in RobotWarehouse and `forced_coord` in OvercookedV2. Since the optimal action is unknown, the decision-relevant information gap $\mathcal{I}_{\text{gap}}$ cannot be estimated directly. We therefore introduce a return-based proxy, $\mathcal{G}_{\text{ret}} = \frac{|R_{\text{guider}} - R_{\text{learner}}|}{|R_{\text{guider}}| + |R_{\text{learner}}| + \epsilon}$, where R_{guider} and R_{learner} denote the evaluation returns of the guider and the learner, respectively, and $\epsilon = 10^{-8}$ ensures numerical stability. A smaller return gap indicates closer evaluation performance and is used as a coarse proxy for a smaller guider–learner information gap. Figure 6 presents ablation results on RobotWarehouse `tiny-4ag-hard` and OvercookedV2 `forced_coord`. The first two panels report episode returns, while the last two show the corresponding information-gap proxy. IB alone can improve task performance, but does not consistently reduce the gap proxy. Among the alignment variants, Align-LG achieves the highest final return on both tasks while maintaining a relatively low late-stage gap proxy. The full IB+Align-LG variant further achieves the strongest overall performance, combining high return with a substantially lower gap proxy. These results suggest complementary roles for the two components: IB promotes compact decision-relevant representations, while bidirectional alignment helps reduce guider–learner information gap. Figure 7 reports hyperparameter sensitivity on `tiny-4ag-hard`, with each component added to MAGPO. An excessively large β_l degrades performance, suggesting that local representations may be more sensitive to over-compression. For alignment, the best performance is obtained with balanced coefficients $\lambda_g = \lambda_l = 10^{-3}$, indicating that the alignment strengths require appropriate tuning to achieve the best performance.

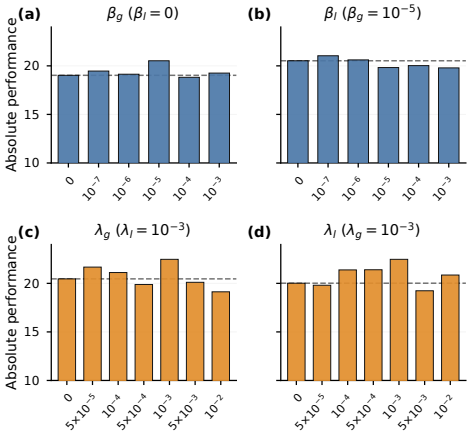

Figure 7: Hyperparameter sensitivity.

6 Conclusion

We reinterpret the guider–learner imitation gap from an information-theoretic perspective as a decision-relevant information gap. Based on this view, we propose MAIGPO, which combines information bottleneck-based extraction with bidirectional representation alignment to improve guider–learner training beyond action-level imitation. Experiments across two benchmark suites show consistent gains over strong CTDE, CTDS, and several CTCE baselines, with larger improvements under greater coordination difficulty and information asymmetry. These results highlight the value of explicitly extracting and aligning decision-relevant representations. One limitation is that both MAGPO and MAIGPO employ a uniform alignment strength across all states, without accounting for state-dependent variations in the information gap. Future work could investigate observation-dependent gap estimation and adaptive alignment mechanisms.

AI USE STATEMENT

Generative AI tools were used to assist with language refinement, and code debugging. The authors independently verified all technical content, experimental results, and analyses, and take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the algorithm and experimental implementation. An anonymized implementation and experimental configurations are available at `https://anonymous.4open.science/r/MAIGPO-C801`. The code will be made publicly available upon acceptance.

REFERENCES

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2623–2631, 2019.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pp. 531–540. PMLR, 2018.

Clément Bonnet, Daniel Luo, Donal Byrne, Shikha Surana, Sasha Abramowitz, Paul Duckworth, Vincent Coyette, Laurence I Midgley, Elshadai Tegegn, Tristan Kalloniatis, et al. Jumanji: a diverse suite of scalable reinforcement learning environments in jax. *arXiv preprint arXiv:2306.09884*, 2023.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

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.

Filippos Christianos, Georgios Papoudakis, Muhammad A Rahman, and Stefano V Albrecht. Scaling multi-agent reinforcement learning with selective parameter sharing. In *International Conference on Machine Learning*, pp. 1989–1998. PMLR, 2021.

Ruan de Kock, Omayma Mahjoub, Sasha Abramowitz, Wiem Khlifi, Callum Rhys Tilbury, Claude Formanek, Andries Smit, and Arnu Pretorius. Mava: a research library for distributed multi-agent reinforcement learning in jax. *arXiv preprint arXiv:2107.01460*, 2021.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

Tobias Gessler, Tin Dizdarevic, Ani Calinescu, Benjamin Ellis, Andrei Lupu, and Jakob Foerster. Overcookedv2: Rethinking overcooked for zero-shot coordination. In *International Conference on Learning Representations*, volume 2025, pp. 62752–62784, 2025.

Rihab Gorsane, Omayma Mahjoub, Ruan John de Kock, Roland Dubb, Siddarth Singh, and Arnu Pretorius. Towards a standardised performance evaluation protocol for cooperative marl. *Advances in Neural Information Processing Systems*, 35:5510–5521, 2022.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Maximilian Hüttenrauch, Adrian Šošić, and Gerhard Neumann. Deep reinforcement learning for swarm systems. *Journal of Machine Learning Research*, 20(54):1–31, 2019.

Maximilian Igl, Kamil Ciosek, Yingzhen Li, Sebastian Tschiatschek, Cheng Zhang, Sam Devlin, and Katja Hofmann. Generalization in reinforcement learning with selective noise injection and information bottleneck. *Advances in neural information processing systems*, 32, 2019.

Yonghyeon Jo, Sunwoo Lee, Junghyuk Yeom, and Seungyul Han. Fox: Formation-aware exploration in multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12985–12994, 2024.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. Llave: Large language and vision embedding models with hardness-weighted contrastive learning. *arXiv preprint arXiv:2503.04812*, 2025.

Chenghao Li, Tonghan Wang, Chengjie Wu, Qianchuan Zhao, Jun Yang, and Chongjie Zhang. Celebrating diversity in shared multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:3991–4002, 2021.

Tianxu Li, Kun Zhu, Juan Li, and Yang Zhang. Learning distinguishable trajectory representation with contrastive loss. *Advances in Neural Information Processing Systems*, 37:64454–64478, 2024a.

Xinran Li, Zifan Liu, Shibo Chen, and Jun Zhang. Individual contributions as intrinsic exploration scaffolds for multi-agent reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 28387–28402, 2024b.

Yueheng Li, Guangming Xie, and Zongqing Lu. Guided policy optimization under partial observability. *arXiv preprint arXiv:2505.15418*, 2025a.

Yueheng Li, Guangming Xie, and Zongqing Lu. Multi-agent guided policy optimization. *arXiv preprint arXiv:2507.18059*, 2025b.

Boyin Liu, Zhiqiang Pu, Yi Pan, Jianqiang Yi, Yanyan Liang, and Du Zhang. Lazy agents: A new perspective on solving sparse reward problem in multi-agent reinforcement learning. In *International Conference on Machine Learning*, pp. 21937–21950. PMLR, 2023.

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.

David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Omayma Mahjoub, Sasha Abramowitz, Ruan de Kock, Wiem Khlifi, Simon du Toit, Jemma Daniel, Louay Ben Nessir, Louise Beyers, Claude Formanek, Liam Clark, et al. Sable: a performant, efficient and scalable sequence model for marl. *arXiv preprint arXiv:2410.01706*, 2024.

Thanh Thi Nguyen, Ngoc Duy Nguyen, and Saeid Nahavandi. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. *IEEE transactions on cybernetics*, 50(9):3826–3839, 2020.

Frans A Oliehoek, Christopher Amato, et al. *A concise introduction to decentralized POMDPs*, volume 1. Springer, 2016.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. *arXiv preprint arXiv:2006.07869*, 2020.

Vihang Patil, Markus Hofmarcher, Elisabeth Rumetshofer, and Sepp Hochreiter. Contrastive abstraction for reinforcement learning. *arXiv preprint arXiv:2410.00704*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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, 2020.

Yara Rizk, Mariette Awad, and Edward W Tunstel. Decision making in multiagent systems: A survey. *IEEE Transactions on Cognitive and Developmental Systems*, 10(3):514–529, 2018.

Alexander Rutherford, Benjamin Ellis, Matteo Gallici, Jonathan Cook, Andrei Lupu, Gardar Ingvarsson Juto, Timon Willi, Ravi Hammond, Akbir Khan, Christian Schroeder de Witt, et al. Jaxmarl: Multi-agent rl environments and algorithms in jax. *Advances in Neural Information Processing Systems*, 37:50925–50951, 2024.

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.

Lukas M Schmidt, Johanna Brosig, Axel Plinge, Bjoern M Eskofier, and Christopher Mutschler. An introduction to multi-agent reinforcement learning and review of its application to autonomous mobility. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1342–1349. IEEE, 2022.

Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Bhavya Sukhija, Stelian Coros, Andreas Krause, Pieter Abbeel, and Carmelo Sferrazza. Maxinforl: Boosting exploration in reinforcement learning through information gain maximization. *arXiv preprint arXiv:2412.12098*, 2024.

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

Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, pp. 330–337, 1993.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Jianhong Wang, Wangkun Xu, Yunjie Gu, Wenbin Song, and Tim C Green. Multi-agent reinforcement learning for active voltage control on power distribution networks. *Advances in Neural Information Processing Systems*, 34:3271–3284, 2021.

Tonghan Wang, Jianhao Wang, Yi Wu, and Chongjie Zhang. Influence-based multi-agent exploration. *arXiv preprint arXiv:1910.05512*, 2019.

Zifeng Wang, Zheng Zhan, Yifan Gong, Yucai Shao, Stratis Ioannidis, Yanzhi Wang, and Jennifer Dy. Dualhsic: Hsic-bottleneck and alignment for continual learning. In *International Conference on Machine Learning*, pp. 36578–36592. PMLR, 2023.

Luca Weihs, Unnat Jain, Iou-Jen Liu, Jordi Salvador, Svetlana Lazebnik, Aniruddha Kembhavi, and Alex Schwing. Bridging the imitation gap by adaptive insubordination. *Advances in Neural Information Processing Systems*, 34:19134–19146, 2021.

Muning Wen, Jakub Kuba, Runji Lin, Weinan Zhang, Ying Wen, Jun Wang, and Yaodong Yang. Multi-agent reinforcement learning is a sequence modeling problem. *Advances in Neural Information Processing Systems*, 35:16509–16521, 2022.

Rushuai Yang, Chenjia Bai, Hongyi Guo, Siyuan Li, Bin Zhao, Zhen Wang, Peng Liu, and Xuelong Li. Behavior contrastive learning for unsupervised skill discovery. In *International conference on machine learning*, pp. 39183–39204. PMLR, 2023.

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.

Yifan Zhong, Jakub Grudzien Kuba, Xidong Feng, Siyi Hu, Jiaming Ji, and Yaodong Yang. Heterogeneous-agent reinforcement learning. *Journal of Machine Learning Research*, 25(32): 1–67, 2024.

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.

Yihe Zhou, Shunyu Liu, Yunpeng Qing, Kaixuan Chen, Tongya Zheng, Yanhao Huang, Jie Song, and Mingli Song. Is centralized training with decentralized execution framework centralized enough for marl? *arXiv preprint arXiv:2305.17352*, 2023.

A CTDS AND MAGPO

In this section, we provide additional details about the **CTDS** (Centralized-Teacher with Decentralized-Students) paradigm and the **MAGPO** algorithm, which serves as a baseline for MAIGPO.

A.1 CTDS PARADIGM

CTDS is a multi-agent reinforcement learning paradigm in which a centralized *teacher* guides multiple decentralized *students*. During training, the teacher has access to global information and can provide richer guidance signals, while the students only observe local information. The students learn to mimic the teacher's decisions via policy alignment. Formally, for each student i:

$$\mathcal{L}^i_{\text{CTDS}} = \text{KL}(\pi_{\text{student}_i}(\cdot|o_i) \,||\, \pi_{\text{teacher}_i}(\cdot|s_{\text{global}}))$$

where π_{teacher_i} is the teacher's policy conditioned on the joint state s_{global}, and π_{student_i} is the student policy conditioned on the local observation o_i.

A.2 MAGPO BASELINE

MAGPO (Multi-Agent Guided Policy Optimization) (Li et al., 2025b) is a multi-agent reinforcement learning framework designed to improve coordination via a guider–learner structure. Training proceeds in two stages: first updating the guider policy, and then updating the learner policies. MAGPO falls under the CTDS (Centralized-Teacher, Decentralized-Student) paradigm. We refer readers to the original paper for detailed derivations and implementation details.

Guider Update. The guider policy μ_ϕ is first optimized to maximize the standard reinforcement learning (RL) objective and then aligned with the learners via a KL-regularized loss:

$$\mathcal{L}(\phi) = -\frac{1}{Tn}\sum_{j=1}^{n}\sum_{t=0}^{T-1}\Big[\min\left(r_t^{ij}(\phi)\hat{A}_t, \mathrm{clip}(r_t^{ij}(\phi),\epsilon,\delta)\hat{A}_t\right) - m_t^{ij}\, D_{\mathrm{KL}}(\pi_\theta^{ij}\|\mu_\phi^{ij})\Big],$$

where $r_t^{ij}(\phi)$ is the probability ratio, $\hat{A}_t$ is the advantage estimate computed via generalized advantage estimation (GAE), and m_t^{ij} is a mask that ensures the KL loss is applied only when the ratio constraint is violated. The double clipping function and the mask prevent the guider from drifting too far from the learners while still providing informative guidance.

Learner Update. Each learner policy π_θ is updated with a combination of (i) behavior cloning toward the guider and (ii) an auxiliary RL objective to improve returns from collected trajectories:

$$\mathcal{L}(\theta) = \frac{1}{Tn}\sum_{j=1}^{n}\sum_{t=0}^{T-1}\Big[D_{\mathrm{KL}}(\pi_\theta^{ij}\|\mu_\phi^{ij}) - \lambda\min(r_t^{ij}(\theta)\hat{A}_t, \mathrm{clip}(r_t^{ij}(\theta),1-\epsilon,1+\epsilon)\hat{A}_t)\Big].$$

This ensures that learners acquire useful behavior from the guider while directly improving their own returns.

Advantages of MAGPO. The main improvement of MAGPO over standard CTDS is Guider Backtracking. When we update the guider, MAGPO adds a double clipping function and a KL-regularized loss at the action level. This keeps the guider's policy from being too far from the student's, so the student can learn better and not get lost.

B Information-Theoretic Analysis of Guider–Learner Framework

Assumption 2 (Ideal Decision-Relevant Encoding)**.** *The representations Z_t^g and Z_t^l preserve all and only decision-relevant information from their respective inputs:*

$$\begin{aligned} I(A_t; S_t^+ \mid Z_t^g) = 0, \quad I(A_t; O_t \mid Z_t^l) = 0, \\ I(S_t^+; Z_t^g \mid A_t) = 0, \quad I(O_t; Z_t^l \mid A_t) = 0. \end{aligned} \tag{14}$$

The first pair of conditions characterizes decision sufficiency: given the corresponding representation, the original input provides no additional information for predicting A_t. The second pair imposes decision minimality, requiring the representation to discard input-specific information that is irrelevant once A_t is known. This idealization is consistent with the objective of information bottleneck-based representation learning, which seeks compact representations that preserve task-relevant information while suppressing redundancy. Although these conditions may not hold exactly in practice, they provide a principled abstraction for analyzing the information asymmetry between the guider and learner.

Proposition 2 (Decision-Relevant Information Asymmetry)**.** *Under Assumption 2 and the Markov relations established above, the following information ordering holds:*

$$I(A_t; Z_t^l) = I(A_t; O_t) \le I(A_t; S_t^+) = I(A_t; Z_t^g). \tag{15}$$

Proof. Since Z_t^g is encoded from S_t^+, we have

$$I(A_t; Z_t^g \mid S_t^+) = 0. \tag{16}$$

Together with the sufficiency condition $I(A_t; S_t^+ \mid Z_t^g) = 0$, the chain rule gives

$$\begin{aligned} I(A_t; S_t^+, Z_t^g) &= I(A_t; S_t^+) + I(A_t; Z_t^g \mid S_t^+) \\ &= I(A_t; S_t^+), \end{aligned} \tag{17}$$

and

$$\begin{aligned} I(A_t; S_t^+, Z_t^g) &= I(A_t; Z_t^g) + I(A_t; S_t^+ \mid Z_t^g) \\ &= I(A_t; Z_t^g). \end{aligned} \tag{18}$$

Therefore,

$$I(A_t; S_t^+) = I(A_t; Z_t^g). \tag{19}$$

Similarly, because Z_t^l is encoded from O_t,

$$I(A_t; Z_t^l \mid O_t) = 0. \tag{20}$$

Combining this relation with $I(A_t; O_t \mid Z_t^l) = 0$ from Assumption 2, we obtain

$$I(A_t; O_t) = I(A_t; Z_t^l). \tag{21}$$

Finally, the Markov relation

$$A_t \longrightarrow S_t^+ \longrightarrow O_t \tag{22}$$

implies, by the data processing inequality, that

$$I(A_t; O_t) \leq I(A_t; S_t^+). \tag{23}$$

Combining the above results yields

$$I(A_t; Z_t^l) = I(A_t; O_t) \leq I(A_t; S_t^+) = I(A_t; Z_t^g), \tag{24}$$

which completes the proof. □

Definition 2 (Decision-Relevant Information Gap)**.** *Under Assumption 2 and the established Markov relations, the guider-to-learner decision-relevant information gap is defined as*

$$\mathcal{I}_{\text{gap}} \triangleq I(A_t; Z_t^g) - I(A_t; Z_t^l) = I(A_t; Z_t^g \mid Z_t^l). \tag{25}$$

Proof. Since Z_t^l is generated from O_t, which is itself generated from S_t^+, we have the Markov relation

$$A_t \longrightarrow S_t^+ \longrightarrow O_t \longrightarrow Z_t^l. \tag{26}$$

By Assumption 2,

$$I(A_t; S_t^+ \mid Z_t^g) = 0. \tag{27}$$

Because Z_t^l is a downstream representation of S_t^+, the conditional data processing inequality gives

$$I(A_t; Z_t^l \mid Z_t^g) \leq I(A_t; S_t^+ \mid Z_t^g) = 0. \tag{28}$$

Hence,

$$I(A_t; Z_t^l \mid Z_t^g) = 0. \tag{29}$$

Applying the chain rule of mutual information in two orders yields

$$\begin{aligned} I(A_t; Z_t^g, Z_t^l) &= I(A_t; Z_t^l) + I(A_t; Z_t^g \mid Z_t^l), \\ &= I(A_t; Z_t^g) + I(A_t; Z_t^l \mid Z_t^g). \end{aligned} \tag{30}$$

Using $I(A_t; Z_t^l \mid Z_t^g) = 0$, we obtain

$$I(A_t; Z_t^g \mid Z_t^l) = I(A_t; Z_t^g) - I(A_t; Z_t^l), \tag{31}$$

which completes the proof. □

Lemma 1 (Conditional Independence of Representations)**.** *Under the dependency structure $Z_t^g \leftarrow S_t^+ \rightarrow O_t \rightarrow Z_t^l$ and Assumption 2, the guider and learner representations are conditionally independent given the target action:*

$$I(Z_t^g; Z_t^l \mid A_t) = 0. \tag{32}$$

Proof. Since Z_t^l is generated downstream from S_t^+, the conditional data processing inequality gives

$$I(Z_t^g; Z_t^l \mid A_t) \leq I(Z_t^g; S_t^+ \mid A_t). \tag{33}$$

By Assumption 2,

$$I(Z_t^g; S_t^+ \mid A_t) = 0. \tag{34}$$

Therefore,

$$I(Z_t^g; Z_t^l \mid A_t) = 0. \tag{35}$$

□

Proposition 3 (Decomposition of the Decision-Relevant Information Gap)**.** *Under Assumption 2 and the established dependency structure, the decision-relevant information gap admits the decomposition*

$$\mathcal{I}_{\text{gap}} = I(Z_t^g; A_t) - I(Z_t^g; Z_t^l). \tag{36}$$

Proof. By the chain rule of mutual information,

$$I(A_t; Z_t^g \mid Z_t^l) = I(Z_t^g; A_t) - I(Z_t^g; Z_t^l) + I(Z_t^g; Z_t^l \mid A_t). \tag{37}$$

By Lemma 1,

$$I(Z_t^g; Z_t^l \mid A_t) = 0. \tag{38}$$

Therefore,

$$\mathcal{I}_{\text{gap}} = I(A_t; Z_t^g \mid Z_t^l) = I(Z_t^g; A_t) - I(Z_t^g; Z_t^l), \tag{39}$$

which completes the proof. □

Under ideal decision-relevant encoding, the guider representation preserves its decision information, and minimizing the information gap therefore reduces to maximizing the decision-relevant information shared between the guider and learner representations. Although the ideal conditions may hold only approximately in practice, the guider typically retains richer decision-relevant information than the learner due to its centralized input. Moreover, as both representations are optimized toward the same action target, their residual dependence conditioned on A_t, namely $I(Z_t^g; Z_t^l \mid A_t)$, is expected to become small during training. Therefore, maximizing $I(Z_t^g; Z_t^l)$ remains a practical and effective surrogate for mitigating the decision-relevant information gap, even beyond the idealized setting. Importantly, such alignment should be bidirectional: the guider representation provides richer decision-relevant knowledge to the learner, while feedback from the learner prevents the guider representation from becoming excessively dependent on information unavailable during decentralized execution. Bidirectional alignment thus promotes more effective guider–learner collaborative training.

C Bidirectional Information Alignment

Contrastive Representation Alignment. According to Proposition 3, reducing the decision-relevant information gap can be approximated by increasing the information shared between the guider and learner representations. To this end, we adopt InfoNCE-based contrastive learning to align their latent features.

Given N paired representations $\{(x_i, y_i)\}_{i=1}^N$, the InfoNCE loss is defined as

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\text{sim}(x_i, y_i)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}(x_i, y_j)/\tau\right)}, \tag{40}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and $\tau > 0$ is a temperature coefficient. The matched pair (x_i, y_i) is treated as a positive pair, while $\{(x_i, y_j)\}_{j\neq i}$ form negative pairs. Minimizing Eq. equation 40 increases the similarity of matched representations while separating mismatched ones, thereby maximizing a lower bound on their mutual information:

$$I(X;Y) \geq \log N - \mathcal{L}_{\text{InfoNCE}}. \tag{41}$$

Guider–Learner Alignment. Let $Z_{b,n}^g$ and $Z_{b,n}^l$ denote the guider and learner representations of the n-th sample in the b-th training sequence, respectively. Here, n indexes the agent–time samples used for contrastive alignment, and N denotes the number of such samples within each sequence. To account for potentially different representation spaces, the source representation is first mapped by a learnable projection head. For learner-to-guider alignment, we write

$$H_{b,n}^l = h_l(Z_{b,n}^l). \tag{42}$$

Both the projected learner features and guider features are ℓ_2-normalized:

$$\bar{H}_{b,n}^l = \frac{H_{b,n}^l}{\|H_{b,n}^l\|_2}, \qquad \bar{Z}_{b,n}^g = \frac{Z_{b,n}^g}{\|Z_{b,n}^g\|_2}. \tag{43}$$

Their pairwise similarity logits are then computed as

$$s^{l\to g}_{b,n,j} = \frac{(\bar{H}^l_{b,n})^\top \bar{Z}^g_{b,j}}{\tau}. \tag{44}$$

For each learner feature $\bar{H}^l_{b,n}$, the corresponding guider feature $\bar{Z}^g_{b,n}$ forms the positive pair, while the remaining guider features $\{\bar{Z}^g_{b,j}\}_{j\neq n}$ within the same sequence serve as negative samples. The learner-to-guider alignment loss is

$$\mathcal{L}^{l\to g}_{\text{align}} = -\frac{1}{BN\log N}\sum_{b=1}^{B}\sum_{n=1}^{N}\log\frac{\exp\left(s^{l\to g}_{b,n,n}\right)}{\sum_{j=1}^{N}\exp\left(s^{l\to g}_{b,n,j}\right)}, \tag{45}$$

where division by $\log N$ normalizes the loss with respect to the number of contrastive candidates.

Similarly, guider-to-learner alignment uses a projection head $H^g_{b,n} = h_g(Z^g_{b,n})$ and treats the learner representations as contrastive targets:

$$s^{g\to l}_{b,n,j} = \frac{(\bar{H}^g_{b,n})^\top \bar{Z}^l_{b,j}}{\tau}, \tag{46}$$

leading to

$$\mathcal{L}^{g\to l}_{\text{align}} = -\frac{1}{BN\log N}\sum_{b=1}^{B}\sum_{n=1}^{N}\log\frac{\exp\left(s^{g\to l}_{b,n,n}\right)}{\sum_{j=1}^{N}\exp\left(s^{g\to l}_{b,n,j}\right)}. \tag{47}$$

The overall bidirectional alignment objective is

$$\mathcal{L}_{\text{align}} = \lambda_l\mathcal{L}^{l\to g}_{\text{align}} + \lambda_g\mathcal{L}^{g\to l}_{\text{align}}, \tag{48}$$

where λ_l and λ_g control the two alignment directions. Learner-to-guider alignment transfers richer decision-relevant information from the guider to the learner, whereas guider-to-learner alignment encourages the guider to retain representations that remain accessible to the decentralized learner. Their combination therefore facilitates guider–learner collaborative training.

D Environment Details

We evaluate our method on two representative multi-agent benchmarks and select a set of representative tasks from each environment: (1) **Robotic Warehouse (RobotWarehouse)**, with 15 tasks; and (2) **OvercookedV2**, with 8 tasks.

The evaluated tasks are listed as follows:

- **RobotWarehouse**: `tiny-2ag`, `tiny-2ag-hard`, `tiny-4ag`, `tiny-4ag-hard`, `small-4ag`, `small-4ag-hard`, `medium-4ag`, `medium-4ag-hard`, `medium-6ag`, `large-4ag`, `large-4ag-hard`, `large-8ag`, `large-8ag-hard`, `xlarge-4ag`, `xlarge-4ag-hard`.
- **OvercookedV2**: `coord_ring`, `counter_circuit`, `cramped_room_v2`, `forced_coord`, `fun_coordination`, `more_fun_coordination`, `two_rooms`, `two_rooms_both`.

D.1 RobotWarehouse

The Robot Warehouse (RobotWarehouse) environment simulates a warehouse where robots autonomously navigate to fetch and deliver requested items from specific shelves to designated workstations and then return. Inspired by real-world autonomous delivery depots, the objective in RobotWarehouse is for a team of robots to deliver as many randomly placed items as possible within a fixed time budget.

In this work, we use a JAX-based implementation of the original RobotWarehouse environment Christianos et al. (2021) from the Jumanji environment suite Bonnet et al. (2023). As a result,

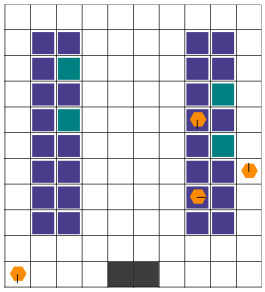

Figure 8: Environment rendering for Robot Warehouse. Task name: `tiny-4ag`.

there are minor differences in collision handling: while the original implementation includes logic to resolve collisions, the Jumanji version simply ends the episode if two agents collide.

Naming convention. Tasks in RobotWarehouse are named according to the format `<size>-<num agents>ag<diff>`, where each field has the following meaning:

- `<size>`: Defines the warehouse size, specifying the number of rows and columns of shelf groups (e.g., `tiny`, `small`, `medium`, `large`).
- `<num agents>`: Number of agents in the environment.
- `<diff>`: Optional difficulty level; `easy` and `hard` correspond to $2N$ and $N/2$ requests respectively, where N is the number of agents. By default, there are N requests.

Observation space. Observations are partially observable: each agent perceives a 3×3 square grid centered on its position. Within this area, agents can observe their own position and orientation, the positions and orientations of other agents, as well as shelves and whether each shelf contains a package for delivery.

Action space. The action space is discrete and consists of five actions that allow navigation and shelf interaction: `stop` (no operation), `turn left`, `turn right`, `move forward`, and `load/unload shelf`.

Reward. Agents receive a reward of 1 for each successful delivery of a requested shelf (colored green in Figure 8) to a designated goal (black), and 0 otherwise. Achieving this reward requires a sequence of successful actions, making the reward notably sparse.

D.2 OVERCOOKEDV2

Overcooked-AI is a fully observable cooperative environment where two cooks (agents) must cooperate to prepare and serve onion soups. It is inspired by the popular videogame Overcooked. OvercookedV2 is an extended version of the original Overcooked environment. It introduces meaningful partial observability and increased stochasticity to enable more complex coordination challenges. In this work, we use a JAX-based implementation of the OvercookedV2 environment from the jaxmarl environment suite Rutherford et al. (2024).

Task layouts. Each task is defined by a predefined kitchen layout, such as `more_fun_coordination`. Different layouts vary in kitchen geometry, object placement, and agent accessibility, thereby inducing different coordination requirements.

Observation space. Observations are structured as a width x height x num_channels tensor, with partial observability based on the configured view radius.

Action space. The discrete action space consists of six actions: moving in four directions, remaining stationary, and interacting with nearby objects. The interaction action is used to pick up or place items, operate pots, and deliver completed dishes.

Reward. Agents receive a shared reward for successfully delivering a correctly prepared dish. Additional shaped rewards may be provided for intermediate subtasks, such as placing ingredients into a pot or collecting a cooked dish.

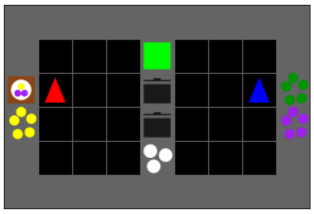

Figure 9: Environment rendering for OvercookedV2. Task name: `more_fun_coordination`.

E Further experimental results

E.1 Experimental Environment

All experiments were conducted on machines equipped with NVIDIA GPUs, including RTX 3090, RTX 4090, RTX 5090, and H800 GPUs. A subset of the comparative experiments was performed on NVIDIA H800 GPUs. Notably, JAX may employ nondeterministic GPU kernels, particularly for parallel reduction operations, such that GPU-level randomness cannot be fully controlled by JAX random seeds alone. To improve reproducibility, we set `os.environ["XLA_FLAGS"] = "--xla_gpu_deterministic_ops=true"` in all experiments, enforcing deterministic GPU operations. Although this setting reduced the training speed by approximately 25% in our experiments, it ensured consistent and reproducible results across repeated runs.

E.2 Per-task sample efficiency results

RobotWarehouse. Figure 10(a-p) presents the per-task learning curves on RobotWarehouse. MAIGPO outperforms the CTDE and CTDS baselines across all 15 tasks and surpasses the CTCE baselines on 7 of them, demonstrating consistently strong performance under challenging coordination settings.

OvercookedV2. Figure 10(p–w) presents the task-wise learning curves on OvercookedV2. MAIGPO outperforms the CTDE and CTDS baselines and surpasses the CTCE baselines on six tasks, demonstrating strong performance in challenging coordination settings. Although MAPPO remains competitive, MAIGPO achieves better overall performance across the benchmark.

F Additional Benchmarks

MPE and SMAX. The Multi-Agent Particle Environments (MPE) consist of cooperative physics-based tasks in a two-dimensional world, where agents interact with landmarks and other agents. We focus on *simple-spread*, a non-adversarial task without communication, in which agents must cooperatively cover landmarks while avoiding collisions. We use the JAX implementation provided by JaxMARL (Rutherford et al., 2024). Beyond the commonly used three-agent setting (Gorsane et al., 2022).

SMAX (Rutherford et al., 2024) is a JAX-based reimplementation of the StarCraft Multi-Agent Challenge (SMAC) (Samvelyan et al., 2019) designed for efficient multi-agent experimentation without requiring the StarCraft II engine. Agents control heterogeneous units and cooperate to defeat an opposing team. Since SMAX uses a distinct implementation, its results are not directly comparable to those obtained on the original SMAC.

Figure 11 reports representative learning curves on MPE `simple_spread_5ag` and SMAX `smacv2_10_units`. MAIGPO improves over MAGPO on both tasks, with a particularly clear gain on `smacv2_10_units`. These results further demonstrate that the performance gains of MAIGPO extend beyond RobotWarehouse and OvercookedV2 to additional cooperative benchmarks.

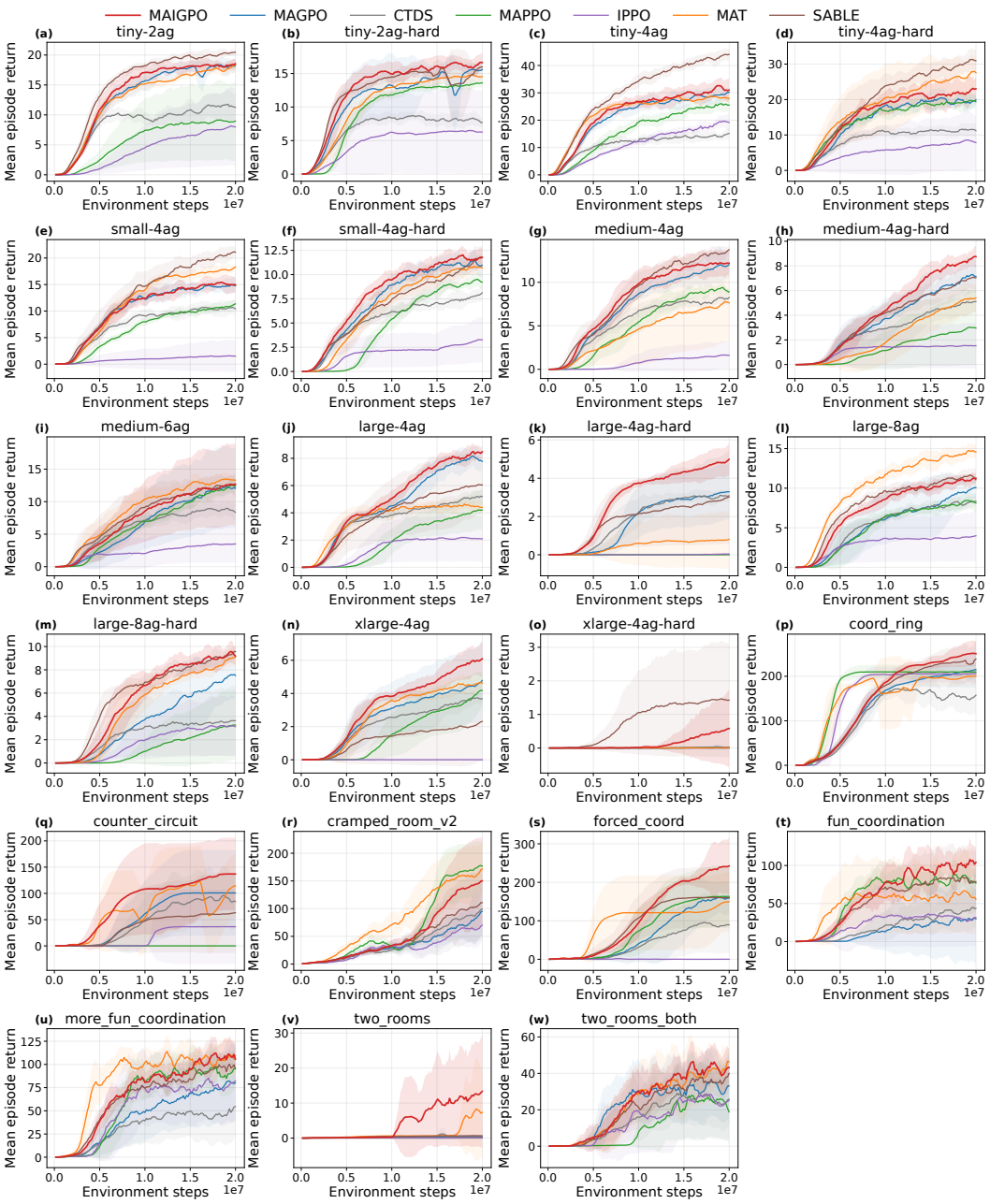

Figure 10: Task-level learning curves.

F.1 HYPERPARAMETERS

Different tasks require different strengths of information extraction and representation alignment. Following prior work, we tune the relevant hyperparameters separately for each task. Since the information bottleneck coefficients are sensitive to task characteristics, we perform a sequential search strategy: we first tune the guider-side IB coefficient β_g, followed by the learner-side IB coefficient β_l. After determining the optimal IB coefficients, we further tune the remaining MAIGPO hyperparameters using a budget of 30 trials with the Tree-structured Parzen Estimator (TPE) implemented in Optuna Akiba et al. (2019). For baseline algorithms, we directly adopt the hyperparameters provided by Mava de Kock et al. (2021) and MAGPO Li et al. (2025b) whenever available. The search space used for MAIGPO is summarized in Table 1.

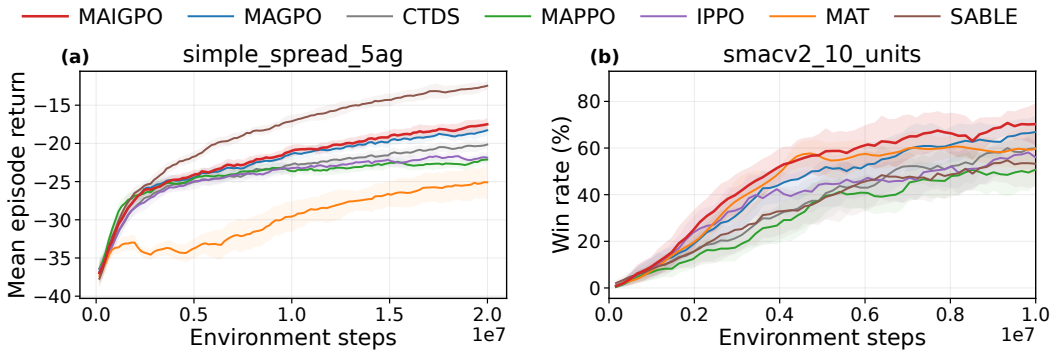

Figure 11: **Learning curves on additional benchmarks.** Left: mean episode return on MPE `simple_spread_5ag`. Right: win rate on SMAX `smacv2_10_units` over the first 10^7 environment steps. Curves show the mean over five random seeds, with shaded 95% confidence intervals.

Table 1: Hyperparameter search space for MAIGPO.

Parameter	**Search Space**
Guider-side IB coefficient β_g	$\{0, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}\}$
Learner-side IB coefficient β_l	$\{0, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}\}$
Guider-side alignment coefficient λ_g	$\{0, 10^{-5}, 5{\times}10^{-5}, 10^{-4}, 5{\times}10^{-4}, 10^{-3}, 5{\times}10^{-3}, 10^{-2}\}$
Learner-side alignment coefficient λ_l	$\{0, 10^{-5}, 5{\times}10^{-5}, 10^{-4}, 5{\times}10^{-4}, 10^{-3}, 5{\times}10^{-3}, 10^{-2}\}$
InfoNCE temperature τ	$\{0.05, 0.1, 0.2\}$