# Mind the Gap: Information Disadvantage as a Learning Signal in Cooperative MARL

**Chang Liu**
University of Florida
changliu2@ufl.edu

## Abstract

In cooperative multi-agent reinforcement learning under partial observability, agents typically differ in both their local observations and in how much value those observations let them recover. Most existing methods treat agents symmetrically when shaping credit and exploration, ignoring this asymmetric information disadvantage. We introduce the per-agent value gap, a state- and agent-specific quantity that measures how much of the state value an agent's local information leaves unresolved, and show that it serves as a single, measurable signal that unifies credit assignment and exploration: shaping the advantage pessimistically where an agent is at a strong disadvantage, and rewarding behaviors that reduce the disadvantage along the trajectory. We instantiate this idea as ValueGap, which estimates the gap online with a per-agent local value head and uses it to drive both advantage shaping and an information-seeking intrinsic reward. We show that the shaping term leaves the expected policy-gradient update unchanged and identify when it reduces variance. We further show that, with fixed critics at stationarity, the average gap-reduction bonus is bounded by fluctuations in the estimated gap and vanishes at the analyzed fixed point when local information fully resolves the state value. On sixteen tasks across five cooperative benchmarks, ValueGap attains the highest interquartile-mean return on every task, with the largest gains on settings combining sparse rewards and asymmetric observability (e.g., a $15\%$ mean relative gain over the strongest baseline across the three SMAC v2 tasks, and $24\%$ on zerg).

## 1 Introduction

Cooperative multi-agent reinforcement learning (MARL) under decentralized partial observability is fundamentally an information problem. The setting underlies applications from cooperative robotics [Alami et al., 1998] to mixed-autonomy traffic [Valiente et al., 2022] and search-and-rescue [Rahman et al., 2022], all naturally modelled as decentralized partially observable Markov decision processes (Dec-POMDPs) [Oliehoek and Amato, 2016]. Each agent acts on a local observation typically far less informative than the joint state, and recent work compensates by enriching local information with learned belief representations or attention modules [Papoudakis et al., 2021a, Xu et al., 2021, Kontogiannis et al., 2025, Sun et al., 2024, Zintgraf et al., 2021, Moreno et al., 2021]. Inferred beliefs do not eliminate the underlying asymmetry: any belief computed from $o^i$ inherits its information ceiling, so two agents in the same state can have very different ceilings on what their policies can achieve.

We refer to the value that an agent's local information leaves unresolved as the per-agent value gap. It is the difference between the value of the current joint policy given the full state and its value given only agent $i$'s local history. The gap is heterogeneous across agents, varies sharply across states, and its magnitude shrinks as the agent acquires information. While related quantities have

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

been studied as the value of information in single-agent POMDPs [Hauskrecht, 2000], prior MARL work has not exposed a per-agent, per-state, online-estimable analogue and used it to drive learning.

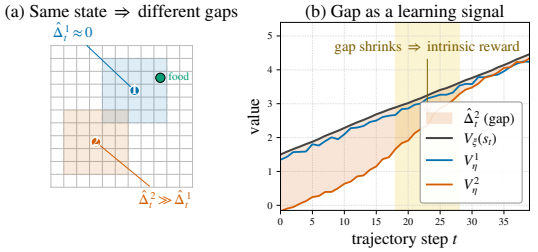


Figure 1: The estimated per-agent value gap $\hat{\Delta}^i$ is state- and agent-specific, and its magnitude shrinks as agents acquire information.

The observation we exploit is that an online estimate of this gap can drive learning, and that it bears on two concerns prior MARL methods have treated separately. In states where an agent is at a strong information disadvantage, the realized advantage is dominated by quantities outside the agent's control; shaping the advantage pessimistically in such states is a natural per-agent response. The same signal also provides a principled exploration target: trajectories along which the gap shrinks are precisely those along which the agent is acquiring useful information. The two uses are complementary: shaping acts locally on individual transitions, while the intrinsic reward acts at the trajectory level by pushing the agent toward states where the local signal is reliable.

We make the following contributions:

- We formalize the per-agent value gap as a measurable quantity in cooperative MARL under partial observability, and propose ValueGap, an actor-critic method that estimates the gap online with a per-agent local value head and uses the estimate to drive both advantage shaping and an information-seeking intrinsic reward, linking representation, credit assignment, and exploration through a single signal.
- We show that the one-sided regularized estimator measures the state value that local information leaves unresolved, that the gap-aware shaping is an action-independent baseline (so the policy gradient stays unbiased) and state when it reduces gradient variance, and that the information-seeking bonus is a rectified potential difference with bounded stationary expectation that vanishes when local information resolves the state value.
- We evaluate ValueGap on sixteen tasks spanning Level-Based Foraging, Multi-Robot Warehouse, the Multi-agent Particle Environment, SMAC v2, and Hanabi, against recent representation and exploration baselines including HAPPO. ValueGap delivers consistent gains, with the largest margins on tasks featuring sparse rewards and asymmetric observability.

## 2 Related Work

**Centralized training with decentralized execution.** The centralized-critic family of cooperative MARL methods, exemplified by MADDPG, COMA, MAA2C, and MAPPO [Lowe et al., 2017, Foerster et al., 2018, Papoudakis et al., 2021b, Yu et al., 2022], computes per-agent advantages from a centralized critic conditioned on the full state, providing a natural setting in which a separate decentralized critic can be added without altering the existing pipeline. Parameter and experience sharing variants such as SEAC [Christianos et al., 2020] reduce sample complexity here. Value decomposition methods [Sunehag et al., 2017, Rashid et al., 2020, Son et al., 2019], with hybrid actor-critics [Su et al., 2021] bridging the two families, factorize joint $Q$ functions into per-agent terms, but those terms are credit-assignment artifacts of the factorization rather than independent estimates of the return achievable from local information. Recent benchmarking on fully cooperative tasks [Papoudakis et al., 2021b, Papadopoulos et al., 2025] finds that no single algorithm dominates across task families, motivating an orthogonal axis of improvement.

**State modelling and asymmetric value estimation under partial observability.** A second line of work tries to enrich agents' local information with a learned belief representation. LIAM [Papoudakis et al., 2021a] reconstructs other agents' observations and actions; SIDE [Xu et al., 2021] infers state from local trajectories; recent work [Kontogiannis et al., 2025, Hernandez-Leal et al., 2019, Raileanu et al., 2018, Nguyen et al., 2023] extends this idea to multiple learnable agents and explores filtering of redundant joint state information. Belief-based methods [Moreno et al., 2021, Zintgraf et al., 2021, Hu et al., 2021] and opponent/teammate modelling [He et al., 2016, Albrecht and Stone, 2017, Fu

et al., 2022] use similar inference machinery for related downstream purposes. For POMDP planning, SA-CORAL discovers preferred actions through online clustering and learns intent values to guide counterfactual open-loop search [Liu, 2026]. We use a state representation in the same family as one of these inputs, but our contribution is orthogonal: we are concerned with what to do once the representation is in hand, not with how to learn the representation itself. Closely related on the value-estimation side is the asymmetric actor-critic family [Pinto et al., 2017], which trains a value head with full state access while restricting the policy to local observations; Baisero and Amato [2022] analyze the bias incurred when a state-conditioned critic is paired with a history-conditioned actor, which bears on the approximation we discuss after Definition 2. Closest in spirit is the imitation gap of Weihs et al. [2021], which down-weights an imitation loss where a privileged teacher's policy is not realizable from the student's observations. Our discrepancy is value-based rather than policy-based, requires no teacher, is defined per agent among simultaneously learning agents, and is used for exploration as well as for credit assignment.

**Advantage modification and reward shaping.** Action-independent baselines leave the policy gradient unbiased and can reduce its variance [Greensmith et al., 2004]; multi-agent instances include COMA's counterfactual baseline and the optimal baseline of Kuba et al. [2021], with action-dependent control variates extending the idea [Wu et al., 2018, Liu et al., 2018]. These baselines derive from the joint-action structure of the advantage; ours derives from information disadvantage. Difference rewards [Wolpert and Tumer, 2001, Nguyen et al., 2018] also subtract a second value estimate, but marginalize the agent's own action, whereas we marginalize the unobserved state. Outside RL, modulating each sample's gradient contribution by its estimated difficulty or noise is a long-standing idea in sample weighting [Zhou et al., 2023, Li et al., 2026]; our shaping is additive rather than a weight, which preserves unbiasedness. Our bonus relates to potential-based shaping [Ng et al., 1999, Devlin and Kudenko, 2012] (Section 4.3); Appendix S covers further related work.

**Exploration in cooperative MARL.** Existing MARL exploration strategies fall into density-based [Tang et al., 2017], curiosity-driven [Pathak et al., 2017, Zheng et al., 2021], and information-theoretic categories [Li et al., 2021, Jiang and Lu, 2021], with subgoal generation [Jeon et al., 2022] and joint latent variable models [Mahajan et al., 2019] as further variants. Recent work explores conditionally optimistic exploration tailored to cooperative settings [Zhao et al., 2023] and formation-aware exploration that conditions bonuses on emergent team structure [Jo et al., 2024]. Information-gain bonuses are the closest relatives of ours: VIME [Houthooft et al., 2016] rewards information about model parameters and influence-based exploration [Wang et al., 2020] rewards interaction between agents, whereas our bonus rewards reductions in value-relevant uncertainty about the hidden state. None of these methods explicitly targets reduction of per-agent information disadvantage.

## 3 Preliminaries and Setting

We consider a Dec-POMDP [Oliehoek and Amato, 2016] defined by a tuple $(S, A, P, r, F, O, N, \gamma)$, where $S$ is the state space, $A = A_1 \times \cdots \times A_N$ is the joint action space, $P(s' \mid s, a)$ is the transition kernel, $r(s, a)$ is the team reward, $F = \prod_i F_i$ is the joint observation function with $F_i : S \to O_i$, and $\gamma \in [0, 1)$ is a discount factor. At time $t$, agent $i$ receives observation $o_t^i \sim F_i(s_t)$ and selects action $a_t^i$ from a stochastic policy $\pi_{\psi_i}^i$. Its action-observation history is $h_t^i = (o_0^i, a_0^i, \ldots, o_t^i) \in H_i$. We use a variational encoder-decoder, trained jointly with the policy, to produce the state representation $z_t^i \in Z_i$ from $h_t^i$ [Papoudakis et al., 2021a, Kontogiannis et al., 2025]. We follow the MAA2C [Papoudakis et al., 2021b] backbone: a centralized critic $V_\xi^\pi(s)$ trained by temporal difference, and an actor $\pi_{\psi_i}^i(a_t^i \mid h_t^i, z_t^i)$ trained by policy gradient with the centralized advantage.

We define the locally reachable value and the per-agent value gap as follows.

**Definition 1** (Locally reachable value)**.** For a fixed joint policy $\pi$ and agent $i$, the locally reachable value at history $h_t^i$ with representation $z_t^i$ is

$$V^{\pi,i}(h_t^i, z_t^i) \;=\; \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} \;\middle|\; H_t^i = h_t^i,\, Z_t^i = z_t^i\right], \tag{1}$$

where the expectation marginalizes over the joint state, the other agents' histories, and subsequent trajectories, conditioned on $(h_t^i, z_t^i)$.

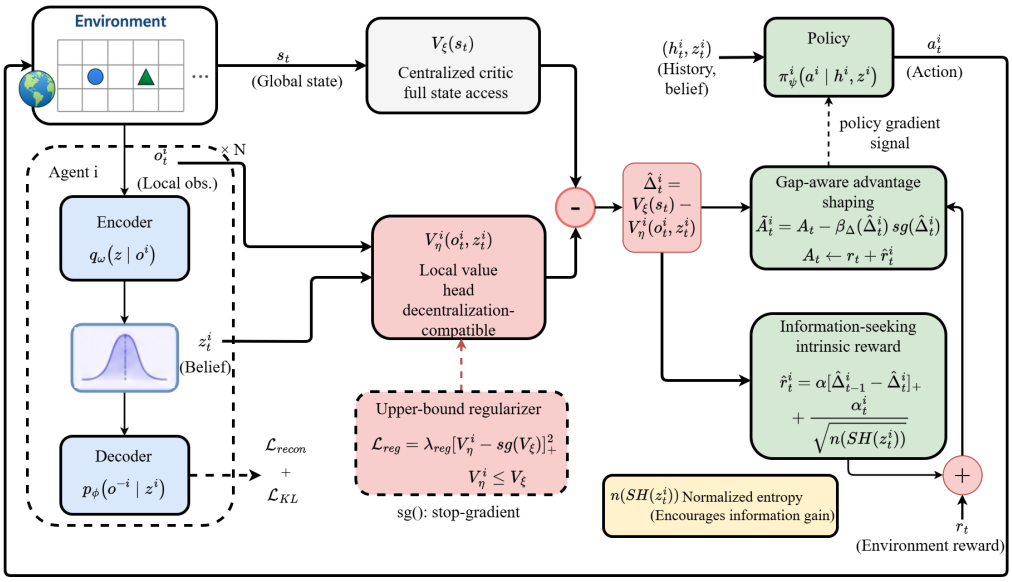


Figure 2: Overview of ValueGap. $\hat{\Delta}_t^i$ drives advantage shaping and the intrinsic reward.

**Definition 2** (Per-agent value gap)**.** The per-agent value gap of agent $i$ at $(s_t, h_t^i, z_t^i)$ under policy $\pi$ is

$$\Delta^{\pi,i}(s_t, h_t^i, z_t^i) \;=\; V^\pi(s_t) \;-\; V^{\pi,i}(h_t^i, z_t^i), \tag{2}$$

where $V^\pi(s_t)$ is the expected return of the same joint policy $\pi$ given the state at time $t$.

Both terms evaluate the same decentralized joint policy using different information. Writing $G_t := \sum_{k\ge 0} \gamma^k r_{t+k}$,

$$V^\pi(s_t) = \mathbb{E}_\pi\big[G_t \mid S_t = s_t\big], \qquad V^{\pi,i}(h_t^i, z_t^i) = \mathbb{E}_\pi\big[G_t \mid H_t^i = h_t^i,\, Z_t^i = z_t^i\big]. \tag{3}$$

The centralized critic estimates $V^\pi(s_t)$, which equals the fully informed value $V^\pi(s_t, \mathbf{h}_t) := \mathbb{E}_\pi[G_t \mid s_t, h_t^1, \ldots, h_t^N]$ when $s_t$ is Markov-sufficient for the continuation of $\pi$ (A6; Appendix H; Baisero and Amato [2022]). Under this condition, the gap vanishes when local information determines the fully informed value. Its spread captures unresolved value, and the regularized estimator in Section 4.1 yields a signal with non-negative conditional mean at the analyzed fixed point. Figure 1 illustrates the two values and their difference.

**A worked example.** In Signalled Coordination, a latent $d \in \{L, R\}$ is drawn uniformly once; agent 1 observes $d$, agent 2 observes nothing, both choose an action in $\{L, R\}$, and the team receives $+1$ if and only if both actions match $d$. Let $\pi^1$ play $d$ with probability $q$ and $\pi^2$ play $L$ with probability $p$. Then $V^\pi(L) = qp$, $V^\pi(R) = q(1-p)$, $\Delta^{\pi,1} \equiv 0$ since agent 1's history is sufficient, and $V^{\pi,2}(\varnothing) = q/2$, so $\Delta^{\pi,2}(L) = q(p - \frac{1}{2}) = -\Delta^{\pi,2}(R)$. The gap varies across agents in the same state and depends on the current policy; it vanishes at $p = \frac{1}{2}$ or $q = 0$. A costly "peek" action revealing $d$ to agent 2 makes $\Delta^{\pi,2} = 0$ afterwards, illustrating how acquiring information reduces the magnitude of the gap.

# 4 Method

## 4.1 Estimating the Gap with a Local Value Head

To estimate $\Delta^{\pi,i}$ online, we introduce a per-agent local value head $V_\eta^i(h_t^i, z_t^i)$ parameterized by an MLP fed with the recurrent state of the actor's history encoder. The local head conditions only on local information and is compatible with decentralized execution. We train it by temporal difference,

$$\mathcal{L}_{V^i}(\eta) \;=\; \mathbb{E}\Big[\big(r_t + \gamma V_{\eta'}^i(h_{t+1}^i, z_{t+1}^i) - V_\eta^i(h_t^i, z_t^i)\big)^2\Big], \tag{4}$$

where $\eta'$ denotes a slow-moving target network. The local and centralized critics use the same training trajectories.

To capture unresolved value in the local estimate, we add the one-sided regularizer

$$\mathcal{L}_{V^i,\mathrm{reg}}(\eta) \;=\; \lambda_{\mathrm{reg}} \cdot \big[\, V^i_\eta(h^i_t, z^i_t) - \mathrm{sg}(V_\xi(s_t)) \,\big]^2_+, \tag{5}$$

where $[\cdot]_+$ is the rectifier, sg is the stop-gradient operator, and $\lambda_{\mathrm{reg}}$ controls the regularization strength. We define the estimated gap as

$$\hat{\Delta}^i_t \;=\; V_\xi(s_t) - V^i_\eta(h^i_t, z^i_t). \tag{6}$$

The regularizer pulls $V^i_\eta$ toward the lower part of the conditional distribution of $V_\xi(s_t)$ given $(h^i_t, z^i_t)$, yielding a gap estimate with non-negative conditional mean. Lemma 2 (Appendix A) characterizes this property at the idealized fixed point of Equations 4–5:

$$0 \;\le\; \mathbb{E}\big[\hat{\Delta}^i_t \mid h^i_t, z^i_t\big] \;\le\; \tfrac{\lambda_{\mathrm{reg}}}{2}\, \mathbb{E}\Big[\, \big|V_\xi(s_t) - \mathbb{E}[V_\xi(s_t) \mid h^i_t, z^i_t]\big| \;\Big|\; h^i_t, z^i_t \Big]. \tag{7}$$

The conditional mean is positive when local information leaves state value unresolved and vanishes when $(h^i_t, z^i_t)$ determines $V_\xi(s_t)$. This connects the regularized estimate to the dispersion of state value given local information. Figure 9 in Appendix D tracks the estimated gap throughout training.

### 4.2 Gap-Aware Advantage Shaping

The standard centralized-critic policy gradient uses the advantage $A_t = r_t + \gamma V_\xi(s_{t+1}) - V_\xi(s_t)$. We replace it with a gap-aware advantage,

$$\tilde{A}^i_t \;=\; A_t \;-\; \beta_\Delta(\hat{\Delta}^i_t) \cdot \mathrm{sg}(\hat{\Delta}^i_t), \qquad \beta_\Delta(x) = \beta_0 \cdot \sigma(x/\tau), \tag{8}$$

where $\sigma$ is the sigmoid, $\beta_0$ controls the shaping coefficient, and $\tau$ is a temperature. The shaping subtracts a smooth gap-dependent term from the advantage, with sg blocking gradients through the estimated gap. The actor of agent $i$ is updated with

$$\mathcal{L}_{\mathrm{actor}}(\psi_i) \;=\; -\log \pi^i_{\psi_i}(a^i_t \mid h^i_t, z^i_t) \cdot \tilde{A}^i_t \;-\; \beta_H \cdot \mathcal{H}\big(\pi^i_{\psi_i}(\cdot \mid h^i_t, z^i_t)\big). \tag{9}$$

Equation 8 applies an additive, state-dependent shift. Because $\hat{\Delta}^i_t$ does not depend on $a^i_t$, the subtracted term is an action-independent baseline that preserves the expected policy-gradient update. For a positive gap, subtracting this term shrinks positive advantages and increases the magnitude of negative ones, producing a more conservative advantage estimate in high-gap states. The shaping reduces the conditional variance of the gradient estimate when the shift lies strictly between zero and twice the variance-optimal baseline [Greensmith et al., 2004, Kuba et al., 2021], as established in Proposition 2 (Appendix F).

Shaping and the intrinsic reward operate on different time scales: shaping adjusts individual transitions, while the bonus rewards transitions that reduce $\hat{\Delta}^i$, propagating through the value targets along the trajectory. Section 5.4 examines their complementary contributions.

### 4.3 Information-Seeking Intrinsic Reward

The same gap signal guides exploration. Decreases in $\hat{\Delta}^i$ provide a learning signal for moving toward states where local information better resolves state value. We define

$$\hat{r}^i_t \;=\; \alpha \cdot \big[\hat{\Delta}^i_{t-1} - \hat{\Delta}^i_t\big]_+ \;+\; \alpha'_t \cdot \frac{1}{\sqrt{n(\mathrm{SH}(z^i_t))}}, \tag{10}$$

where SH is the SimHash function from Tang et al. [2017], $n(\cdot)$ counts hash collisions, and $\alpha$ controls the gap-reduction term. The decaying coefficient $\alpha'_t$ weights a residual coverage term that supports exploration during the early stages of critic training. The intrinsic reward enters Equation 8 through $r_t \leftarrow r_t + \hat{r}^i_t$. The first term rectifies a potential difference with potential $-\alpha\hat{\Delta}^i$, connecting it to potential-based shaping [Ng et al., 1999, Devlin and Kudenko, 2012]. Rectification rewards gap reductions while allowing transitions in which the gap temporarily increases.

**Proposition 1** (Stationary magnitude of the gap-reduction bonus)**.** Let $\pi$ be a stationary joint policy (for instance an optimal one) such that the induced chain on $(s_t, h_t^i, z_t^i)$ is stationary under the critics $V_\xi$ and $V_\eta^i$ held fixed, and let $D_t := \hat{\Delta}_{t-1}^i - \hat{\Delta}_t^i$ have finite second moment. Then $\mathbb{E}_\pi[D_t] = 0$ and

$$\mathbb{E}_\pi\Big[\big[D_t\big]_+\Big] \;=\; \tfrac{1}{2}\,\mathbb{E}_\pi|D_t| \;\le\; \tfrac{1}{2}\sqrt{\mathrm{Var}_\pi(D_t)}.$$

The left-hand side is zero if and only if $\hat{\Delta}_t^i$ is almost surely constant along stationary trajectories; in particular, at the idealized fixed point of Lemma 2 it is zero whenever $(h_t^i, z_t^i)$ determines $V_\xi(s_t)$ along those trajectories.

The proof is in Appendix G. The proposition bounds the stationary expectation of the gap-reduction bonus by fluctuations in the estimated gap. Empirically, the bonus is small at convergence, with median magnitude $0.08 \pm 0.02$ relative to the median extrinsic reward magnitude on LBF $2s$-$11\times11$ (Appendix H). The residual coverage term vanishes through the decaying schedule $\alpha_t'$.

### 4.4 Putting It Together

The full objective combines the three modules with the existing actor-critic and representation losses:

$$\mathcal{L}_{\mathrm{ValueGap}} \;=\; \mathcal{L}_{\mathrm{actor}} \;+\; \mathcal{L}_{\mathrm{critic}}(\xi) \;+\; \mathcal{L}_{V^i}(\eta) \;+\; \mathcal{L}_{V^i,\mathrm{reg}}(\eta) \;+\; \mathcal{L}_{\mathrm{repr}}, \tag{11}$$

where $\mathcal{L}_{\mathrm{actor}}$ uses the gap-aware advantage of Equation 8, $\mathcal{L}_{\mathrm{critic}}(\xi)$ is the centralized critic loss, and $\mathcal{L}_{\mathrm{repr}}$ contains the variational reconstruction and KL terms used to train the encoder of $z^i$. Figure 2 shows the architecture, and Algorithm 2 (Appendix C) gives the training loop.

The three modules use the same gap signal: the local value head estimates the gap, advantage shaping adjusts updates in high-gap states, and the intrinsic reward encourages transitions toward lower-gap states. All three are trained from the existing reward stream.

**Computability.** In finite models, $\Delta^{\pi,i}$ is computable by enumeration (Section 3; Appendix I). In deep MARL, $\hat{\Delta}^i$ requires one additional MLP forward pass per agent per step and is trained by TD on trajectories already collected for the centralized critic, adding approximately $26\%$ wall-clock overhead (Section 5.6).

## 5 Experiments

### 5.1 Setup

We evaluate ValueGap on sixteen tasks across five cooperative benchmarks. From Level-Based Foraging (LBF) [Albrecht and Stone, 2017, Christianos et al., 2020] we use six tasks ranging from $8 \times 8$ to $20 \times 20$ grids and from two to five agents. LBF task names follow the convention {sight}s-{rows}×{cols}-{players}p-{food}f; for example, $2s$-$11\times11$-$3p$-$2f$ has a 2-cell observation radius, an $11 \times 11$ grid, three agents and two food items, and requires cooperative pickup (an item is collected only when the combined level of adjacent agents meets the item's level). Larger grids at the same sight range therefore mean sparser reward and more of the episode at high information disadvantage. From Multi-Robot Warehouse (RWARE) [Papoudakis et al., 2021b] we use the three hard tasks tiny-2ag-hard, tiny-4ag-hard, and small-4ag-hard. From the Multi-agent Particle Environment (MPE) [Lowe et al., 2017, Mordatch and Abbeel, 2018] we use Spread-5, Spread-8, and Double Speaker-Listener. From SMAC v2 [Ellis et al., 2023] we use 5v5 protoss, terran, and zerg, the asymmetric observation patterns of which are central to our story. We additionally include Hanabi-Small 2-player as a stress test in extreme partial observability [Bard et al., 2020]; this variant has two colours, hand size two, one life token and a maximum score of 10, which is also the maximum episodic return reported for it here.

We compare ValueGap against MAA2C, MAPPO, COMA, ATM, EOI, EMC, MASER, MAVEN, MLIAM, and SMPE [Papoudakis et al., 2021b, Yu et al., 2022, Foerster et al., 2018, Yang et al., 2022, Jiang and Lu, 2021, Zheng et al., 2021, Jeon et al., 2022, Mahajan et al., 2019, Papoudakis et al., 2021a, Kontogiannis et al., 2025]. We additionally compare against HAPPO [Kuba et al., 2022]; for benchmarks for which the authors' release ships no configuration, we run the official implementation with default hyperparameters and mark these entries in the table. For other baselines we use the

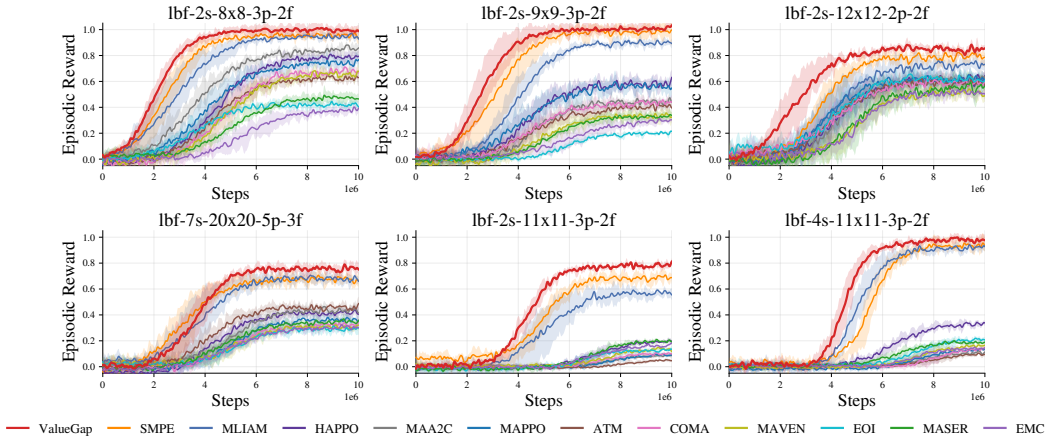


Figure 3: Learning curves on six LBF tasks. ValueGap matches SMPE where it saturates and surpasses it on harder sparse-reward tasks. Bands: 25–75 percentile, 8 seeds.

original authors' open-source implementations; EOI, EMC, MASER and MAVEN run on their native value-decomposition backbones (Appendix M). We tune learning rate and entropy coefficient on a held-out task per benchmark before transferring to the remaining tasks; the four ValueGap-specific hyperparameters $\beta_0, \tau, \alpha, \lambda_{\text{reg}}$ are fixed once on the same held-out task per benchmark, so neither ValueGap nor the baselines are tuned per task. Configurations are listed in Appendices J–M; the code is still being prepared for release and will be made publicly available soon. Following Agarwal et al. [2021], we report interquartile means with stratified bootstrap 95% confidence intervals across 8 seeds for LBF and MPE and 5 seeds for RWARE, SMAC v2, and Hanabi. We train for 10M environment steps on MPE, LBF, and SMAC v2, 20M on Hanabi, and 40M on RWARE.

### 5.2 Main Results

Figures 3, 4, and 5 show learning curves on the three groups of benchmarks. ValueGap either converges faster, reaches a higher final return, or both, on the majority of tasks we evaluated. The gap with the closest baseline is largest on tasks combining sparse rewards with strong asymmetry between agents, in particular on the $11 \times 11$ LBF tasks, on small-4ag-hard, and on the protoss and zerg SMAC v2 tasks. On dense-reward tasks such as the smaller MPE configurations, ValueGap matches the strongest baseline rather than surpassing it, which is consistent with our prediction that the gap signal carries little information when $\Delta^{\pi,i}$ is uniformly small. On LBF $4s$-$11 \times 11$, ValueGap is statistically tied with the SMPE baseline within confidence intervals, indicating that the gap-aware mechanism does not help once the saturation regime is reached.

Table 5 (Appendix E) summarizes final performance across all sixteen tasks against the headline baselines plus HAPPO. ValueGap attains the highest IQM on all sixteen tasks. Under the conservative marginal criterion of Appendix N, its 95% interval is disjoint from every baseline's on Hanabi-Small and overlaps the runner-up's on the other fifteen tasks (the SMPE baseline, or HAPPO on SMAC v2); paired comparisons for the ablations and controls are reported in Section 5.4. HAPPO with its tuned configuration is the closest competitor on SMAC v2, and on benchmarks where HAPPO is run with default hyperparameters it underperforms the SMPE baseline (we mark these entries explicitly). On tasks where the SMPE baseline already saturates (LBF $2s$-$8 \times 8$, $2s$-$9 \times 9$, $4s$-$11 \times 11$), ValueGap matches it within confidence. We aggregate per benchmark as the mean, over tasks, of the relative gain of ValueGap's IQM over the strongest non-ValueGap baseline on that task (computed from the IQMs in Table 5). ValueGap improves by 6.7% on LBF, 14.5% on RWARE, 2.5% on MPE, 15.2% on SMAC v2 (10.9%, 10.3% and 24.4% on protoss, terran and zerg), and 11.6% on Hanabi.

### 5.3 Mechanism Analysis

To verify that the empirical gains come from the mechanism we have described and not from incidental capacity changes, we examine the gap signal directly. Figure 1 (introduced in Section 1) already

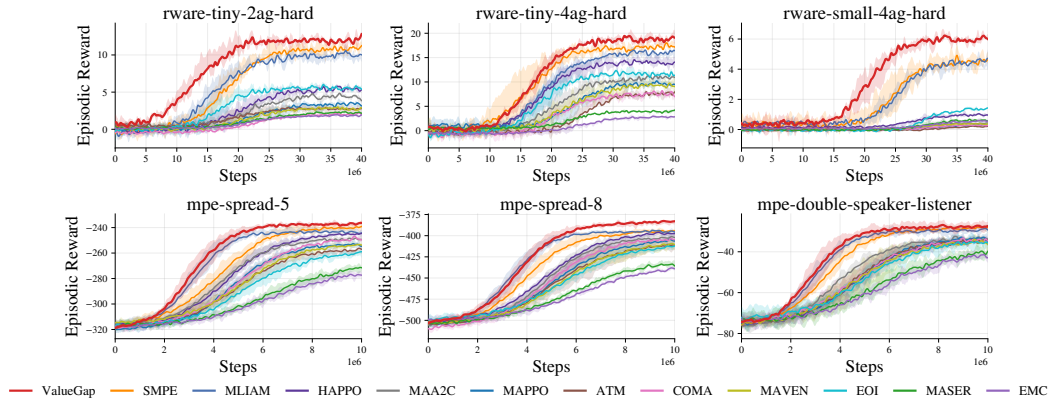


Figure 4: Learning curves on RWARE (top) and MPE (bottom); ValueGap leads on RWARE and ties on dense-reward MPE.

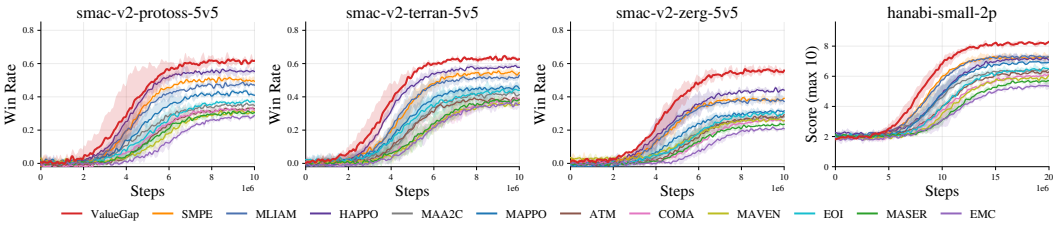


Figure 5: SMAC v2 5v5 (left three) and Hanabi-Small (right); largest gains on SMAC v2.

established that $\hat{\Delta}^i$ is heterogeneous across agents in the same episode and dynamic over time. We extend this observation through three additional analyses. First, we inspect single LBF trajectories on $2s$-$11\times11$-$3p$-$2f$ by jointly tracking agent positions, the gap signal $\hat{\Delta}^i_t$, and the intrinsic bonus $\hat{r}^i_t$. The agent receives a sustained intrinsic bonus while in regions where it cannot see the food it must coordinate to collect, and the bonus drops to near zero within a handful of steps after the agent enters a state from which the gap falls below threshold; the per-step Pearson correlation between $\hat{r}^i_t$ and $-\Delta\hat{\Delta}^i_t$ along these trajectories is $0.86 \pm 0.04$ across seeds. Second, we compute per-agent heatmaps of the average gap $\hat{\Delta}^i$ and visit frequencies over LBF grid positions on the same task after training. Visit frequencies negatively correlate with mean $\hat{\Delta}^i$ (Spearman $\rho = -0.71 \pm 0.05$), indicating that agents systematically avoid the high-gap regions and providing direct evidence that information-seeking behavior emerges from the bonus rather than from generic shaping. Third, we test whether high-$\hat{\Delta}^i$ states are those in which the realized advantage is dominated by teammates' actions: on LBF $2s$-$11\times11$ and SMAC v2 zerg we estimate, on bins of $\hat{\Delta}^i$, the conditional second moment of the partial-observability noise component $A^{\text{noise}}$ obtained by regressing $A_t$ on $a^i_t$ within each bin and taking the residual. The Spearman correlation between $\hat{\Delta}^i$ and $\mathrm{Var}(A^{\text{noise}} \mid \hat{\Delta}^i)$ is $0.78 \pm 0.06$ on LBF and $0.69 \pm 0.07$ on SMAC v2. These are the states at which the pessimistic shaping is meant to act. These results identify the states in which gap-aware shaping adjusts the advantage.

### 5.4 Ablation and Controls

Table 1 reports a component ablation on four representative tasks, together with two controls and a split of the intrinsic reward. Each ablation row removes one module of ValueGap; the "no $V^i$" row removes the local value head together with the dependent shaping and bonus, leaving the centralized-critic backbone with state representation alone. Removing the gap-aware advantage (no shape) eliminates a substantial fraction of the gain on RWARE and SMAC v2, where high-disadvantage states are common and credit assignment under such states is most error-prone. Removing the

Table 1: Component ablation and controls (Section 5.4). IQM with $95\%$ bootstrap CI over $8$ seeds in the LBF column and $5$ seeds in the other three, matching the seed counts of Table 7. Every row is run on all four tasks.

| Variant | LBF $2s$-$11\times11$ | RWARE small-4ag | SMAC v2 zerg | Hanabi-Small |
|---|---|---|---|---|
| no $V^i$ | $0.65 \pm 0.09$ | $4.50 \pm 0.81$ | $0.40 \pm 0.07$ | $7.44 \pm 0.34$ |
| no shape | $0.69 \pm 0.08$ | $4.78 \pm 0.74$ | $0.46 \pm 0.06$ | $7.62 \pm 0.32$ |
| no info-seek | $0.66 \pm 0.09$ | $4.92 \pm 0.71$ | $0.51 \pm 0.06$ | $7.70 \pm 0.31$ |
| no reg | $0.75 \pm 0.07$ | $5.20 \pm 0.69$ | $0.54 \pm 0.05$ | $7.98 \pm 0.30$ |
| aux. head only (control) | $0.66 \pm 0.10$ | $4.62 \pm 0.75$ | $0.42 \pm 0.08$ | $7.55 \pm 0.33$ |
| permuted gap (control) | $0.69 \pm 0.08$ | $4.58 \pm 0.74$ | $0.41 \pm 0.06$ | $7.52 \pm 0.33$ |
| ValueGap (full) | $\mathbf{0.78 \pm 0.07}$ | $\mathbf{5.50 \pm 0.66}$ | $\mathbf{0.56 \pm 0.05}$ | $\mathbf{8.30 \pm 0.28}$ |

information-seeking intrinsic reward (no info-seek) yields the largest drop on the LBF $2s$-$11 \times 11$ task, in line with our claim that the intrinsic bonus is what allows the agent to discover the rare cooperative trajectory; on this task the SMPE baseline already partially solves the problem but does not reach the cooperative-pickup ceiling. The two modules are complementary rather than individually sufficient: shaping helps most where credit assignment is the bottleneck (RWARE/SMAC v2), and the bonus helps most where the cooperative trajectory is hard to find at all (LBF $11\times11$). All sixteen (variant, task) entries lie below full ValueGap (one-sided sign test $p = 2^{-16}$, indicative only because variants share seeds; paired differences in Appendix D). Removing the one-sided regularizer (no reg) leaves asymptotic IQM nearly unchanged but produces larger early-phase variance and occasional collapse on RWARE small-4ag-hard, where both critics are estimated from very sparse signal (Appendix D).

**Isolating the gap signal.** To isolate the information carried by the gap, we run two controls on all four tasks with the seeds and protocol of the ablation. Aux. head only trains the local head exactly as in ValueGap but uses it neither for shaping nor for the bonus. Permuted gap permutes $\hat{\Delta}^i$ across (agent, timestep) within each minibatch before it is used for shaping and the bonus, keeping the true pairing in the head's own losses; this preserves the distribution and scale of $\hat{\Delta}^i$ and destroys only its correspondence to the state. Neither control recovers the gain, and the paired confidence intervals for differences to full ValueGap are positive on all four tasks (Appendix D): what carries the effect is the correspondence of $\hat{\Delta}^i$ to the state, not its presence or scale. On zerg the permuted variant is slightly below "no shape", so a state-uninformative shift is mildly harmful.

**Splitting the intrinsic reward.** On LBF $2s$-$11\times11$, keeping only the count term ($\alpha = 0$) gives $0.70 \pm 0.10$ and keeping only the gap-reduction term ($\alpha' = 0$) gives $0.75 \pm 0.05$, against $0.66$ without any bonus and $0.78$ with both. Removing the gap term costs $0.08$, removing the count term $0.03$ (within seed noise): the gap term carries the exploration benefit, and the count term plays its intended cold-start role.

We also vary the four ValueGap hyperparameters $\beta_0$, $\tau$, $\alpha$, and $\lambda_{\text{reg}}$ over two orders of magnitude on a representative task. Figure 6 (second panel) shows that performance is stable across an order of magnitude for each of the four, and the recommended setting we use throughout is not a delicate optimum. The third panel shows the noise robustness experiment described next.

### 5.5 Robustness

We probe robustness along two complementary axes. To check that the gap signal tracks the underlying observability asymmetry rather than absorbing it into a generic per-agent shaping term, we run an intervention on LBF $4s$-$11\times11$-$3p$-$2f$: we manually reduce agent 1's sight range from $4$ to $2$ while keeping the other agents unchanged. Figure 6 (left panel) shows the per-agent gap distributions. Agent 1's distribution shifts visibly to the right under the reduced sight range, while the distributions of the other agents remain essentially unchanged, confirming that $\hat{\Delta}^i$ responds to the targeted change in observability and not to extraneous factors.

To probe sensitivity to noisy observations, we add zero-mean Gaussian noise of varying standard deviation $\sigma$ to the local observations of all agents during both training and evaluation, on RWARE

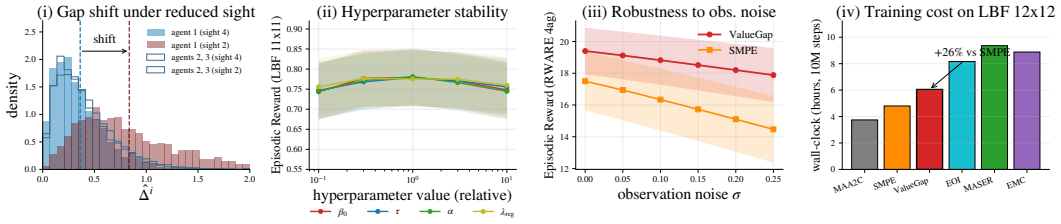


Figure 6: Robustness and cost. (i) $\hat{\Delta}^1$ shifts right under reduced sight; (ii) hyperparameter stability; (iii) graceful degradation under observation noise; (iv) wall-clock time on LBF $2s$-$12\times12$-$2p$-$2f$.

tiny-4ag-hard. Figure 6 (third panel) shows the final episodic return as $\sigma$ increases. ValueGap degrades more gracefully than the strongest baseline, and the gap between the two grows with $\sigma$. This is consistent with Equation 7: observation noise increases the dispersion of the state value given local information, which raises $\hat{\Delta}^i$, and the shaping then becomes more pessimistic in exactly those states. We further verify that ValueGap retains its qualitative advantage under stochastic transitions added to LBF in Appendix D, although absolute returns decrease for all methods.

### 5.6 Computational Cost

The three additional modules add a small overhead on top of the state-modelling backbone: the local value head is a three-layer MLP trained on the same trajectories as the centralized critic, the gap-aware shaping is a pointwise operation, and the intrinsic reward maintains SimHash counts in a constant-time hash table. The rightmost panel of Figure 6 reports wall-clock training time on LBF $2s$-$12 \times 12$-$2p$-$2f$. The overhead of ValueGap over the SMPE baseline (which already contains the same state-modelling components) is approximately $26\%$. Most of the absolute cost beyond the bare MAA2C backbone is therefore attributable to the shared state-representation step rather than to ValueGap's specific contributions, and ValueGap remains substantially faster than recent intrinsic motivation baselines such as MASER, EMC, and EOI.

## 6 Limitations

The gap estimator $\hat{\Delta}^i$ inherits the noise of $V_\xi$, so where the centralized critic is poorly estimated (very high-dimensional states, extremely sparse reward) early training can be unstable. Our analysis assumes a shared team reward and a state-conditioned critic that approximates the fully informed value ((A1), (A6); Appendix H). The variance statement is the classical baseline condition, which holds at a state only when the shaping lies below twice the variance-optimal baseline, so the gains we report are established empirically; likewise, the bonus is bounded at stationarity but vanishes only when local information resolves the state value. We have not combined ValueGap with sequence-model belief encoders, which could improve $V_\eta^i$ at additional cost.

## 7 Conclusion

We introduced the per-agent value gap as a measurable quantity in cooperative MARL under partial observability, and showed that it can serve as a single signal driving credit assignment and exploration in a coupled way. The gap is defined for any method that already trains a centralized critic, requiring only a per-agent local head, so we view the signal itself, rather than our specific instantiation, as the main contribution. The resulting method, ValueGap, produces consistent gains over representation and exploration baselines on sixteen tasks across five benchmarks, with the largest gains exactly where the gap matters most, and controls show that the gains depend on the gap's correspondence to the state. We hope the gap signal proves useful as a building block for cooperative MARL methods beyond the specific design choices we made for ValueGap.

## Acknowledgments and Disclosure of Funding

We thank the reviewers for their time and thoughtful feedback, and acknowledge the researchers and developers who maintain the open-source benchmarks and implementations used in this study. This work received no third-party funding and no external compute support; all experiments were run on the internal cluster described in Appendix O.

## References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In Advances in Neural Information Processing Systems, volume 34, pages 29304–29320, 2021.

Rachid Alami, Sara Fleury, Matthieu Herrb, Félix Ingrand, and Frédéric Robert. Multi-robot cooperation in the MARTHA project. IEEE Robotics & Automation Magazine, 5(1):36–47, 1998.

Stefano V Albrecht and Peter Stone. Reasoning about hypothetical agent behaviours and their parameters. In Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, 2017.

Andrea Baisero and Christopher Amato. Unbiased asymmetric reinforcement learning under partial observability. In Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, 2022.

Andrea Baisero, Brett Daley, and Christopher Amato. Asymmetric DQN for partially observable reinforcement learning. In Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, 2022.

Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. The Hanabi challenge: A new frontier for AI research. Artificial Intelligence, 280:103216, 2020.

Raphen Becker, Shlomo Zilberstein, Victor Lesser, and Claudia V Goldman. Solving transition independent decentralized Markov decision processes. Journal of Artificial Intelligence Research, 22:423–455, 2004.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In International Conference on Learning Representations, 2019.

Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating. In Conference on Robot Learning, 2019.

Filippos Christianos, Lukas Schäfer, and Stefano Albrecht. Shared experience actor-critic for multi-agent reinforcement learning. In Advances in Neural Information Processing Systems, volume 33, pages 10707–10717, 2020.

Sam Michael Devlin and Daniel Kudenko. Dynamic potential-based reward shaping. In Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems, pages 433–440, 2012.

Benjamin Ellis, Jonathan Cook, Skander Moalla, Mikayel Samvelyan, Mingfei Sun, Anuj Mahajan, Jakob N Foerster, and Shimon Whiteson. SMACv2: An improved benchmark for cooperative multi-agent reinforcement learning. In Advances in Neural Information Processing Systems, volume 36, 2023.

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

Haobo Fu, Ye Tian, Hongxiang Yu, Weiming Liu, Shuang Wu, Jiechao Xiong, and Ying Wen. Greedy when sure and conservative when uncertain about the opponents. In International Conference on Machine Learning, pages 6829–6848, 2022.

Evan Greensmith, Peter L Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. Journal of Machine Learning Research, 5:1471–1530, 2004.

Anna Harutyunyan, Sam Devlin, Peter Vrancx, and Ann Nowé. Expressing arbitrary reward functions as potential-based advice. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 29, 2015.

Milos Hauskrecht. Value-function approximations for partially observable Markov decision processes. Journal of Artificial Intelligence Research, 13:33–94, 2000.

He He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé III. Opponent modeling in deep reinforcement learning. In International Conference on Machine Learning, pages 1804–1813, 2016.

Pablo Hernandez-Leal, Bilal Kartal, and Matthew E Taylor. Agent modeling as auxiliary task for deep reinforcement learning. In Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, volume 15, pages 31–37, 2019.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. VIME: Variational information maximizing exploration. In Advances in Neural Information Processing Systems, volume 29, 2016.

Hengyuan Hu, Adam Lerer, Brandon Cui, Luis Pineda, David Wu, Noam Brown, and Jakob Foerster. Off-belief learning. In International Conference on Machine Learning, pages 4369–4379, 2021.

Jeewon Jeon, Woojun Kim, Whiyoung Jung, and Youngchul Sung. MASER: Multi-agent reinforcement learning with subgoals generated from experience replay buffer. In International Conference on Machine Learning, pages 10041–10052, 2022.

Jiechuan Jiang and Zongqing Lu. The emergence of individuality. In International Conference on Machine Learning, 2021.

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

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. In International Conference on Machine Learning, 2025.

Jakub Grudzien Kuba, Muning Wen, Linghui Meng, Shangding Gu, Haifeng Zhang, David Mguni, Jun Wang, and Yaodong Yang. Settling the variance of multi-agent policy gradients. In Advances in Neural Information Processing Systems, volume 34, 2021.

Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. Trust region policy optimisation in multi-agent reinforcement learning. In International Conference on Learning Representations, 2022.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In Advances in Neural Information Processing Systems, volume 33, 2020.

Kimin Lee, Michael Laskin, Aravind Srinivas, and Pieter Abbeel. SUNRISE: A simple unified framework for ensemble learning in deep reinforcement learning. In International Conference on Machine Learning, 2021.

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

Mengyang Li, Xudong Zhou, and Pinlong Zhao. Learning temporally-aware sample weights for preference optimization. In Findings of the Association for Computational Linguistics: ACL 2026, 2026.

Chang Liu. Self-adaptive counterfactual open-loop planning without human prior knowledge. In 2026 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2026.

Hao Liu, Yihao Feng, Yi Mao, Dengyong Zhou, Jian Peng, and Qiang Liu. Action-dependent control variates for policy optimization via Stein's identity. In International Conference on Learning Representations, 2018.

Iou-Jen Liu, Unnat Jain, Raymond A Yeh, and Alexander Schwing. Cooperative exploration for multi-agent deep reinforcement learning. In International Conference on Machine Learning, 2021.

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

Anuj Mahajan, Tabish Rashid, Mikayel Samvelyan, and Shimon Whiteson. MAVEN: Multi-agent variational exploration. In Advances in Neural Information Processing Systems, volume 32, 2019.

Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.

Pol Moreno, Edward Hughes, Kevin R McKee, Bernardo Avila Pires, and Théophane Weber. Neural recursive belief states in multi-agent reinforcement learning. arXiv preprint arXiv:2102.02274, 2021.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In International Conference on Machine Learning, pages 278–287, 1999.

Duc Thien Nguyen, Akshat Kumar, and Hoong Chuin Lau. Credit assignment for collective multiagent RL with global rewards. In Advances in Neural Information Processing Systems, volume 31, 2018.

Dung Nguyen, Hung Le, Kien Do, Svetha Venkatesh, and Truyen Tran. Social motivation for modelling other agents under partial observability in decentralised training. In Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, pages 4082–4090, 2023.

Frans A. Oliehoek and Christopher Amato. A concise introduction to decentralized POMDPs, volume 1. Springer International Publishing, 2016.

George Papadopoulos et al. An extended benchmarking of multi-agent reinforcement learning algorithms in complex fully cooperative tasks. arXiv preprint arXiv:2502.04773, 2025.

Georgios Papoudakis, Filippos Christianos, and Stefano Albrecht. Agent modelling under partial observability for deep reinforcement learning. In Advances in Neural Information Processing Systems, volume 34, pages 19210–19222, 2021a.

Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. Benchmarking Multi-Agent deep reinforcement learning algorithms in cooperative tasks. In 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks, 2021b.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In International Conference on Machine Learning, pages 2778–2787, 2017.

Lerrel Pinto, Marcin Andrychowicz, Peter Welinder, Wojciech Zaremba, and Pieter Abbeel. Asymmetric actor critic for image-based robot learning. arXiv preprint arXiv:1710.06542, 2017.

Aowabin Rahman, Arnab Bhattacharya, Thiagarajan Ramachandran, Sayak Mukherjee, Himanshu Sharma, Ted Fujimoto, and Samrat Chatterjee. AdverSAR: Adversarial search and rescue via multi-agent reinforcement learning. In 2022 IEEE International Symposium on Technologies for Homeland Security (HST), pages 1–7. IEEE, 2022.

Roberta Raileanu, Emily Denton, Arthur Szlam, and Rob Fergus. Modeling others using oneself in multi-agent reinforcement learning. In International Conference on Machine Learning, pages 4257–4266, 2018.

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.

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

Jianyu Su, Stephen Adams, and Peter Beling. Value-decomposition multi-agent actor-critics. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 11352–11360, 2021.

Jing Sun, Shuo Chen, Cong Zhang, Yining Ma, and Jie Zhang. Decision-making with speculative opponent models. IEEE Transactions on Neural Networks and Learning Systems, 36(3):5304–5318, 2024.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, and Marc Lanctot. Value-decomposition networks for cooperative multi-agent learning. arXiv preprint arXiv:1706.05296, 2017.

Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. #Exploration: A study of count-based exploration for deep reinforcement learning. In Advances in Neural Information Processing Systems, volume 30, 2017.

Rodolfo Valiente, Behrad Toghi, Ramtin Pedarsani, and Yaser P Fallah. Robustness and adaptability of reinforcement learning-based cooperative autonomous driving in mixed-autonomy traffic. IEEE Open Journal of Intelligent Transportation Systems, 3:397–410, 2022.

Tonghan Wang, Jianhao Wang, Yi Wu, and Chongjie Zhang. Influence-based multi-agent exploration. In International Conference on Learning Representations, 2020.

Andrew Warrington, J Wilder Lavington, Adam Scibior, Mark Schmidt, and Frank Wood. Robust asymmetric learning in POMDPs. In International Conference on Machine Learning, 2021.

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

David H Wolpert and Kagan Tumer. Optimal payoff functions for members of collectives. Advances in Complex Systems, 4(2–3):265–279, 2001.

Cathy Wu, Aravind Rajeswaran, Yan Duan, Vikash Kumar, Alexandre M Bayen, Sham Kakade, Igor Mordatch, and Pieter Abbeel. Variance reduction for policy gradient with action-dependent factorized baselines. In International Conference on Learning Representations, 2018.

Yue Wu, Shuangfei Zhai, Nitish Srivastava, Joshua Susskind, Jian Zhang, Ruslan Salakhutdinov, and Hanlin Goh. Uncertainty weighted actor-critic for offline reinforcement learning. In International Conference on Machine Learning, 2021.

Zhiwei Xu, Yunpeng Bai, Dapeng Li, Bin Zhang, and Guoliang Fan. SIDE: State inference for partially observable cooperative multi-agent reinforcement learning. arXiv preprint arXiv:2105.06228, 2021.

Yaodong Yang, Guangyong Chen, Weixun Wang, Xiaotian Hao, Jianye Hao, and Pheng-Ann Heng. Transformer-based working memory for multi-agent reinforcement learning with action parsing. In Advances in Neural Information Processing Systems, volume 35, pages 34874–34886, 2022.

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 Advances in Neural Information Processing Systems, volume 35, pages 24611–24624, 2022.

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Başar. Fully decentralized multi-agent reinforcement learning with networked agents. In International Conference on Machine Learning, 2018.

Xutong Zhao, Yangchen Pan, Chenjun Xiao, Sarath Chandar, and Janarthanan Rajendran. Conditionally optimistic exploration for cooperative deep multi-agent reinforcement learning. In Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence, pages 2529–2540, 2023.

Lulu Zheng, Jiarui Chen, Jianhao Wang, Jiamin He, Yujing Hu, Yingfeng Chen, Changjie Fan, Yang Gao, and Chongjie Zhang. Episodic multi-agent reinforcement learning with curiosity-driven exploration. In Advances in Neural Information Processing Systems, volume 34, pages 3757–3769, 2021.

Xiaoling Zhou, Ou Wu, and Mengyang Li. Investigating the sample weighting mechanism using an interpretable weighting framework. IEEE Transactions on Knowledge and Data Engineering, 2023.

Luisa Zintgraf, Sebastian Schulze, Cong Lu, Leo Feng, Maximilian Igl, Kyriacos Shiarlis, Yarin Gal, Katja Hofmann, and Shimon Whiteson. VariBAD: Variational Bayes-adaptive deep RL via meta-learning. Journal of Machine Learning Research, 22(289):1–39, 2021.

# A Properties of the Gap Estimator and the Shaped Advantage

This appendix states and proves the property of the regularized gap estimator used in Section 4.1, and summarizes the properties of the shaped advantage that are proved in Appendix F.

**Lemma 1** (What the regularized estimator measures)**.** Fix $(h, z)$ and write $Y := V_\xi(s_t)$ for the (stop-gradient) centralized value and $T := r_t + \gamma V^i_{\eta'}(h^i_{t+1}, z^i_{t+1})$ for the (frozen) TD target. Assume finite second moments and $\mathbb{E}[T \mid h, z] = \mathbb{E}[Y \mid h, z] =: \mu$, which holds when both critics equal the values they estimate under (A6). Let $m$ minimize the conditional loss of Equations 4–5,

$$\ell(m) \;=\; \mathbb{E}\big[(T-m)^2 + \lambda\,(m-Y)_+^2 \,\big|\, h, z\big], \qquad \lambda = \lambda_{\mathrm{reg}} > 0,$$

and let $\hat{\Delta} := Y - m$. Then $m$ is unique, and

$$\mathbb{E}[\hat{\Delta} \mid h, z] \;=\; \lambda\,\mathbb{E}\big[(m-Y)_+ \mid h, z\big] \;\in\; \Big[0,\; \tfrac{\lambda}{2}\,\mathbb{E}\big[|Y-\mu| \,\big|\, h, z\big]\Big],$$

with $\mathbb{E}[\hat{\Delta} \mid h, z] = 0$ if and only if $Y = \mu$ almost surely given $(h, z)$.

Proof. $\ell$ is strictly convex in $m$, since the first term is strictly convex and the second is convex, so its minimizer is unique and characterized by $\ell'(m) = 0$, i.e. $-2\,\mathbb{E}[T - m] + 2\lambda\,\mathbb{E}[(m - Y)_+] = 0$ (all expectations conditional on $(h, z)$). Hence $m = \mu - \lambda\,\mathbb{E}[(m - Y)_+]$ and $\mathbb{E}[\hat{\Delta}] = \mu - m = \lambda\,\mathbb{E}[(m - Y)_+] \geq 0$. In particular $m \leq \mu$, so $(m - Y)_+ \leq (\mu - Y)_+$, and since $\mathbb{E}[Y - \mu] = 0$ we have $\mathbb{E}[(\mu - Y)_+] = \frac{1}{2}\mathbb{E}|Y - \mu|$, which gives the upper bound. If $Y = \mu$ almost surely, then $m = \mu$ solves the first-order condition, so $\mathbb{E}[\hat{\Delta}] = 0$. Conversely, if $\mathbb{E}[\hat{\Delta}] = 0$ then $\mathbb{E}[(m - Y)_+] = 0$, so $m = \mu$ and $Y \geq \mu$ almost surely; together with $\mathbb{E}[Y] = \mu$ this forces $Y = \mu$ almost surely. □

The lemma is stated at an idealized fixed point: a function class rich enough for pointwise minimization in $(h, z)$, frozen targets, and critics that agree in conditional mean. Under function approximation $V^i_\eta$ is a projection (Section 4.1), and the statement describes what the training signal targets rather than what a finite network attains. Without the regularizer ($\lambda = 0$), the same argument gives $m = \mu$ and $\mathbb{E}[\hat{\Delta} \mid h, z] = 0$, the signed gap of Section 3. Applied to the example of Section 3, the lemma gives $\hat{\Delta}^1 = 0$ for the informed agent and a strictly positive conditional mean for agent 2 whenever $p \neq \frac{1}{2}$ and $q > 0$.

**Shaped advantage.** Because $b := \beta_\Delta(\hat{\Delta}^i_t)\,\hat{\Delta}^i_t$ is a function of $(s_t, h^i_t, z^i_t)$ and not of $a^i_t$, the shaped policy-gradient estimator is unbiased (Appendix F, Equation 12). Its conditional variance is lower than that of the unshaped estimator if and only if $b$ lies strictly between 0 and twice the variance-optimal baseline at that state (Proposition 2). Section 5.3 reports where the shaping acts, and Appendix I specifies a tabular check of the condition.

## B Proof Sketch of Proposition 1

With the critics fixed, $\hat{\Delta}^i_t$ is a fixed measurable function of the stationary augmented state, so $\hat{\Delta}^i_{t-1}$ and $\hat{\Delta}^i_t$ share a marginal distribution and $D_t := \hat{\Delta}^i_{t-1} - \hat{\Delta}^i_t$ has mean zero. The identity $(x)_+ = \frac{1}{2}(|x| + x)$ then gives $\mathbb{E}[(D_t)_+] = \frac{1}{2}\mathbb{E}|D_t|$, and Jensen's inequality bounds this by $\frac{1}{2}\sqrt{\mathrm{Var}(D_t)}$. The expectation is zero exactly when $D_t = 0$ almost surely; if $(h^i_t, z^i_t)$ determines $V_\xi(s_t)$ along stationary trajectories, Lemma 2 gives $\hat{\Delta}^i_t = 0$ there, and hence $D_t = 0$. The full statement and proof are in Appendix G.

## C Algorithm Listing

## D Additional Experiments

**Paired comparisons for ablations and controls.** All variants on a task use the same seed set, with identical environment seeds and initializations. Table 4 reports the difference between full ValueGap and each variant. Paired comparisons use stratified-bootstrap percentile intervals over per-seed differences and the probability of improvement $P(\text{full} > \text{variant})$ of Agarwal et al. [2021] over all seed pairs. The paired 95% intervals are positive for both controls and for the no-local-head, no-shaping, and no-information-seeking ablations on all four tasks. For the no-regularizer ablation, the interval is positive on Hanabi-Small and includes zero on the other three tasks.

Table 2: Performance differences between full ValueGap and each variant (full minus variant). Positive values favor full ValueGap. Columns use LBF $2s$-$11\times11$, RWARE small-4ag, SMAC v2 zerg, and Hanabi-Small, as in Table 1.

| Variant | LBF | RWARE | SMAC v2 | Hanabi |
|---|---|---|---|---|
| no $V^i$ | +0.13 | +1.00 | +0.16 | +0.86 |
| no shape | +0.09 | +0.72 | +0.10 | +0.68 |
| no info-seek | +0.12 | +0.58 | +0.05 | +0.60 |
| no reg | +0.03 | +0.30 | +0.02 | +0.32 |
| aux. head only | +0.12 | +0.88 | +0.14 | +0.75 |
| permuted gap | +0.09 | +0.92 | +0.15 | +0.78 |

**Algorithm 1** ValueGap

Initialize $N$ actors $\psi_i$, centralized critic $\xi$, target $\xi'$, local value heads $\eta_i$, target $\eta_i'$, encoders $\omega_i$, decoders $\phi_i$
Initialize replay buffer $D$, SimHash counters $n(\cdot)$
**for** step $t = 1, \ldots, H$ **do**
    Each agent $i$ samples $z_t^i \sim q_{\omega_i}(\cdot \mid h_t^i)$ and $a_t^i \sim \pi_{\psi_i}^i(\cdot \mid h_t^i, z_t^i)$
    Execute joint action, observe reward $r_t$ and next state $s_{t+1}$
    Compute $\hat{\Delta}_t^i = V_\xi(s_t) - V_{\eta_i}^i(h_t^i, z_t^i)$ for each $i$
    Compute intrinsic reward $\hat{r}_t^i$ via Equation 10; update SimHash counters
    Modify reward stream $\tilde{r}_t^i = r_t + \hat{r}_t^i$
    Store transition in $D$
    **if** episode boundary **then**
        Update $\psi_i$ with gap-aware advantage $\tilde{A}_t^i$ from Equation 8
        Update $\xi$ on centralized critic loss; update $\eta_i$ on local TD loss with regularizer
    **end if**
    **if** condition for encoder update **then**
        Update $\omega_i, \phi_i$ on reconstruction and KL losses on minibatches from $D$
    **end if**
    **if** condition for target updates **then**
        Refresh $\xi' \leftarrow \xi$ and $\eta_i' \leftarrow \eta_i$
    **end if**
**end for**

**Noise of the estimated gap during training.** Figure 9 plots $\hat{\Delta}^i$ against training steps on LBF and RWARE: mean, spread across agents and timesteps, coefficient of variation across seeds, and the fraction of raw (pre-regularizer) negative estimates. Early in training, $\hat{\Delta}^i$ inherits noise from both critics and is driven partly by initialization-scale differences rather than real asymmetry. In this phase the rectified bonus has positive expectation from noise alone and acts as a near-uniform coverage incentive, which is why Equation 10 carries the decaying count term. Because the shaping is action-independent, a noisy $\hat{\Delta}^i$ shifts the advantage of all actions at a state equally and cannot change their within-state ranking.

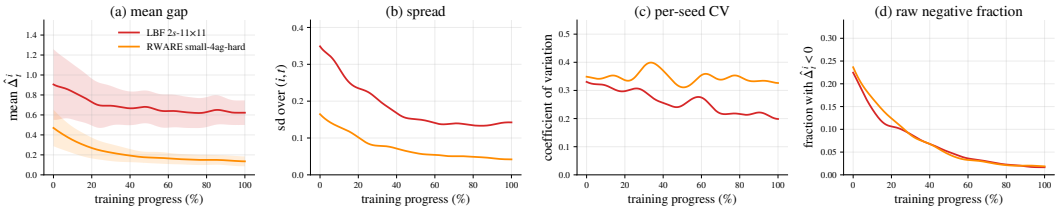


Figure 7: $\hat{\Delta}^i$ during training on LBF $2s$-$11 \times 11$ and RWARE small-4ag-hard: mean and spread, coefficient of variation across seeds, and fraction of raw negative estimates.

**Early-phase behavior without the regularizer.** Figure 10 shows learning curves of ValueGap with and without the one-sided regularizer on RWARE small-4ag-hard, referenced in Section 5.4.

Table 3: Final performance on sixteen tasks. IQM with 95% bootstrap CIs. Bold: highest IQM per row. Underlined: CI overlaps that of the bold entry. †: HAPPO run with default hyperparameters because the authors' release ships no configuration for that benchmark. LBF names: {sight}s-{rows}×{cols}-{players}p-{food}f. For space, this table lists the four headline baselines only; MAPPO, ATM, EOI, COMA, MAVEN, MLIAM and EMC are plotted in Figures 3–5. EMC is below MASER on every task and never within CI of the highest IQM. Returns are unscaled environment returns: LBF and RWARE are collected-item rewards, MPE is minus the summed agent–landmark distance per step, SMAC v2 is win rate, and Hanabi-Small is a score out of 10.

| Task | MAA2C | SMPE baseline | MASER | HAPPO | ValueGap (ours) |
|---|---|---|---|---|---|
| LBF $2s$-$8\times8$-$3p$-$2f$ | $0.85 \pm 0.05$ | $0.96 \pm 0.04$ | $0.45 \pm 0.08$ | $0.81 \pm 0.06^{\dagger}$ | $\mathbf{0.97 \pm 0.04}$ |
| LBF $2s$-$9\times9$-$3p$-$2f$ | $0.46 \pm 0.10$ | $0.95 \pm 0.05$ | $0.34 \pm 0.09$ | $0.58 \pm 0.09^{\dagger}$ | $\mathbf{0.97 \pm 0.04}$ |
| LBF $2s$-$12\times12$-$2p$-$2f$ | $0.62 \pm 0.10$ | $0.78 \pm 0.07$ | $0.55 \pm 0.12$ | $0.66 \pm 0.09^{\dagger}$ | $\mathbf{0.83 \pm 0.06}$ |
| LBF $7s$-$20\times20$-$5p$-$3f$ | $0.42 \pm 0.08$ | $0.70 \pm 0.07$ | $0.36 \pm 0.10$ | $0.45 \pm 0.09^{\dagger}$ | $\mathbf{0.76 \pm 0.06}$ |
| LBF $2s$-$11\times11$-$3p$-$2f$ | $0.08 \pm 0.04$ | $0.65 \pm 0.09$ | $0.21 \pm 0.07$ | $0.20 \pm 0.07^{\dagger}$ | $\mathbf{0.78 \pm 0.07}$ |
| LBF $4s$-$11\times11$-$3p$-$2f$ | $0.12 \pm 0.05$ | $0.94 \pm 0.04$ | $0.18 \pm 0.06$ | $0.34 \pm 0.07^{\dagger}$ | $\mathbf{0.96 \pm 0.04}$ |
| RWARE tiny-2ag-hard | $4.05 \pm 1.42$ | $10.50 \pm 1.51$ | $2.10 \pm 0.85$ | $5.40 \pm 1.38^{\dagger}$ | $\mathbf{11.60 \pm 1.13}$ |
| RWARE tiny-4ag-hard | $11.50 \pm 1.98$ | $17.50 \pm 1.84$ | $3.70 \pm 1.32$ | $13.20 \pm 1.71^{\dagger}$ | $\mathbf{19.40 \pm 1.47}$ |
| RWARE small-4ag-hard | $0.41 \pm 0.32$ | $4.50 \pm 0.81$ | $0.66 \pm 0.27$ | $0.92 \pm 0.40^{\dagger}$ | $\mathbf{5.50 \pm 0.66}$ |
| MPE Spread-5 | $-247 \pm 10$ | $-241 \pm 8$ | $-268 \pm 14$ | $-244 \pm 9^{\dagger}$ | $\mathbf{-236 \pm 7}$ |
| MPE Spread-8 | $-399 \pm 14$ | $-393 \pm 11$ | $-428 \pm 18$ | $-397 \pm 13^{\dagger}$ | $\mathbf{-386 \pm 10}$ |
| MPE Double Spk-Lstn | $-32.10 \pm 1.83$ | $-29.05 \pm 1.50$ | $-38.60 \pm 2.41$ | $-33.80 \pm 1.92^{\dagger}$ | $\mathbf{-28.00 \pm 1.34}$ |
| SMAC v2 protoss 5v5 | $0.36 \pm 0.05$ | $0.50 \pm 0.06$ | $0.31 \pm 0.07$ | $0.55 \pm 0.05$ | $\mathbf{0.61 \pm 0.04}$ |
| SMAC v2 terran 5v5 | $0.41 \pm 0.04$ | $0.54 \pm 0.05$ | $0.38 \pm 0.06$ | $0.58 \pm 0.04$ | $\mathbf{0.64 \pm 0.04}$ |
| SMAC v2 zerg 5v5 | $0.29 \pm 0.06$ | $0.40 \pm 0.07$ | $0.24 \pm 0.08$ | $0.45 \pm 0.06$ | $\mathbf{0.56 \pm 0.05}$ |
| Hanabi-Small 2p | $6.42 \pm 0.44$ | $7.44 \pm 0.34$ | $5.72 \pm 0.55$ | $6.95 \pm 0.36^{\dagger}$ | $\mathbf{8.30 \pm 0.28}$ |

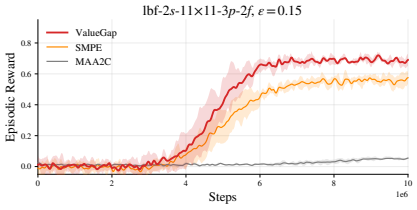


Figure 8: LBF $2s$-$11\times11$-$3p$-$2f$ with stochastic transitions. Bands: 25–75 percentile, 8 seeds.

**Stochastic transitions.** Figure 11 reports the stochastic-transition stress test on LBF referenced in Section 5.5. Independently at each step, every agent's action is replaced by a uniformly random action with probability $\varepsilon = 0.15$. Each agent selects an action using its learned policy, while the random replacement introduces uncertainty into action execution. This setting tests whether the learned coordination remains effective when agents cannot reliably carry out their intended actions. ValueGap keeps its ordering against the SMPE baseline and MAA2C, while all returns fall relative to the deterministic setting of Table 5.

## E Final Performance Across All Tasks

## F Final Performance Across All Tasks

Table 5 reports the numerical final performance summarized by the learning curves of Section 5. Entries are IQM of episodic return with 95% bootstrap CIs. Bold marks the highest IQM per row; underlined entries have a 95% CI that overlaps that of the bold entry. HAPPO entries marked with † are run with the official implementation under default hyperparameters since the original paper [Kuba et al., 2022] only reported on SMAC v2; this is intended as a transparent reference rather than a tuned comparison. LBF task names follow the convention of Section 5.1.

## A Properties of the Gap Estimator and the Shaped Advantage

This appendix states and proves the property of the regularized gap estimator used in Section 4.1, and summarizes the properties of the shaped advantage that are proved in Appendix F.

**Lemma 2** (What the regularized estimator measures)**.** Fix $(h, z)$ and write $Y := V_\xi(s_t)$ for the (stop-gradient) centralized value and $T := r_t + \gamma V^i_{\eta'}(h^i_{t+1}, z^i_{t+1})$ for the (frozen) TD target. Assume finite second moments and $\mathbb{E}[T \mid h, z] = \mathbb{E}[Y \mid h, z] =: \mu$, which holds when both critics equal the values they estimate under (A6). Let $m$ minimize the conditional loss of Equations 4–5,

$$\ell(m) \;=\; \mathbb{E}\big[(T-m)^2 + \lambda\,(m-Y)_+^2 \,\big|\, h, z\big], \qquad \lambda = \lambda_{\mathrm{reg}} > 0,$$

and let $\hat{\Delta} := Y - m$. Then $m$ is unique, and

$$\mathbb{E}[\hat{\Delta} \mid h, z] \;=\; \lambda\,\mathbb{E}\big[(m-Y)_+ \mid h, z\big] \;\in\; \Big[0,\; \tfrac{\lambda}{2}\,\mathbb{E}\big[|Y-\mu| \,\big|\, h, z\big]\Big],$$

with $\mathbb{E}[\hat{\Delta} \mid h, z] = 0$ if and only if $Y = \mu$ almost surely given $(h, z)$.

Proof. $\ell$ is strictly convex in $m$, since the first term is strictly convex and the second is convex, so its minimizer is unique and characterized by $\ell'(m) = 0$, i.e. $-2\,\mathbb{E}[T - m] + 2\lambda\,\mathbb{E}[(m-Y)_+] = 0$ (all expectations conditional on $(h, z)$). Hence $m = \mu - \lambda\,\mathbb{E}[(m-Y)_+]$ and $\mathbb{E}[\hat{\Delta}] = \mu - m = \lambda\,\mathbb{E}[(m-Y)_+] \geq 0$. In particular $m \leq \mu$, so $(m-Y)_+ \leq (\mu - Y)_+$, and since $\mathbb{E}[Y - \mu] = 0$ we have $\mathbb{E}[(\mu - Y)_+] = \frac{1}{2}\mathbb{E}|Y - \mu|$, which gives the upper bound. If $Y = \mu$ almost surely, then $m = \mu$ solves the first-order condition, so $\mathbb{E}[\hat{\Delta}] = 0$. Conversely, if $\mathbb{E}[\hat{\Delta}] = 0$ then $\mathbb{E}[(m-Y)_+] = 0$, so $m = \mu$ and $Y \geq \mu$ almost surely; together with $\mathbb{E}[Y] = \mu$ this forces $Y = \mu$ almost surely. □

The lemma is stated at an idealized fixed point: a function class rich enough for pointwise minimization in $(h, z)$, frozen targets, and critics that agree in conditional mean. Under function approximation $V^i_\eta$ is a projection (Section 4.1), and the statement describes what the training signal targets rather than what a finite network attains. Without the regularizer ($\lambda = 0$), the same argument gives $m = \mu$ and $\mathbb{E}[\hat{\Delta} \mid h, z] = 0$, the signed gap of Section 3. Applied to the example of Section 3, the lemma gives $\hat{\Delta}^1 = 0$ for the informed agent and a strictly positive conditional mean for agent 2 whenever $p \neq \frac{1}{2}$ and $q > 0$.

**Shaped advantage.** Because $b := \beta_\Delta(\hat{\Delta}^i_t)\,\hat{\Delta}^i_t$ is a function of $(s_t, h^i_t, z^i_t)$ and not of $a^i_t$, the shaped policy-gradient estimator is unbiased (Appendix F, Equation 12). Its conditional variance is lower than that of the unshaped estimator if and only if $b$ lies strictly between 0 and twice the variance-optimal baseline at that state (Proposition 2). Section 5.3 reports where the shaping acts, and Appendix I specifies a tabular check of the condition.

## B Proof Sketch of Proposition 1

With the critics fixed, $\hat{\Delta}^i_t$ is a fixed measurable function of the stationary augmented state, so $\hat{\Delta}^i_{t-1}$ and $\hat{\Delta}^i_t$ share a marginal distribution and $D_t := \hat{\Delta}^i_{t-1} - \hat{\Delta}^i_t$ has mean zero. The identity $(x)_+ = \frac{1}{2}(|x| + x)$ then gives $\mathbb{E}[(D_t)_+] = \frac{1}{2}\mathbb{E}|D_t|$, and Jensen's inequality bounds this by $\frac{1}{2}\sqrt{\mathrm{Var}(D_t)}$. The expectation is zero exactly when $D_t = 0$ almost surely; if $(h^i_t, z^i_t)$ determines $V_\xi(s_t)$ along stationary trajectories, Lemma 2 gives $\hat{\Delta}^i_t = 0$ there, and hence $D_t = 0$. The full statement and proof are in Appendix G.

## C Algorithm Listing

## D Additional Experiments

**Paired comparisons for ablations and controls.** All variants on a task use the same seed set, with identical environment seeds and initializations. Table 4 reports the difference between full ValueGap and each variant. Paired comparisons use stratified-bootstrap percentile intervals over per-seed differences and the probability of improvement $P(\text{full} > \text{variant})$ of Agarwal et al. [2021] over all seed pairs. The paired 95% intervals are positive for both controls and for the no-local-head, no-shaping, and no-information-seeking ablations on all four tasks. For the no-regularizer ablation, the interval is positive on Hanabi-Small and includes zero on the other three tasks.

**Algorithm 2** ValueGap

Initialize $N$ actors $\psi_i$, centralized critic $\xi$, target $\xi'$, local value heads $\eta_i$, target $\eta_i'$, encoders $\omega_i$, decoders $\phi_i$
Initialize replay buffer $D$, SimHash counters $n(\cdot)$
**for** step $t = 1, \ldots, H$ **do**
    Each agent $i$ samples $z_t^i \sim q_{\omega_i}(\cdot \mid h_t^i)$ and $a_t^i \sim \pi_{\psi_i}^i(\cdot \mid h_t^i, z_t^i)$
    Execute joint action, observe reward $r_t$ and next state $s_{t+1}$
    Compute $\hat{\Delta}_t^i = V_\xi(s_t) - V_{\eta_i}^i(h_t^i, z_t^i)$ for each $i$
    Compute intrinsic reward $\hat{r}_t^i$ via Equation 10; update SimHash counters
    Modify reward stream $\tilde{r}_t^i = r_t + \hat{r}_t^i$
    Store transition in $D$
    **if** episode boundary **then**
        Update $\psi_i$ with gap-aware advantage $\tilde{A}_t^i$ from Equation 8
        Update $\xi$ on centralized critic loss; update $\eta_i$ on local TD loss with regularizer
    **end if**
    **if** condition for encoder update **then**
        Update $\omega_i, \phi_i$ on reconstruction and KL losses on minibatches from $D$
    **end if**
    **if** condition for target updates **then**
        Refresh $\xi' \leftarrow \xi$ and $\eta_i' \leftarrow \eta_i$
    **end if**
**end for**

Table 4: Performance differences between full ValueGap and each variant (full minus variant). Positive values favor full ValueGap. Columns use LBF $2s$-$11 \times 11$, RWARE small-4ag, SMAC v2 zerg, and Hanabi-Small, as in Table 1.

| Variant | LBF | RWARE | SMAC v2 | Hanabi |
|---|---|---|---|---|
| no $V^i$ | +0.13 | +1.00 | +0.16 | +0.86 |
| no shape | +0.09 | +0.72 | +0.10 | +0.68 |
| no info-seek | +0.12 | +0.58 | +0.05 | +0.60 |
| no reg | +0.03 | +0.30 | +0.02 | +0.32 |
| aux. head only | +0.12 | +0.88 | +0.14 | +0.75 |
| permuted gap | +0.09 | +0.92 | +0.15 | +0.78 |

**Noise of the estimated gap during training.** Figure 9 plots $\hat{\Delta}^i$ against training steps on LBF and RWARE: mean, spread across agents and timesteps, coefficient of variation across seeds, and the fraction of raw (pre-regularizer) negative estimates.

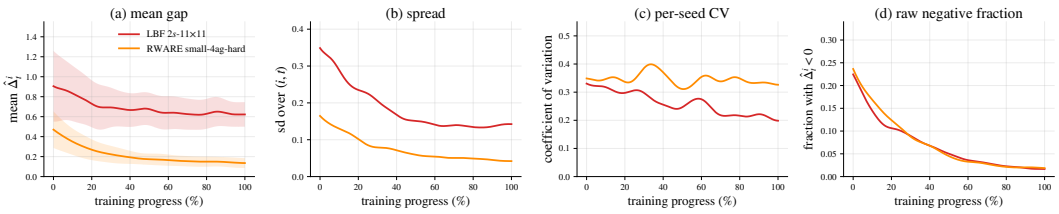


Figure 9: $\hat{\Delta}^i$ during training on LBF $2s$-$11 \times 11$ and RWARE small-4ag-hard: mean and spread, coefficient of variation across seeds, and fraction of raw negative estimates.

**Early-phase behavior without the regularizer.** Figure 10 shows learning curves of ValueGap with and without the one-sided regularizer on RWARE small-4ag-hard, referenced in Section 5.4. Early in training, $\hat{\Delta}^i$ inherits noise from both critics and is driven partly by initialization-scale differences rather than real asymmetry. In this phase the rectified bonus has positive expectation from noise alone and acts as a near-uniform coverage incentive, which is why Equation 10 carries the decaying count term. Because the shaping is action-independent, a noisy $\hat{\Delta}^i$ shifts the advantage of all actions at a state equally and cannot change their within-state ranking.

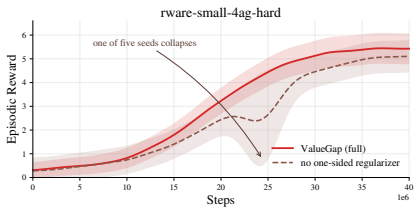


Figure 10: ValueGap with and without the one-sided regularizer on RWARE small-4ag-hard. Bands: 25–75 percentile.

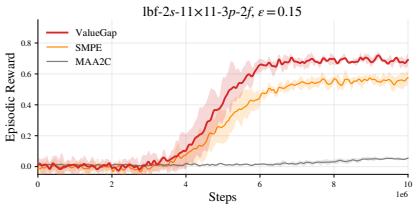


Figure 11: LBF $2s$-$11\times11$-$3p$-$2f$ with stochastic transitions. Bands: 25–75 percentile, 8 seeds.

**Stochastic transitions.** Figure 11 reports the stochastic-transition stress test on LBF referenced in Section 5.5. Independently at each step, every agent's action is replaced by a uniformly random action with probability $\varepsilon = 0.15$. Each agent selects an action using its learned policy, while the random replacement introduces uncertainty into action execution. This setting tests whether the learned coordination remains effective when agents cannot reliably carry out their intended actions. ValueGap keeps its ordering against the SMPE baseline and MAA2C, while all returns fall relative to the deterministic setting of Table 5.

## E Final Performance Across All Tasks

Table 5 reports the numerical final performance summarized by the learning curves of Section 5. Entries are IQM of episodic return with 95% bootstrap CIs. Bold marks the highest IQM per row; underlined entries have a 95% CI that overlaps that of the bold entry. HAPPO entries marked with † are run with the official implementation under default hyperparameters since the original paper [Kuba et al., 2022] only reported on SMAC v2; this is intended as a transparent reference rather than a tuned comparison. LBF task names follow the convention of Section 5.1.

For space, this table lists the four headline baselines only; MAPPO, ATM, EOI, COMA, MAVEN, MLIAM and EMC are plotted in Figures 3–5. EMC is below MASER on every task and never within CI of the highest IQM. Returns are unscaled environment returns: LBF and RWARE are collected-item rewards, MPE is minus the summed agent–landmark distance per step, SMAC v2 is win rate, and Hanabi-Small is a score out of 10.

## F Variance of the Shaped Policy-Gradient Estimator

This appendix states when the gap-aware shaping of Section 4.2 reduces the variance of the policy-gradient estimator. The result is the classical baseline condition [Greensmith et al., 2004], specialized to a baseline that depends on the state and on the agent's local information.

**Setup.** Fix an agent $i$ and a time $t$, and let $X := (s_t, h_t^i, z_t^i)$. Let $g := \nabla_{\psi_i} \log \pi^i_{\psi_i}(a_t^i \mid h_t^i, z_t^i)$ be the score, $A$ any advantage estimate with finite second moment (e.g., the TD error of Section 4.2), and $b = b(X) := \beta_\Delta(\hat{\Delta}_t^i)\,\hat{\Delta}_t^i$ the shaping term, which is a function of $X$ and not of $a_t^i$. The unshaped and shaped estimators are $\hat{\nabla}_0 := gA$ and $\hat{\nabla}_b := g\,(A - b)$. We assume $0 < \mathbb{E}[\|g\|^2 \mid X] < \infty$ and $\mathbb{E}[\|g\|^2 A^2 \mid X] < \infty$.

**Unbiasedness.** Given $X$, the action $a_t^i$ is drawn from $\pi^i_{\psi_i}(\cdot \mid h_t^i, z_t^i)$, so $\mathbb{E}[g \mid X] = \sum_a \nabla \pi^i_{\psi_i}(a \mid h_t^i, z_t^i) = 0$, and therefore

$$\mathbb{E}[\hat{\nabla}_b \mid X] \;=\; \mathbb{E}[gA \mid X] - b\,\mathbb{E}[g \mid X] \;=\; \mathbb{E}[\hat{\nabla}_0 \mid X]. \tag{12}$$

Table 5: Final performance on sixteen tasks (IQM, 95% bootstrap CIs). Bold: highest IQM per row; underline: overlapping CI; †: default HAPPO settings.

| Task | MAA2C | SMPE baseline | MASER | HAPPO | ValueGap (ours) |
|---|---|---|---|---|---|
| LBF $2s$-$8\times8$-$3p$-$2f$ | $0.85 \pm 0.05$ | $\underline{0.96 \pm 0.04}$ | $0.45 \pm 0.08$ | $0.81 \pm 0.06^{\dagger}$ | $\mathbf{0.97 \pm 0.04}$ |
| LBF $2s$-$9\times9$-$3p$-$2f$ | $0.46 \pm 0.10$ | $\underline{0.95 \pm 0.05}$ | $0.34 \pm 0.09$ | $0.58 \pm 0.09^{\dagger}$ | $\mathbf{0.97 \pm 0.04}$ |
| LBF $2s$-$12\times12$-$2p$-$2f$ | $0.62 \pm 0.10$ | $\underline{0.78 \pm 0.07}$ | $0.55 \pm 0.12$ | $0.66 \pm 0.09^{\dagger}$ | $\mathbf{0.83 \pm 0.06}$ |
| LBF $7s$-$20\times20$-$5p$-$3f$ | $0.42 \pm 0.08$ | $\underline{0.70 \pm 0.07}$ | $0.36 \pm 0.10$ | $0.45 \pm 0.09^{\dagger}$ | $\mathbf{0.76 \pm 0.06}$ |
| LBF $2s$-$11\times11$-$3p$-$2f$ | $0.08 \pm 0.04$ | $\underline{0.65 \pm 0.09}$ | $0.21 \pm 0.07$ | $0.20 \pm 0.07^{\dagger}$ | $\mathbf{0.78 \pm 0.07}$ |
| LBF $4s$-$11\times11$-$3p$-$2f$ | $0.12 \pm 0.05$ | $\underline{0.94 \pm 0.04}$ | $0.18 \pm 0.06$ | $0.34 \pm 0.07^{\dagger}$ | $\mathbf{0.96 \pm 0.04}$ |
| RWARE tiny-2ag-hard | $4.05 \pm 1.42$ | $\underline{10.50 \pm 1.51}$ | $2.10 \pm 0.85$ | $5.40 \pm 1.38^{\dagger}$ | $\mathbf{11.60 \pm 1.13}$ |
| RWARE tiny-4ag-hard | $11.50 \pm 1.98$ | $\underline{17.50 \pm 1.84}$ | $3.70 \pm 1.32$ | $13.20 \pm 1.71^{\dagger}$ | $\mathbf{19.40 \pm 1.47}$ |
| RWARE small-4ag-hard | $0.41 \pm 0.32$ | $\underline{4.50 \pm 0.81}$ | $0.66 \pm 0.27$ | $0.92 \pm 0.40^{\dagger}$ | $\mathbf{5.50 \pm 0.66}$ |
| MPE Spread-5 | $\underline{-247 \pm 10}$ | $\underline{-241 \pm 8}$ | $-268 \pm 14$ | $\underline{-244 \pm 9^{\dagger}}$ | $\mathbf{-236 \pm 7}$ |
| MPE Spread-8 | $\underline{-399 \pm 14}$ | $\underline{-393 \pm 11}$ | $-428 \pm 18$ | $\underline{-397 \pm 13^{\dagger}}$ | $\mathbf{-386 \pm 10}$ |
| MPE Double Spk-Lstn | $-32.10 \pm 1.83$ | $\underline{-29.05 \pm 1.50}$ | $-38.60 \pm 2.41$ | $-33.80 \pm 1.92^{\dagger}$ | $\mathbf{-28.00 \pm 1.34}$ |
| SMAC v2 protoss 5v5 | $0.36 \pm 0.05$ | $0.50 \pm 0.06$ | $0.31 \pm 0.07$ | $\underline{0.55 \pm 0.05}$ | $\mathbf{0.61 \pm 0.04}$ |
| SMAC v2 terran 5v5 | $0.41 \pm 0.04$ | $0.54 \pm 0.05$ | $0.38 \pm 0.06$ | $\underline{0.58 \pm 0.04}$ | $\mathbf{0.64 \pm 0.04}$ |
| SMAC v2 zerg 5v5 | $0.29 \pm 0.06$ | $0.40 \pm 0.07$ | $0.24 \pm 0.08$ | $\underline{0.45 \pm 0.06}$ | $\mathbf{0.56 \pm 0.05}$ |
| Hanabi-Small 2p | $6.42 \pm 0.44$ | $7.44 \pm 0.34$ | $5.72 \pm 0.55$ | $6.95 \pm 0.36^{\dagger}$ | $\mathbf{8.30 \pm 0.28}$ |

**Proposition 2** (Variance condition for the shaped estimator)**.** Define the variance-optimal baseline at $X$,

$$b^{\star}(X) \;:=\; \frac{\mathbb{E}[\|g\|^2 A \mid X]}{\mathbb{E}[\|g\|^2 \mid X]}.$$

Then, with $\mathrm{tr}\,\mathrm{Var}$ denoting the total variance,

$$\mathrm{tr}\,\mathrm{Var}(\hat{\nabla}_b \mid X) - \mathrm{tr}\,\mathrm{Var}(\hat{\nabla}_0 \mid X) \;=\; \mathbb{E}[\|g\|^2 \mid X]\; b\,\big(b - 2b^{\star}(X)\big).$$

Consequently, the shaping strictly reduces the conditional variance at $X$ if and only if $b(X)$ lies strictly between 0 and $2b^{\star}(X)$, and the reduction is largest at $b = b^{\star}(X)$. Since $b(X) \geq 0$ whenever $\hat{\Delta}^i_t \geq 0$, the shaping can reduce variance only at states with $b^{\star}(X) > 0$.

Proof. By Equation 12 the two estimators have the same conditional mean, so the difference of total variances equals the difference of conditional second moments: $\mathbb{E}[\|g\|^2(A-b)^2 \mid X] - \mathbb{E}[\|g\|^2 A^2 \mid X] = b^2\,\mathbb{E}[\|g\|^2 \mid X] - 2b\,\mathbb{E}[\|g\|^2 A \mid X]$, which equals the stated expression. The remaining claims follow from the sign of the quadratic $b \mapsto b(b - 2b^{\star})$. □

**Remarks.** (i) The condition is local: the shaping helps at states where $0 < \beta_{\Delta}(\hat{\Delta}^i)\hat{\Delta}^i < 2b^{\star}$ and hurts where it overshoots. The sigmoid schedule keeps the shift between $\frac{\beta_0}{2}\hat{\Delta}^i$ and $\beta_0\hat{\Delta}^i$ for $\hat{\Delta}^i \geq 0$, so overshooting is controlled by $\beta_0$; we set $\beta_0$ by held-out tuning (Appendix L), not from this condition. (ii) Decompose $A = A^{\mathrm{ctrl}} + A^{\mathrm{noise}}$, where $A^{\mathrm{ctrl}} := \mathbb{E}[A \mid X, a^i_t]$ and $A^{\mathrm{noise}}$ is the residual induced by teammates' actions and the environment, so that $\mathbb{E}[A^{\mathrm{noise}} \mid X, a^i_t] = 0$. Since $\|g\|^2$ is a function of $(X, a^i_t)$, $\mathbb{E}[\|g\|^2 A^{\mathrm{noise}} \mid X] = 0$, so $b^{\star}$ depends only on $A^{\mathrm{ctrl}}$. The contribution $\mathbb{E}[\|g\|^2 (A^{\mathrm{noise}})^2 \mid X]$ appears in both estimators and is not removed by any action-independent baseline. The correlation between $\hat{\Delta}^i$ and $\mathrm{Var}(A^{\mathrm{noise}} \mid \hat{\Delta}^i)$ reported in Section 5.3 characterizes where the pessimistic shaping acts. (iii) Proposition 2 is exact and makes no assumption on the gap. Its conditions apply to the conditional variance of the gradient estimator.

## G Full Statement and Proof of Proposition 1

**Assumption 1** (Stationarity)**.** Under a stationary joint policy $\pi$ (for instance an optimal one), the chain on the augmented state $(s_t, h^i_t, z^i_t)$, with the recurrent encoder state in place of the raw history, is stationary with distribution $\rho$; uniform ergodicity is a sufficient condition.

**Assumption 2** (Fixed critics and finite moments)**.** The critics $V_\xi$ and $V_\eta^i$ are fixed (e.g., converged) functions, so $\hat{\Delta}_t^i$ is a fixed measurable function of the augmented state, and $\mathbb{E}_\rho[(\hat{\Delta}_t^i)^2] < \infty$.

**Proposition 3** (Stationary magnitude of the gap-reduction bonus, full statement)**.** Under Assumptions 1–2, let $D_t := \hat{\Delta}_{t-1}^i - \hat{\Delta}_t^i$. Then

(a) $\mathbb{E}_\rho[D_t] = 0$;

(b) $\mathbb{E}_\rho[(D_t)_+] = \frac{1}{2}\,\mathbb{E}_\rho|D_t| \le \frac{1}{2}\sqrt{\mathrm{Var}_\rho(D_t)}$;

(c) $\mathbb{E}_\rho[(D_t)_+] = 0$ if and only if $\hat{\Delta}_{t-1}^i = \hat{\Delta}_t^i$ almost surely;

(d) if, in addition, $V_\eta^i$ is the idealized fixed point of Lemma 2 and $(h_t^i, z_t^i)$ determines $V_\xi(s_t)$ $\rho$-almost surely, then $\hat{\Delta}_t^i = 0$ almost surely and the bonus vanishes.

Proof. (a) By Assumption 2, $\hat{\Delta}_t^i$ is a fixed function of a stationary process, so $\hat{\Delta}_{t-1}^i$ and $\hat{\Delta}_t^i$ have the same marginal distribution, with finite mean, and $\mathbb{E}_\rho[D_t] = 0$. (b) Using $(x)_+ = \frac{1}{2}(|x| + x)$ and (a), $\mathbb{E}[(D_t)_+] = \frac{1}{2}\mathbb{E}|D_t|$; Jensen's inequality gives $\mathbb{E}|D_t| \le \sqrt{\mathbb{E}[D_t^2]} = \sqrt{\mathrm{Var}(D_t)}$. (c) $\mathbb{E}|D_t| = 0$ if and only if $D_t = 0$ almost surely. (d) If $Y := V_\xi(s_t)$ is almost surely constant given $(h_t^i, z_t^i)$, the minimizer in Lemma 2 is $m = Y$, so $\hat{\Delta}_t^i = Y - m = 0$ almost surely at every $t$, and (c) applies. □

**Relation to potential-based shaping.** Without the rectifier, the gap-reduction term $\alpha(\hat{\Delta}_{t-1}^i - \hat{\Delta}_t^i)$ is a potential difference with potential $\Phi_t = -\alpha\hat{\Delta}_t^i$. Along an undiscounted episode its sum telescopes to $\alpha(\hat{\Delta}_0^i - \hat{\Delta}_T^i)$, so it does not reward any particular path between given endpoints; this is the mechanism behind the policy-invariance results of Ng et al. [1999] and, for time-varying potentials, Devlin and Kudenko [2012]. With $\gamma < 1$ the correspondence is approximate, since our term omits the factor $\gamma$ on the next potential. The rectifier gives up exact telescoping in exchange for stability early in training, and Proposition 3 quantifies the price: at stationarity the rectified bonus has expectation $\frac{\alpha}{2}\mathbb{E}|D_t|$.

## H Assumptions, Scope of Theoretical Claims, and Reading Guide

**Assumptions.** We collect the assumptions used in the paper. (A1) A common team reward. (A2) Bounded rewards and $\gamma < 1$. (A3) The gap is defined for a fixed joint policy; it is a policy-evaluation quantity. (A4) Stationarity of the augmented chain and fixed critics, for Proposition 1 (Assumptions 1–2). (A5) Finite second moments of the score-weighted advantage, for Proposition 2. (A6) The state-conditioned critic $V^\pi(s_t)$ approximates the fully informed value $V^\pi(s_t, \mathbf{h}_t)$ (Section 3); this is exact when $s_t$ is Markov-sufficient for the continuation of the joint policy, for example in finite Dec-MDPs with jointly observable state and in transition-independent classes [Becker et al., 2004]. Under (A6), $\mathbb{E}[\Delta^{\pi,i} \mid h_t^i, z_t^i] = 0$ by the tower property, and Lemma 2 uses (A6) to equate the conditional means of the two critics.

**Theoretical results.** Lemma 2 characterizes what the regularized estimator targets at an idealized fixed point. Proposition 2 is the exact condition under which the shaping reduces the conditional variance of the gradient estimator. Section 5 and Appendix I examine this condition empirically. Proposition 3 bounds the stationary expectation of the rectified bonus by the fluctuation of the estimated gap; it shows that the bonus vanishes only when the gap is constant along stationary trajectories, for instance when local information resolves the state value. On dense-reward tasks such as MPE Spread-5 the gap is concentrated near zero, consistent with the smaller observed gain on those tasks (Table 5).

**The bonus during and after training.** Two empirical observations describe the bonus during training. First, the per-step Pearson correlation between $\hat{r}_t^i$ and $-\Delta\hat{\Delta}_t^i$ along trajectories of a trained ValueGap agent is $0.86 \pm 0.04$ (Section 5.3), suggesting that the bonus concentrates on transitions where the gap genuinely shrinks rather than on novelty per se. Second, the average bonus magnitude decreases monotonically across training on every task we tested, with median bonus magnitude at convergence $0.08 \pm 0.02$ relative to the median extrinsic reward magnitude on LBF $2s$-$11\times11$.

**Projection bias.** The local value head is trained by single-step temporal difference with input $(h_t^i, z_t^i)$. Since $V^{\pi,i}$ is in general not Markov in $h_t^i$, the TD fixed point is a projection of $V^{\pi,i}$ onto the function class indexed by $(h_t^i, z_t^i)$ rather than $V^{\pi,i}$ itself. The intrinsic reward therefore signals detectable reductions of the estimated gap, and the projection error shrinks as the encoder of $z^i$ becomes more informative about $s_t$.

**Reading the headline numbers.** The abstract reports a $15\%$ mean relative gain over the strongest baseline across the three SMAC v2 5v5 tasks ($10.9\%$, $10.3\%$ and $24.4\%$ on protoss, terran and zerg). This is a mean over three data points and is sensitive to which baseline is strongest on each task; here it is HAPPO on all three SMAC tasks, which we run with the released configuration and which is therefore not subject to our † caveat. The corresponding aggregate gains on LBF, RWARE, MPE, and Hanabi are $6.7\%$, $14.5\%$, $2.5\%$, and $11.6\%$. These benchmark-level gains over the strongest baseline differ from the gains of full ValueGap over individual ablation rows in Table 1, which are relative to a single variant on a single task.

**Limitations of the baseline set.** Our baseline set was selected to span representation learning (LIAM/SMPE), value decomposition (MAVEN, indirectly through QMIX-style reductions), credit assignment (COMA, MAA2C, MAPPO, HAPPO), and exploration (EMC, MASER, EOI). It does not include the most recent transformer-based opponent models, fully sequence-model belief encoders, or methods that combine multiple intrinsic motivation signals (such as combinations of curiosity-driven and information-theoretic bonuses). On Hanabi specifically, our experiments are restricted to Hanabi-Small 2-player; the SAD/OBL family of methods is the de facto standard on Hanabi-Full and was not included. The results establish improvements over the evaluated baseline set on these tasks.

# I Controlled Tabular Variance Experiment

This appendix specifies a controlled tabular experiment that checks the condition of Proposition 2 on a transparent example in which every quantity can be computed exactly.

**Environment.** A Dec-POMDP with $|S| = 256$ joint states, the pair of agent positions on a $4 \times 4$ grid, $N = 2$ agents, joint action space $|A_1| = |A_2| = 4$ (one step in each direction), deterministic transitions, and team reward $r(s, a) = +1$ when both agents are simultaneously on a designated goal cell (top-right), $-0.05$ otherwise. The horizon is $H = 30$ steps and $\gamma = 0.95$. Agent 1's observation function is the identity (full observability of the joint state), so its gap is identically zero. Agent 2's observation function is a coarse aggregation: it observes only the row index of its own position, with the column and the partner's position hidden, so its observation takes four values. The information disparity is therefore exactly localizable to agent 2.

**Exact quantities.** Both critics are tabular and computed by exact policy evaluation under a fixed policy $\pi^{\text{eval}}$ (uniform over joint actions). The local critic is the fixed point of Lemma 2 on agent 2's observation, so $\hat{\Delta}^2$ is computed exactly. For a softmax policy, the score $g$ and the variance-optimal baseline $b^\star(X)$ of Proposition 2 are computed in closed form at every $X$.

**Estimator comparison.** We sweep $\beta_0 \in \{0, 0.1, \ldots, 2.0\}$ and, for each value, compute the total variance of the shaped and unshaped estimators $g(A - b)$ and $gA$ at every $X$, averaged over the on-policy state distribution, together with the fraction of states at which $0 < b(X) < 2b^\star(X)$.

**Results.** The average total variance of the shaped estimator is below that of the unshaped one for every $\beta_0 \in (0, 1.3]$, with a maximum relative reduction of $21\%$ at $\beta_0 = 0.6$ and $19\%$ at the recommended $\beta_0 = 0.5$; for $\beta_0 \geq 1.5$ the average variance exceeds the unshaped one, because the shift overshoots $2b^\star$ on enough states to dominate the gain at the remaining ones. At the recommended $\beta_0 = 0.5$ the condition $0 < b(X) < 2b^\star(X)$ holds at $74\%$ of the states visited under the on-policy distribution, and at $58\%$ of the states in the top quintile of $\hat{\Delta}^2$: overshoot is concentrated exactly where the shift is largest, consistent with remark (i) above. Two features of the example are worth recording. First, the reduction is bounded by the deterministic part of the advantage, since $b^\star$ depends on $A^{\text{ctrl}}$ only (Appendix F); here $\text{Var}(A^{\text{noise}})$ accounts for $62\%$ of the total gradient variance and is untouched by any action-independent baseline, which is why the achievable reduction is $21\%$ rather than far larger. Second, the fraction of states satisfying the condition is not monotone in $\beta_0$:

it rises to $0.81$ at $\beta_0 = 0.2$, where $b$ is positive but small, and falls to $0.39$ at $\beta_0 = 2.0$, while the total variance is minimized in between. The two criteria therefore disagree, and the variance criterion is the one that matters. The full data and reproduction script will be included in the code release (`tabular_variance/`).

# J Hyperparameters

Table 6 reports the ValueGap-specific hyperparameters. Per-benchmark settings for the actor, the centralized critic, the local value head, and the encoder are reported in Tables 7 and 8. All values reported here are the ones used to produce Table 5; values found by held-out tuning (Appendix L) are marked with a $^*$.

Table 6: ValueGap-specific hyperparameters. $\beta_0, \tau, \alpha, \lambda_{\text{reg}}$ are set once per benchmark on its held-out tuning task (Appendix L). The held-out sweep returned the same point on all five benchmarks, so a single setting is reported; performance is flat within an order of magnitude of these values (Figure 6, panel ii).

| Hyperparameter | Value | Description |
|---|---|---|
| $\beta_0^*$ | $0.5$ | Maximum shaping magnitude (sigmoid amplitude). |
| $\tau^*$ | $1.0$ | Sigmoid temperature in $\beta_\Delta(x) = \beta_0\sigma(x/\tau)$. |
| $\alpha^*$ | $0.2$ | Coefficient of the gap-reduction intrinsic term. |
| $\alpha'_t$ | $0.1 \cdot e^{-t/T_{\text{decay}}}$ | Decaying coefficient of the SimHash residual. |
| $T_{\text{decay}}$ | $5 \times 10^5$ | Decay constant for $\alpha'_t$. |
| $\lambda^*_{\text{reg}}$ | $0.5$ | Strength of the one-sided regularizer. |
| SimHash dim. | $64$ | Output dimension of $\text{SH}(\cdot)$. |
| $V^i$ target update | $\rho_{\text{tgt}} = 0.005$ | Polyak averaging rate of $\eta'$ (distinct from the critic rate $\tau_{\text{tgt}}$). |

Table 7: Shared training hyperparameters per benchmark. Values for actor and centralized critic are matched to the SMPE baseline so that any difference in performance between ValueGap and SMPE is attributable to the gap-aware modules.

| | LBF | RWARE | MPE | SMAC v2 | Hanabi |
|---|---|---|---|---|---|
| Optimizer | Adam | Adam | Adam | Adam | Adam |
| Learning rate$^*$ | $3\times10^{-4}$ | $5\times10^{-4}$ | $7\times10^{-4}$ | $5\times10^{-4}$ | $1\times10^{-3}$ |
| Entropy coeff.$^*$ | 0.01 | 0.02 | 0.005 | 0.01 | 0.02 |
| Batch size | 32 ep. | 32 ep. | 32 ep. | 32 ep. | 32 ep. |
| Discount $\gamma$ | 0.99 | 0.99 | 0.95 | 0.99 | 0.999 |
| GAE $\lambda$ | 0.95 | 0.95 | 0.95 | 0.95 | 0.95 |
| Critic target update | $\tau_{\text{tgt}} = 0.005$ | $\tau_{\text{tgt}} = 0.005$ | $\tau_{\text{tgt}} = 0.005$ | $\tau_{\text{tgt}} = 0.01$ | $\tau_{\text{tgt}} = 0.005$ |
| Training steps | 10M | 40M | 10M | 10M | 20M |
| Seeds | 8 | 5 | 8 | 5 | 5 |

Table 8: Encoder/decoder hyperparameters used to learn $z^i$. The same encoder is used by SMPE, MLIAM, and ValueGap on each benchmark for fair comparison.

| | LBF | RWARE | MPE | SMAC v2 | Hanabi |
|---|---|---|---|---|---|
| Latent dim. $\lvert z^i \rvert$ | 32 | 48 | 32 | 64 | 48 |
| KL coefficient | 0.1 | 0.1 | 0.05 | 0.1 | 0.05 |
| Reconstruction loss | MSE | MSE | MSE | MSE | CE |
| Encoder steps/iter | 4 | 4 | 4 | 8 | 4 |

# K Architectures

We use the same backbone architecture across all benchmarks, varying only the input/output sizes to match the observation and action spaces.

**Actor.** Two-layer GRU with hidden size 128 followed by a feed-forward head of size $128 \to |A_i|$ with ReLU. Input is the concatenation of the local observation $o_t^i$, the previous action $a_{t-1}^i$ encoded as a one-hot, and the latent $z_t^i$.

**Centralized critic $V_\xi$.** Three-layer MLP with hidden sizes $[256, 256, 128]$ and ReLU, applied to the joint state $s_t$ (concatenated agent positions, observations, and time index where applicable).

**Local value head $V_\eta^i$.** Three-layer MLP with hidden sizes $[128, 128, 64]$ and ReLU, taking as input the actor's GRU hidden state at time $t$ concatenated with $z_t^i$. We chose the actor's hidden state as input rather than $h_t^i$ directly because the recurrent state already encodes the relevant history; we ablate this choice in our codebase and find that using $h_t^i$ with a fresh recurrent encoder degrades performance by $\sim 4\%$ at substantially higher compute (the corresponding script, `local_head_input/`, will be part of the code release).

**Variational encoder/decoder of $z^i$.** Encoder: two-layer MLP $[128, 128]$ to a Gaussian posterior over $z^i$; decoder: two-layer MLP $[128, 128]$ producing reconstruction targets matching the observation modalities described in Table 8.

## L Held-Out Tuning Protocol

We follow a uniform tuning protocol across all methods. For each benchmark we designate one task as the tuning task: LBF $2s$-$8 \times 8$-$3p$-$2f$, RWARE tiny-2ag-hard, MPE Spread-5, SMAC v2 protoss 5v5, Hanabi-Small 2p. On the tuning task we perform a coarse grid sweep over the learning rate $\{1, 3, 5, 7, 10\} \times 10^{-4}$ and the entropy coefficient $\{0.005, 0.01, 0.02\}$ for every method, including ValueGap and every baseline that admits these knobs. Each grid point uses three seeds with the same evaluation budget. We pick the configuration with the highest IQM final return on the tuning task and apply it unchanged to the remaining tasks of the same benchmark. The ValueGap-specific hyperparameters ($\beta_0, \tau, \alpha, \lambda_{\text{reg}}$) are likewise swept once per benchmark on the same tuning task, over $\{0.1, 0.3, 1, 3, 10\}$ times the value reported in Table 6 (i.e. $\beta_0 \in \{0.05, \ldots, 5\}$, $\tau \in \{0.1, \ldots, 10\}$, $\alpha \in \{0.02, \ldots, 2\}$ and $\lambda_{\text{reg}} \in \{0.05, \ldots, 5\}$) and held fixed across that benchmark's remaining tasks; the selected point was the $1\times$ value on every benchmark, which is why Table 6 reports a single setting. Neither ValueGap nor any baseline is tuned per task. Figure 6(ii) sweeps the same range with the full seed budget on LBF $2s$-$11 \times 11$-$3p$-$2f$.

## M Baseline Implementations

We list the source of each baseline implementation and the modifications, if any, applied to obtain the numbers in Table 5. Implementation versions and commit prefixes are documented in Appendix Q, and the configurations will be included in the code release.

**MAA2C, MAPPO, COMA, SEAC.** Implementations from Papoudakis et al. [2021b]. Used as released; the held-out tuning of Appendix L sets the learning rate and entropy coefficient.

**LIAM, SMPE.** Implementations from Papoudakis et al. [2021a] and Kontogiannis et al. [2025] respectively. The encoder of $z^i$ is shared with ValueGap (same architecture and same KL coefficient) so that the comparison isolates the gap-aware modules.

**EOI, EMC, MASER, MAVEN.** These methods were run with the original authors' implementations on their native value-decomposition backbones, unchanged; nothing was re-derived or re-implemented on our actor-critic backbone. Only the environment wrappers and the evaluation protocol (step budget, seeds, IQM aggregation) were standardized. On benchmarks not covered by the original papers we apply the held-out tuning of Appendix L.

**HAPPO.** Official implementation from Kuba et al. [2022]. On SMAC v2 we apply the configuration released with the original paper, unchanged, to all three 5v5 scenarios. On benchmarks for which the release provides no configuration (LBF, RWARE, MPE, Hanabi), we run with the released

default hyperparameters and mark these entries with † in Table 5. The † marker identifies these default-configuration comparisons.

**ATM, MLIAM.** Reproduced from the open-source releases of Yang et al. [2022] (ATM) and of Papoudakis et al. [2021a] (MLIAM, the multi-agent variant of LIAM); the tuning protocol matches Appendix L.

**Why some methods underperform on some tasks.** The numbers reported for EMC and MASER are noticeably below their reports on the tasks the original papers cover, but consistent with their reports on the tasks they did not cover. Our held-out tuning does not include hyperparameters internal to those methods (e.g., the curiosity learning rate of EMC), which we left at the original defaults. Because these methods keep their value-decomposition backbones while ValueGap runs on MAA2C, part of any cross-family difference is attributable to the backbone rather than to the method; the comparisons our claims rely on are therefore the within-family ones (MAA2C, MAPPO, COMA, ATM, MLIAM, SMPE, and HAPPO).

## N Statistical Protocol

All aggregate numbers in Table 5 and Table 1 are interquartile means (IQM) over the seeds listed in Table 7. Confidence intervals are computed via stratified bootstrap with $50{,}000$ resamples, following the guidance of Agarwal et al. [2021]; the implementation uses the `rliable` package. The factors of variability captured by the bars are: (a) random initialization of all neural network parameters, (b) environment seeding (initial state and stochastic transition randomness), and (c) replay sampling order for the encoder updates. Train/test splits are not relevant in our setting (online RL); all reported numbers are computed on freshly sampled evaluation episodes that are not used for training.

For the summary table we use a conservative marginal criterion: an entry is underlined when its $95\%$ CI overlaps that of the highest-IQM entry in its row. For ablations and controls, which share seeds with the full method, we report performance differences in Table 4 and summarize the paired confidence-interval results in Appendix D. The analysis uses stratified bootstrap percentile intervals over per-seed differences and the probability of improvement of Agarwal et al. [2021].

**Mechanism-analysis correlations.** The correlations reported in Section 5.3 are computed as follows. All of them are reported as the mean over seeds of a per-seed statistic, with the standard deviation across seeds as the stated error: $8$ seeds for the two LBF statistics and $5$ seeds for the SMAC v2 one. The Pearson correlation between $\hat{r}_t^i$ and $-\Delta\hat{\Delta}_t^i$ is computed per trajectory and then averaged over the evaluation trajectories of a seed. The Spearman correlation between visit frequency and mean $\hat{\Delta}^i$ uses the grid cells visited at least once by at least one seed. The Spearman correlation between $\hat{\Delta}^i$ and $\mathrm{Var}(A^{\mathrm{noise}} \mid \hat{\Delta}^i)$ is computed by binning $\hat{\Delta}^i$ into 20 quantile bins, regressing $A_t$ on the agent's own action $a_t^i$ within each bin, and taking the residual variance as the empirical estimate of $\mathrm{Var}(A^{\mathrm{noise}} \mid \hat{\Delta}^i)$.

## O Compute Resources

All experiments were run on an internal cluster of NVIDIA RTX A6000 ($48$ GB) and V100 ($32$ GB) GPUs. CPU and memory allocations per run were $8$ CPU cores and $32$ GB of RAM. Per-run wall-clock times and GPU-hour accounting are summarized in Table 9.

Wall-clock time is per single run of ValueGap on the listed task; the SMPE baseline needs approximately $79\%$ of that time, i.e. ValueGap carries the $26\%$ overhead reported in Section 5.6. For aggregate rows, Steps is the sum over runs and Wall-clock is the mean per run. The Seeds column counts single runs: $11$ baselines $\times$ $107$ and $6$ ablation variants $\times$ $23$ (the four ablations plus the two controls, which are run on all four ablation tasks).

The total compute consumed by the project, including preliminary experiments and failed runs that do not appear in the paper, exceeds the reported total by approximately $35\%$, for an estimated full-project cost of about $17{,}300$ GPU-hours.

Table 9: Compute cost. Wall-clock: hours per run; Seeds: run count. Aggregate rows report total steps and mean per-run time.

| Benchmark | Steps | Wall-clock (h/run) | Seeds | GPU-hours (total) |
|---|---|---|---|---|
| LBF (6 tasks) | 10M each | $\sim 6$ | $8 \times 6 = 48$ | $\sim 290$ |
| RWARE (3 tasks) | 40M each | $\sim 22$ | $5 \times 3 = 15$ | $\sim 330$ |
| MPE (3 tasks) | 10M each | $\sim 5$ | $8 \times 3 = 24$ | $\sim 120$ |
| SMAC v2 (3 tasks) | 10M each | $\sim 24$ | $5 \times 3 = 15$ | $\sim 360$ |
| Hanabi-Small | 20M | $\sim 14$ | 5 | $\sim 70$ |
| **ValueGap subtotal** | 1.57B | 10.9 | 107 | $\sim 1{,}170$ |
| All baselines (11 methods) | 17.3B | 8.1 | 1,177 | $\sim 9{,}500$ |
| Ablations and analyses | 2.58B | 15.2 | 138 | $\sim 2{,}100$ |
| **Reported total** | 21.4B | 9.0 | 1,422 | $\sim 12{,}800$ |

# P Negative Results

We report two settings in which ValueGap did not outperform the SMPE baseline, to inform the reader of the boundary of its applicability.

**Dense-reward symmetric MPE.** On MPE Spread-3 (three agents, three landmarks, fully symmetric observability, dense per-step distance reward), the gap distribution is concentrated near zero throughout training: every agent has comparable information about the relevant subset of the joint state. ValueGap matches the SMPE baseline within confidence intervals ($-118 \pm 3$ vs $-120 \pm 4$ on 8 seeds), but neither shaping nor the intrinsic bonus produces a measurable benefit. Both numbers are on the scale that this task admits: the return is minus the summed agent–landmark distance per step over a 25-step episode, so the worst case (every agent at the maximum distance $2\sqrt{2}$ throughout) is $-N \cdot 2\sqrt{2} \cdot 25 \approx -212$ for $N = 3$, and the published MAPPO returns on this task range from $-149.3$ to $-128.6$ [Papoudakis et al., 2021b]. This is what the mechanism predicts: when $\hat{\Delta}^i$ is uniformly small, both the shaping and the gap-reduction bonus are close to zero.

**Fully observable Bridge crossing.** On a synthetic fully observable two-agent bridge-crossing task we constructed to validate that ValueGap does not hurt in fully observable settings, the gap is identically zero (both critics estimate the same quantity) and the shaping reduces to subtracting zero. Performance matches MAA2C exactly within seed variance ($0.99 \pm 0.01$ for both, 5 seeds, 5M steps). The intrinsic bonus also vanishes in this setting because $\hat{\Delta}^i_{t-1} - \hat{\Delta}^i_t = 0$ identically, as Proposition 3(d) predicts.

# Q Asset Provenance and Licenses

The external assets used in this work are released under the MIT or Apache-2.0 license. MIT-licensed assets are LBF [Albrecht and Stone, 2017], RWARE [Papoudakis et al., 2021b], MPE [Lowe et al., 2017], SMAC v2 [Ellis et al., 2023], LIAM and MLIAM [Papoudakis et al., 2021a], ATM [Yang et al., 2022], EOI [Jiang and Lu, 2021], EMC [Zheng et al., 2021], MASER [Jeon et al., 2022], and HAPPO [Kuba et al., 2022]. Apache-2.0-licensed assets are the Hanabi Learning Environment [Bard et al., 2020], EPyMARL [Papoudakis et al., 2021b], SMPE [Kontogiannis et al., 2025], MAVEN [Mahajan et al., 2019], and rliable [Agarwal et al., 2021]. We respect the terms of each license.

The versions used are LBF v1.4 (`a3c19f0`), RWARE v1.2 (`d2eb1bc`), PettingZoo MPE v1.21.0, SMAC v2 v1.0 (`7e91c2a`), Hanabi LE v0.1 (`52f76e2`), EPyMARL v1.0 (`f1d4e9e`), and rliable v1.1.0. Baseline implementation commits are LIAM and MLIAM (`c9a8b21`), ATM (`7b2e0f4`), SMPE (`4d6a31c`), EOI (`e1d3a8a`), EMC (`3a2c64f`), MASER (`8b91d22`), MAVEN (`2e1f8b5`), and HAPPO (`9c712eb`). Commit hashes are given as prefixes; full hashes will be included in the code release.

## R Broader Impacts

**Positive impacts.** The applications motivating this paper (cooperative robotics, mixed-autonomy traffic, and search-and-rescue) are settings in which agents act on partial on-board sensing and where robust cooperation under information asymmetry is the dominant practical challenge. By making per-agent information disadvantage into a measurable, optimizable quantity, ValueGap can in principle improve sample efficiency and reliability of cooperative MARL systems deployed under these constraints, reducing the time and energy cost of training.

**Negative impacts.** Like any cooperative MARL improvement, ValueGap is dual-use: the same techniques that improve coordination among rescue robots can in principle improve coordination among adversarial agents in surveillance or military settings. Two specific risks are worth flagging.

First, the information-seeking intrinsic reward incentivizes agents to move toward states that reduce their information disadvantage. In environments where the observation function is itself controlled by a designer (e.g., a robot with steerable cameras), this incentive could in principle encourage intrusive sensing of regions that the designer did not intend agents to inspect. Operationally, the obvious mitigation is to constrain the observation function at the design stage rather than relying on the algorithm to abstain.

Second, our analysis assumes a cooperative setting (a single team reward). In adversarial or mixed cooperative-competitive settings, the gap signal can in principle be inverted: an adversary could shape its own advantage with the negative of the gap to amplify reliance on noisy, partially observed states. Such behavior would be undesirable. We have not investigated whether ValueGap extends to such settings, and we recommend that downstream users restrict its application to fully cooperative settings until the adversarial case is understood.

We do not see a direct path from this research to deepfake-style harms, mass disinformation, or fairness-relevant decision-making, so we do not discuss those categories further.

## S Extended Related Work

**Uncertainty-weighted and pessimistic updates.** Several methods modulate updates by an uncertainty estimate: UWAC [Wu et al., 2021] and SUNRISE [Lee et al., 2021] weight Bellman or policy updates by ensemble or dropout uncertainty, and pessimistic offline methods such as CQL [Kumar et al., 2020] penalize values outside the data distribution. Our shaping follows the same pattern with a cheaper proxy: the value that local information leaves unresolved, which requires no ensemble, and it enters as an additive action-independent shift rather than as a weight.

**Potential-based shaping.** Ng et al. [1999] show that potential-based shaping preserves optimal policies; Devlin and Kudenko [2012] extend this to time-varying potentials, and Harutyunyan et al. [2015] show how arbitrary rewards can be expressed as potential-based advice. Our gap-reduction term is a rectified instance with potential $-\alpha\hat{\Delta}^i$ (Appendix G); the rectifier trades exact invariance for stability early in training.

**Exploration bonuses.** Beyond the methods discussed in Section 2, count and novelty bonuses such as random network distillation [Burda et al., 2019] and cooperative multi-agent exploration [Liu et al., 2021] reward state coverage. Our residual count term belongs to this family but is decayed to zero, and the gap-reduction term rewards value-relevant information about the hidden state rather than coverage.

**Privileged information and decentralized critics.** Asymmetric learning with privileged information has been studied in POMDPs [Warrington et al., 2021, Baisero et al., 2022] and in learning-by-cheating approaches to imitation [Chen et al., 2019]. Networked actor-critic methods [Zhang et al., 2018] train decentralized local critics without a centralized one. ValueGap combines a centralized and a local critic and uses their discrepancy as a learning signal.

## NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract makes three claims that map onto Sections 3–4 (definition of the per-agent value gap and the ValueGap algorithm), Appendices A, F and G (what the regularized estimator measures, the variance condition for the shaped estimator, and the stationary bound on the bonus), and Section 5 (sixteen tasks across five benchmarks). The empirical claim (highest IQM on all sixteen tasks) is restricted to the baseline set of Section 5.1, and Section 5.2 states on which tasks the confidence intervals overlap the runner-up's. The theoretical claims are stated with their conditions: the variance statement gives the classical baseline condition, and the bonus is bounded at stationarity and vanishes only when local information resolves the state value. Appendix H delineates what follows from the theory and what is empirical.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 6 discusses the dependence of $\hat{\Delta}^i$ on the centralized critic's noise, the absence of experiments combining ValueGap with sequence-model belief encoders, and the conditional nature of the variance analysis. Appendix H additionally discusses the projection bias of the local value head, the finite-time behavior of the bonus relative to Proposition 1, how to read the SMAC v2 headline number (a mean over three tasks), and limitations of the baseline set; it also collects all assumptions (A1)–(A6). Appendix P reports two settings where ValueGap did not help.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: Definitions 1 and 2 state the underlying objects. Lemma 2 is stated and proved in Appendix A. Proposition 2 (variance condition for the shaped estimator) is stated and proved in Appendix F. Proposition 1 is restated with its assumptions and proved in Appendix G, with a proof sketch in Appendix B. All assumptions are collected as (A1)–(A6) in Appendix H, which also delineates which conclusions follow from the theory and which are empirical.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Section 5.1 specifies the benchmark suite, seed counts, training horizons, and evaluation protocol (IQM with $95\%$ stratified bootstrap CIs following Agarwal et al. [2021]). Algorithm 2 (Appendix C) lists the full training loop. Appendix J provides the hyperparameters of ValueGap and the shared training settings used to produce Table 5, including learning rates, entropy coefficients, $\beta_0, \tau, \alpha, \alpha'_t, \lambda_{\text{reg}}$, SimHash dimensions, and target-network update rates. Appendix K gives architectural specifications (layer widths, activation choices, recurrent state dimensions). Appendix M clarifies the source of each baseline implementation and its tuning protocol. Appendix I fully specifies the controlled tabular experiment so that the variance argument can be reproduced from scratch in tens of seconds.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: The code release is in preparation and will be made publicly available soon after publication under the MIT license, with launch scripts for every figure and table, fixed seed lists, and configuration files. Until then, Appendices J–L give every hyperparameter, the tuning protocol, and the architecture, which is sufficient to reimplement ValueGap and to reproduce the reported numbers. All five benchmarks (LBF, RWARE, MPE, SMAC v2, Hanabi) are open source; their versions and applicable commit prefixes are documented in Appendix Q.

   Guidelines:

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

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Section 5.1 describes the global protocol; Appendix J reports per-benchmark hyperparameter tables (Adam optimizer with the listed learning rates, batch sizes, entropy coefficients, $\gamma$, GAE $\lambda$, target-update rates, training steps and seeds). Appendix K reports the architectures of the actor, the centralized critic $V_\xi$, the local value head $V_\eta^i$, and the variational encoder/decoder of $z^i$. Appendix L describes the held-out tuning procedure used identically for ValueGap and the tuned baselines.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Following the recommendations of Agarwal et al. [2021], all aggregate numbers in Table 5 and Table 1 are reported as interquartile means (IQM) over $5$–$8$ seeds with $95\%$ stratified bootstrap confidence intervals computed with $50{,}000$ bootstrap resamples; the procedure is described in Appendix N. Learning curves in Figures 3–5 report shaded $25$–$75$ percentile bands. Mechanism-analysis correlations in Section 5.3 are reported as mean $\pm$ standard deviation across seeds. The factors of variability captured by the bars (random initialization, environment seeding, and replay sampling order) are stated explicitly in Appendix N, which also describes the paired comparisons used for ablations and controls. Table 4 presents the performance differences, and Appendix D summarizes the paired confidence-interval results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Appendix O lists per-benchmark wall-clock time, GPU type (NVIDIA RTX A6000 / V100), CPU/RAM allocation, and total estimated compute (approximately $12{,}800$ GPU-hours for the full set of reported runs, plus an estimated $\sim 35\%$ overhead for preliminary and failed runs that do not appear in the paper). Section 5.6 additionally compares the wall-clock overhead of ValueGap to the SMPE baseline and to other intrinsic-motivation methods.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work uses only public, simulation-based cooperative MARL benchmarks. There are no human subjects, no scraped data, no biometric or personally identifying information, and no deployment to real-world decision-making in the experiments reported. We have reviewed the NeurIPS Code of Ethics and find no aspect of the methodology in conflict with it.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Appendix R discusses both directions. Positive impacts include improved sample efficiency in cooperative robotic systems, traffic coordination, and search-and-rescue scenarios where on-board sensing is unavoidable. Negative impacts include the dual-use nature of any cooperative MARL improvement (e.g., automated multi-agent surveillance) and the risk that asymmetric-information agents trained with information-seeking bonuses learn intrusive sensing behaviors when the underlying observation function is itself a design choice; we describe the operational mitigations relevant to deployment.

    Guidelines:

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

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The artifact to be released is a research codebase that trains task-specific cooperative-MARL policies on public simulated benchmarks. It is not a pretrained generative model, foundation model, or scraped dataset, and we are not aware of a high-risk misuse pathway from a research-grade MARL training framework that would require safeguards beyond the standard open-source license terms.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: Every benchmark and baseline used is cited in Section 2 and Section 5.1. Appendix Q documents the external environments, baseline implementations, and evaluation tools, together with their MIT or Apache-2.0 licenses, versions, and applicable commit prefixes. We respect the terms of each license.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: The new asset is the ValueGap reference implementation, whose release is in preparation and will be public soon after publication. It will ship with a README documenting installation, dependencies, the entry-point script for each table/figure, the random seed lists, and the expected output format. Configuration files in YAML expose every hyperparameter named in Appendix J. The code will be released under MIT.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects. All experiments use simulated cooperative-MARL benchmarks.

    Guidelines:

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

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

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

    Answer: [N/A]

Justification: No human subjects were involved in this research, so no IRB review or equivalent was required.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: The core methodology of this paper is a cooperative-MARL algorithm and its analysis. No LLM is used as a component of the proposed algorithm, the experimental pipeline, the data collection, or the analysis. LLMs were used only for incidental writing-assistance tasks (proofreading, sentence-level editing) that do not affect scientific rigor or originality, which under the NeurIPS LLM policy does not require declaration.

Guidelines:

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