Fair Cooperation in Mixed-Motive Games via Conflict-Aware Gradient Adjustment

Woojun Kim
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
woojunk@andrew.cmu.edu

Katia Sycara
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
sycara@andrew.cmu.edu

Abstract

Multi-agent reinforcement learning in mixed-motive settings presents a fundamental challenge: agents must balance individual interests with collective goals, which are neither fully aligned nor strictly opposed. To address this, reward restructuring methods such as gifting and intrinsic motivation have been proposed. However, these approaches primarily focus on promoting cooperation by managing the trade-off between individual and collective returns, without explicitly addressing fairness with respect to agents' task-specific rewards. In this paper, we propose an adaptive conflict-aware gradient adjustment method that promotes cooperation while ensuring fairness in individual rewards. The proposed method dynamically balances policy gradients derived from individual and collective objectives in situations where the two objectives are in conflict. By explicitly resolving such conflicts, our method improves collective performance while preserving fairness across agents. We provide theoretical results that guarantee monotonic non-decreasing improvement in both the collective and individual objectives and ensure fairness. Empirical results in sequential social dilemma environments demonstrate that our approach outperforms baselines in terms of social welfare, while maintaining fairness.

1 Introduction

Multi-agent reinforcement learning (MARL) aims to train multiple agents to maximize cumulative rewards in a given task. Depending on the reward structure, MARL is typically categorized into three settings: cooperative, adversarial, and mixed-motive. In the mixed-motive setting, agents' rewards are neither fully aligned (as in cooperative settings) nor entirely opposed (as in adversarial settings), necessitating that each agent balances self-interest with the collective interest. This mixed-motive setting is frequently encountered in real-world applications. For example, in traffic control systems, each agent (e.g., a local intersection controller) may aim to minimize local congestion, which can conflict with global traffic flow optimization if not coordinated. A similar tension happens in sequential social dilemmas (SSDs) such as Cleanup or Harvest [18], where agents must invest in public goods (e.g., cleaning waste or harvesting resources judiciously) that benefit the group but do not yield immediate individual rewards.

However, achieving such a balance in mixed-motive settings is inherently challenging. Excessively selfish behavior by agents can deteriorate collective welfare, which, in turn, negatively impacts each agent's own return—creating a vicious cycle that ultimately harms all participants. Additionally, in some scenarios, certain agents must sacrifice their own returns to improve the collective outcome, potentially leading to unfairness. Conversely, an excessive focus on fairness can hinder learning in tasks that require cooperation. Therefore, it is crucial to enhance collective outcome while ensuring fairness by appropriately balancing individual and collective interests.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

In mixed-motive settings, many approaches adopt reward restructuring by incorporating intrinsic rewards such as social influence [12], formal contracts [9], gifting [23, 17], and inequity aversion [10]. These methods primarily aim to maximize the collective return by mediating the trade-off between self-interest and collective interests. For example, gifting mechanisms promote cooperation by enabling agents to share a portion of their rewards with others. However, despite their effectiveness in inducing cooperation, such reward restructuring may raise fairness concerns, for example, the gifted reward is intrinsic and not part of the task-defined reward that agents are fundamentally trained to maximize. Consider the Cleanup environment: agents only receive extrinsic rewards for collecting apples, yet apples will only regrow if waste is cleaned. It is often observed that some agents specialize in cleaning waste while others collect apples and subsequently gift a portion of their reward to those who sacrificed their own gain. Although this leads to improved collective performance, the agents engaged in waste cleaning never directly receive task rewards from apple collection. This becomes even worse if the agents are trained with the collective return, since some agents are encouraged to clean the waste all the time. Aside from reward restructuring, an approach has been proposed to align individual and collective objectives by adjusting policy gradients toward stable fixed points of the collective return, while still considering individual interests [20]. However, this method does not adequately consider fairness, as it primarily focuses on stability without explicitly addressing the conflict between individual and collective objectives.

In order to enhance cooperation while ensuring fairness, we propose a fair and conflict-aware gradient adjustment method (FCGrad) that dynamically balances gradients derived from individual and collective objectives by explicitly handling conflicts between them. FCGrad first detects the presence of conflicts, and when conflicts are found, it projects one gradient onto the normal plane of the other—preserving one objective's direction while avoiding interference with the other. Notably, FCGrad prioritizes the gradient associated with the lower objective value. For example, if the individual objective is lower than the collective objective, indicating that the agent is in an unfair situation, we project the individual gradient onto the normal plane of the collective gradient and use the result as the final update. This enables cooperation to be enhanced while maintaining fairness by resolving conflicts. We provide theoretical results showing that, under certain assumptions, the proposed gradient method guarantees monotonic non-decreasing improvement in both collective and individual objectives. We further show that the two objectives converge to the same value, leading to all agents' objectives aligning—thus ensuring individual fairness. In addition, we empirically demonstrate the effectiveness of FCGrad in terms of α-fairness [25], which captures both performance and fairness, in the Unfair Coin Game and two sequential social dilemma environments: Cleanup and Harvest.

2 Background and Related Works

2.1 Partially Observable Stochastic Game

A *Partially Observable Markov Game* (POMG) models multi-agent decision-making under uncertainty [21, 4]. A POMG is defined as a tuple $(N, S, \{A_i\}_{i=1}^N, T, \{O_i\}_{i=1}^N, \{R_i\}_{i=1}^N, \gamma)$, where N is the number of agents, S is the set of states, A_i is the action set of agent i, $T : S \times A_1 \times \cdots \times A_N \to \Delta(S)$ is the transition function, $O_i : S \to \Delta(\mathcal{O}_i)$ is the observation function, $R_i : S \times A_1 \times \cdots \times A_N \to \mathbb{R}$ is the reward function for agent i, and $\gamma \in [0, 1)$ is the discount factor. Here, depending on the reward structure, a POMG can represent various types of multi-agent settings: *cooperative* settings [13, 14, 15, 16], where all agents share an identical reward function (i.e., $r^1 = \cdots = r^N$); *adversarial* settings [8, 31], where agents have directly opposing objectives, often modeled as zero-sum (i.e., $\sum_{i=1}^N r^i = 0$); or *mixed-motive* settings [24, 17], where agents' rewards are neither fully aligned nor strictly opposed, creating simultaneous incentives for both cooperation and competition.

2.2 Mixed-motive Coordination in Multi-Agent RL

We consider mixed-motive settings, where agents' self-interest often conflicts with collective outcomes. Let us define the *collective return* as the average of *individual returns*: $R_{col} = \frac{1}{N}\sum_{i=1}^N R^i(s, a)$, where $R^i(s, a)$ is the individual return of Agent i. In the context of gradient-based learning, a conflict occurs when the local and collective return gradients are misaligned, that is, when $\nabla_{\theta_i}\mathbb{E}\left[R^i\right] \cdot \nabla_{\theta_i}\mathbb{E}\left[R_{\text{col}}\right] < 0$, where θ_i denotes the parameters of Agent i's policy.

To enhance cooperation (i.e. maximize collective reward) while avoiding conflicts, a variety of approaches have been proposed, including inequity aversion [10, 30], social influence [12], reciprocal reward shaping [33], formal contract mechanisms [9], and gifting-based cooperation [23, 17]. Many of these approaches are studied in the context of *Sequential Social Dilemmas (SSDs)* [18], a prominent class of mixed-motive settings in which agents repeatedly arbitrate between short-term selfish actions and long-term collective returns. For example, [17] proposed a gift-based method that balances altrusim and self-interest based based on social relationships between agents. [12] proposed an intrinsic motivation method that rewards agents for exerting causal influence over others' actions, thereby improving coordination in SSDs. [10] introduced inequity-averse agents that learn to cooperate by assigning temporal credit to prosocial behavior and penalizing inequitable outcomes. The aforementioned methods can be broadly viewed as forms of reward shaping, wherein additional intrinsic or socially-informed rewards guide agents toward cooperative behavior.

In contrast to reward shaping approaches, recent work [20] has explored direct optimization in the gradient space to reconcile individual and collective objectives. Specifically, the Altruistic Gradient Adjustment (AgA) method [20] modifies the policy gradients of both the collective and individual objectives, pulling agents toward stable fixed points of the collective objective and pushing them away from unstable ones. The adjusted gradient for Agent i is defined as $g^i_{aga} = g_{col} + \lambda(g^i_{ind} + H^T_{col} g_{col})$, where g_{col} and g_{ind} are the gradients of the collective and individual objectives for Agent i, H^T_{col} is the Hessian of the collective return with respect to the policy parameters, and λ is the adjustment coefficient and its sign is determined by $\text{sign}[(g_{\text{col}} \cdot H^T_{col} g_{col})\ [(g^i_{ind} \cdot H^T_{col} g_{col}) + \|H^T_{col} \cdot g_{col}\|^2]]$. This adjustment steers the update direction according to the local stability of the collective objective. Despite its effectiveness, AgA incurs additional computational complexity, focuses on the stability of the collective objective rather than directly resolving gradient conflicts, and provides no guarantees of monotonic improvement or fairness.

2.3 Gradient Adjustment

Gradient adjustment approaches have been actively investigated in multi-task learning [32, 22, 26, 28]. For example, CAGrad [22] formulates a quadratic program to compute a conflict-averse convex combination of gradients, achieving better trade-offs at the cost of increased complexity, and Nash-MTL [26] frames the task-weighting problem as a bargaining game, using the Nash bargaining solution to promote fairness and efficiency across tasks. Another method that inspires our work is PCGrad [32], which mitigates conflicts by projecting each conflicting gradient onto the normal plane of the other, offering a simple yet effective solution with low computational overhead. Specifically, when two gradients g_1 and g_2 are conflicted, PCGrad adjusts them by projecting one onto the normal plane of the other, i.e., $\tilde{g}_1^{PCGrad} = g_1 - \frac{g_1 \cdot g_2}{\|g_2\|^2} g_2$, and then uses the average of $\tilde{g}_1^{PCGrad}$ and $\tilde{g}_2^{PCGrad}$ as the final update.

2.4 Fairness in Multi-agent RL

Fairness concerns how returns are distributed among agents rather than how large the total return is, making it complementary, but often orthogonal to cooperation and efficiency. Fairness has been considered in multi-agent RL literature in both cooperative and mixed-motive settings [34, 6, 1, 29]. For example, in cooperative settings, [34] formulates fairness as the optimization of a fair social welfare function and [6] proposes a method for achieving team fairness by enforcing permutation-equivariant policies, which mitigate emergent unfairness caused by asymmetric role assignment. In mixed-motive settings, [17] shows enhanced fairness when measuring the sum of individual rewards and gifts, whereas in this paper we evaluate fairness using individual rewards only. [10], inspired by the literature on inequality in economics [5], explicitly leverages fairness by adding both disadvantage and advantage inequality terms to the reward of each agent to improve cooperation in SSD. Specifically, the shaped reward for Agent is $r^i = r^i - \alpha_{IA}/(N-1) \sum_{j \neq i} \max(r_j - r_i, 0) - \beta_{IA}/(N-1) \sum_{j \neq i} max(r_i - r_j, 0)$, where α_{IA} and β_{IA} weight disadvantage and advantage inequity, respectively.

Note that throughout this paper, we define fairness in terms of task-defined extrinsic individual rewards, the quantities that agents are fundamentally trained to maximize, and do not consider intrinsic rewards such as gifting, as they do not directly reflect actual participation in the underlying task. A more detailed discussion on this assumption is provided in Appendix A.

Algorithm 1: FCGrad

Input: Policy parameters θ, learning rate η, weighting factor β

Compute $g_{\text{ind}} := \nabla_\theta V_{\text{ind}}(\theta)$, $g_{\text{col}} := \nabla_\theta V_{\text{col}}(\theta)$

if $\langle g_{ind}, g_{col} \rangle \geq 0$ **then**

  $g_{\text{FCGrad}} \leftarrow (1-\beta) g_{\text{ind}} + \beta g_{\text{col}}$;

else

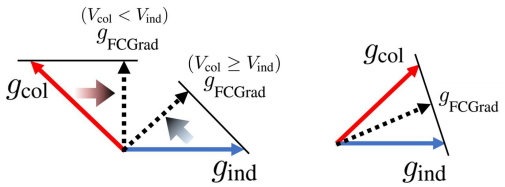
  if $V_{col}(\theta) \geq V_{ind}(\theta)$ **then**

    $g_{\text{FCGrad}} \leftarrow g_{\text{ind}} - \frac{\langle g_{\text{col}}, g_{\text{ind}} \rangle}{\|g_{\text{col}}\|^2} g_{\text{col}}$;

  else

    $g_{\text{FCGrad}} \leftarrow g_{\text{col}} - \frac{\langle g_{\text{ind}}, g_{\text{col}} \rangle}{\|g_{\text{ind}}\|^2} g_{\text{ind}}$;

Return $\theta \leftarrow \theta + \eta g_{\text{FCGrad}}$;



(a) Conflict (b) Non-conflict

Figure 1: FCGrad illustration: (a) When conflicts occur, the gradient corresponding to the lower objective—either individual or collective—is projected onto the normal plane of the gradient of the higher objective; (b) When no conflict is detected, a task-dependent weighted sum of the two gradients is applied.

3 Methodology

In mixed-motive settings, the individual and collective objectives may be either aligned or in conflict. When they are aligned, optimizing both objectives is sufficient, as neither interferes with the other. In such cases, an appropriately weighted combination of the two can be effective. However, when the objectives are in conflict, it becomes essential to explicitly address the interference between them, as prioritizing one may hinder the other. This is because focusing solely on the individual objective may hinder learning in tasks where cooperative behavior is essential for maximizing individual returns, while focusing solely on the collective objective may compromise fairness among agents. Therefore, it is important to (1) recognize when such conflicts arise and (2) correspondingly adjust the individual and collective objectives, appropriately considering both fairness and cooperation.

To this end, we propose a fair and conflict-aware gradient adjustment method, called FCGrad, which guarantees the monotonic non-decrease of both individual and collective objectives, while preserving fairness across individual objectives. Specifically, when the individual and collective gradients are in conflict, FCGrad projects the gradient associated with the lower expected return onto the normal plane of the other. This projected gradient remains a valid ascent direction for its own objective while avoiding interference with the other, and is then used as the final update. For example, if the individual expected return is lower than the collective expected return, indicating that the agent is disadvantaged in terms of fairness, we project the gradient of the individual objective onto the normal plane of the collective gradient and use it as the update direction. The detailed procedure and a visual illustration of FCGrad are provided in Algorithm 1 and Fig. 1, respectively. In the following, we present the detailed method along with its theoretical analysis.

3.1 FCGrad: Fair and Conflict-aware Gradient Adjustment

We now describe how FCGrad operates from the perspective of Agent i. Let $\theta \in \mathbb{R}^d$ denote the parameters of the policy π_θ for Agent i. Let us define $V_{\text{ind}}(\theta)$ and $V_{\text{col}}(\theta)$ as the expected individual and collective returns, respectively, computed under the initial state distribution. Note that $V_{\text{ind}}(\theta)$ and $V_{\text{col}}(\theta)$ are the individual and collective objectives, respectively. Let $g_{\text{ind}} := \nabla_\theta V_{\text{ind}}(\theta)$ and $g_{\text{col}} := \nabla_\theta V_{\text{col}}(\theta)$ denote the gradients of the individual and collective objectives, respectively. These represent ascent directions for $V_{\text{ind}}(\theta)$ and $V_{\text{col}}(\theta)$, meaning that for a sufficiently small $\eta > 0$, the following holds: $V_{\text{ind}}(\theta + \eta g_{\text{ind}}) > V_{\text{ind}}(\theta)$ and $V_{\text{col}}(\theta + \eta g_{\text{col}}) > V_{\text{col}}(\theta)$.

FCGrad proceeds as follows: (1) check whether g_{ind} and g_{col} are in conflict by examining the sign of their inner product, where a negative inner product indicates the presence of a conflict. (2) if $\langle g_{\text{ind}}, g_{\text{col}} \rangle \geq 0$ (i.e., non-conflict), FCGrad uses the weighted sum of two gradients: $g = (1-\beta) g_{\text{ind}} + \beta g_{\text{col}}$, (3) $\langle g_{\text{ind}}, g_{\text{col}} \rangle < 0$ (i.e., conflict), FCGrad places more weight on the individual (collective) gradient when the collective (individual) objective is greater, in order to ensure fairness.

The corresponding gradient is given by

$$g_{\text{FCGrad}} = \begin{cases} \tilde{g}_{ind} & \text{if } (V_{\text{col}} \geq V_{\text{ind}}) \\ \tilde{g}_{col} & \text{if } (V_{\text{col}} < V_{\text{ind}}) \end{cases} \tag{1}$$

where $\tilde{g}_{\text{col}}$ and $\tilde{g}_{\text{ind}}$ are the projections of g_{col} and g_{ind}, respectively, onto the normal plane of another gradient vector, given by

$$\tilde{g}_{\text{col}} := g_{\text{col}} - \frac{\langle g_{\text{ind}}, g_{\text{col}} \rangle}{\|g_{\text{ind}}\|^2} g_{\text{ind}}, \qquad \tilde{g}_{\text{ind}} := g_{\text{ind}} - \frac{\langle g_{\text{col}}, g_{\text{ind}} \rangle}{\|g_{\text{col}}\|^2} g_{\text{col}} \tag{2}$$

(4) update the policy parameter with the step size η: $\theta \leftarrow \theta + \eta g$. Note that $\tilde{g}_{\text{col}}$ projects g_{col} onto the normal plane of g_{ind}. Thus, $\tilde{g}_{\text{col}}$ is still a valid ascent direction for the collective objective while preserving the individual reward. This indicates that FCGrad prioritizes the individual objective without compromising the collective one when the agent is in an unfair situation, i.e., when the individual objective is lower. Conversely, when the collective objective is lower, FCGrad prioritizes the collective objective without compromising the individual one.

3.2 Theoretical Analysis

In this section, we prove that FCGrad guarantees monotonically non-decreasing improvements in both the collective and individual objectives, and that both objectives converge to the same value. This ensures that the expected individual returns across agents also converge to the same value.

Theorem 3.1 *Assume $V_{ind}(\theta)$ and $V_{col}(\theta)$ are differentiable and L-smooth. Let the update direction g be defined as in Equation 1. Then, for a sufficiently small step size $\eta > 0$, the update $\theta \leftarrow \theta + \eta g$ yields monotonically non-decreasing improvements in both $V_{col}(\theta)$ and $V_{int}(\theta)$.*

Proof. See Appendix B.

Theorem 3.1 states that FCGrad ensures monotonic non-decreasing improvements in both $V_{\text{ind}}(\theta_t)$ and $V_{\text{col}}(\theta_t)$ under certain assumptions. Note that all agents are updated using FCGrad, so both the individual objectives of all agents and the collective objective, defined as the expected return averaged across agents, are improved accordingly. However, monotonic improvement alone does not guarantee fairness. To establish fairness, it is necessary to further show that the individual and collective values converge to the same value over time, which in turn implies that all agents' individual values also become equal. The next theorem formalizes this result by proving that the gap between $V_{\text{ind}}(\theta_t)$ and $V_{\text{col}}(\theta_t)$ vanishes under mild conditions.

Theorem 3.2 *Assume $V_{ind}(\theta)$ and $V_{col}(\theta)$ be L-smooth, and let $\delta_t := V_{ind}(\theta_t) - V_{col}(\theta_t)$ denote the value gap at iteration t. Assume the step size satisfies the Robbins–Monro conditions: $0 < \eta_t \leq |\delta_t|/L$ with $\sum_t \eta_t = \infty$ and $\sum_t \eta_t^2 < \infty$. Also assume conflict recurrence, meaning that for any $\epsilon > 0$ and any t, if $|\delta_t| \geq \epsilon$, then there exists $t' \geq t$ such that $(g_{ind,t'} \cdot g_{col,t'}) < 0$. Then, the value gap converges to zero:*

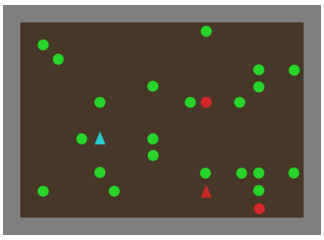
$$\lim_{t \to \infty} |V_{ind}(\theta_t) - V_{col}(\theta_t)| = 0. \tag{3}$$

Proof. See Appendix B.

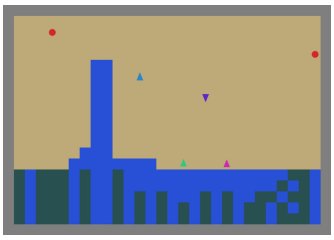
Theorem 3.2 states that the gap between the collective and individual objectives converges to zero under certain assumptions, including conflict recurrence, where conflicts occur continuously. This assumption is reasonable in mixed-motive settings, especially near equilibrium, because agents face inherent tensions between cooperation and self-interest, and as they approach equilibrium, misalignments in their objectives can continue to induce conflicts, even with small policy updates. Under the assumption that all agents use FCGrad, the individual objectives of all agents converge to the collective objective, and thus all individual objectives converge to the same value. This, in turn, implies that individual fairness is achieved.

3.3 Practical Algorithm

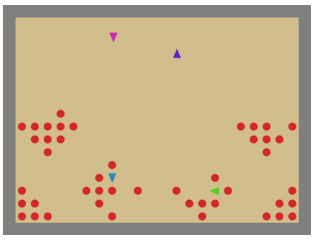
We now introduce a practical FCGrad-based multi-agent RL algorithm for mixed-motive settings. We consider decentralized training and execution, where each agent does not have access to other



(a) Unfair coins



(b) Cleanup



(c) Harvest

Figure 2: The environments considered in our experiments: (a) Unfair Coins—green coins appear more frequently than red coins, inducing fairness challenges; (b) Cleanup with distinct spawn positions—two agents (cyan and pink) spawn near waste areas, while the rest (blue and purple) spawn farther away; and (c) Harvest with distinct spawn positions—two agents (blue and green) spawn near apple (red) regions, while the rest spawn farther away.

agents' information but shares rewards, as commonly assumed in gifting mechanisms [23, 17]. Each agent trains its policy and value functions for both individual and collective returns solely based on its own local observations and the shared rewards. For this, we construct two separate value networks for the individual and collective objectives, while sharing a common encoder between them. Each value function is trained using generalized advantage estimation [27] to compute the corresponding advantage estimates. Using the two value functions, we compute the policy gradients of the individual and collective objectives, denoted as g_{ind} and g_{col}, respectively, via the PPO policy gradient. These gradients are then combined using FCGrad to determine the final update direction.

4 Experimental Results

4.1 Experimental Setup

Environments We conduct our experiments using the JAX-based codebase and environments provided by the SocialJAX suite [7]. We modify the existing environments—Coins, Cleanup, and Harvest—to incorporate a fairness perspective. Specifically, since Cleanup and Harvest already involve inherent fairness dilemmas, we introduce only minor changes by assigning distinct respawn positions to the agents. For the Coin Game, which originally focuses on the conflict between individual and collective objectives, we introduce asymmetry in the potential rewards that agents can obtain, creating a disparity in individual incentives. Fig.2 illustrates the considered environments. We provide detailed descriptions in the following sections.

Metric As our goal is to maximize returns while ensuring fairness, both performance and fairness metrics should be jointly considered for evaluation. We use α-*fairness* [25] as the evaluation metric, where, given individual returns $(r_1, \cdots, r_N)$, the fairness utility is defined as

$$U_\alpha(r_1, \cdots, r_N) = \begin{cases} \sum_{i=1}^{N} \frac{r_i^{1-\alpha}}{1-\alpha}, & \text{if } \alpha \neq 1, \\ \sum_{i=1}^{N} \log(r_i), & \text{if } \alpha = 1. \end{cases} \tag{4}$$

Notably, the fairness utility recovers several well-known objectives for specific values of α: it corresponds to the collective return when $\alpha = 0$, the geometric mean of individual rewards—also known as Nash Social Welfare—when $\alpha = 1$, and the minimum individual reward when $\alpha \to \infty$. Thus, $\alpha = 0$ reflects no consideration of fairness, and as α increases, the evaluation increasingly prioritizes fairness over aggregate performance. In summary, we consider the following three representative instances of α-fairness return in our evaluation: (i) average return (**Mean**, $\alpha = 0$), (ii) geometric mean return (**GeoMean**, $\alpha = 1$), and (iii) minimum individual return (**Min**, $\alpha \to \infty$). Note that α-*fairness return considers both performance and fairness*, where α determines the trade-off between them. The reported results are averaged over four random seeds.

Baselines We evaluate FCGrad with six baselines: (a) collective reward optimization (Col), (b) individual reward optimization (Ind), (c) inequity aversion reward restructuring (IA) [10], (d) weighted gradient combination of g_{ind} and g_{col} (denoted as Weighted), which corresponds to FCGrad without conflict handling, (e) PCGrad [32], and (f) Altruistic Gradient Adjustment (AgA) [20]. Note

that baselines (d)-(f) use the same architecture as FCGrad, where two separate value functions are trained for individual and collective objectives; they differ only in the policy update rule based on g_{ind} and g_{col}. All methods are implemented on top of the IPPO [3].

Hyperparameter We introduce a hyperparameter β for FCGrad, which determines the weight between the collective and individual objectives when there is no conflict. β plays a particularly important role in tasks that require high-level cooperation. We set β to 0.5, 0.7, and 0.8 for the Unfair Coin Game, Cleanup, and Harvest, respectively. The same values of β are used for the baseline method, Weighted. Additional hyperparameters for IPPO are provided in Appendix C.

4.2 Unfair Coins

The Coins environment [19] consists of two agents (green and red) and two types of coins, each associated with one of the agents. When a coin appears, it is assigned a color with probabilities p_{green} and p_{red}. An agent receives a reward of 1 for collecting any coin, regardless of its color. However, collecting a coin of the opposite color imposes a penalty of -2 on the other agent, creating a conflict between individual gain and cooperative behavior. In contrast to the original setting [19], where p_{green} and p_{red} are both set to 0.5—so that collecting coins matching each agent's color naturally aligns with fairness and also maximizes the collective reward—we consider an unfair variant where $p_{\text{green}} = 15/16$ and $p_{\text{red}} = 1/16$, introducing an inherent asymmetry in coin appearances. Although optimal collective performance still requires agents to collect coins matching their own color, this setup raises a fairness concern: the green agent receives substantially more rewards due to the higher frequency of green coins. To mitigate this imbalance and achieve a fairer outcome, the green agent must occasionally yield coins to the red agent, sacrificing some collective reward in favor of equity.

Results. In the Unfair Coin environment, achieving fairness requires the green agent to yield some of its coins to the red agent, thereby reducing its own reward. In other words, there exists a strong trade-off between collective performance and fairness. Therefore, we particularly focus on the performance trend with respect to α, as well as the **Min** performance, which places greater emphasis on fairness—the return of the most disadvantaged agent.

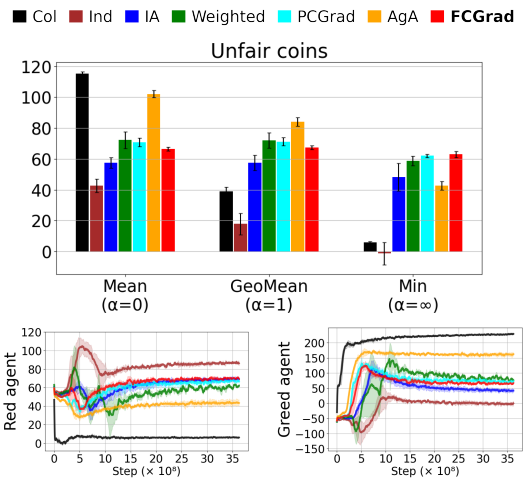
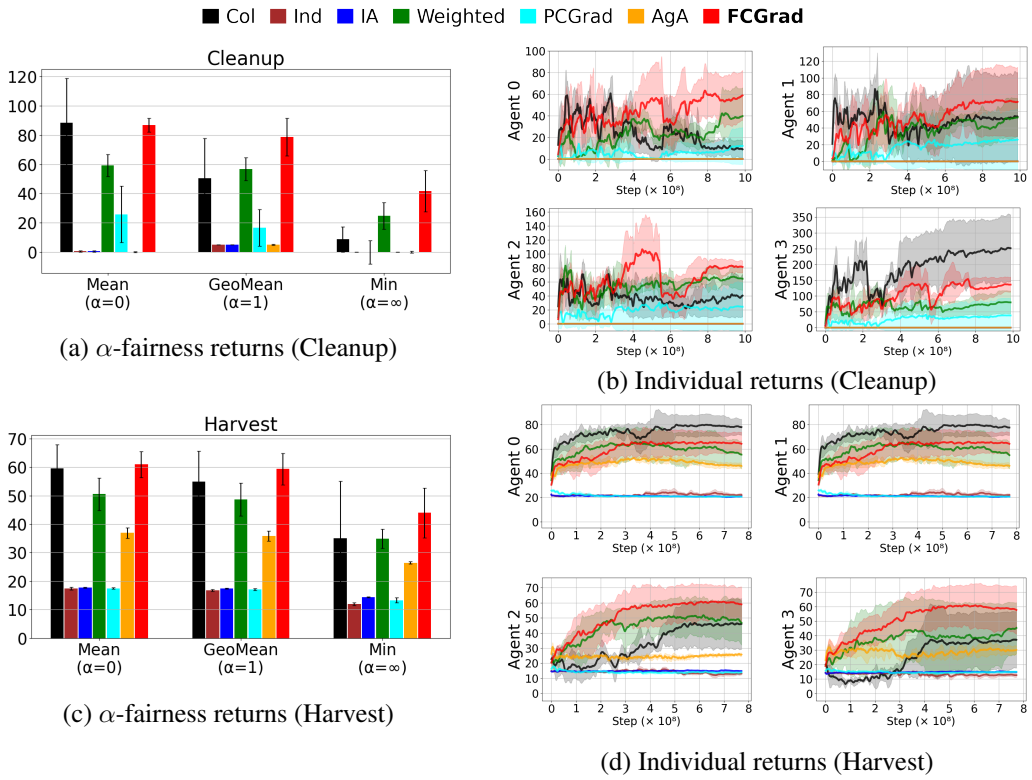


Figure 3: Top: Performance across agents in terms of mean, geometric mean, and minimum. Note that a higher value of α places more emphasis on fairness. Bottom: Agent-wise returns—Red and Green Agents.

Fig. 3 presents the α-fairness returns in the unfair coin environment (top) and the individual return of the green and red agents (bottom). The performance of Col, Ind, and AgA is observed to decrease more dramatically as α increases compared to FCGrad and PCGrad, which are conflict-aware methods. Interestingly, both the collective and individual approaches result in extremely unfair outcomes, but in opposite directions. Col, which maximizes collective reward, trains both agents to collect their own coins. As a result, the green agent, with more coin opportunities, gains higher returns, leading to unfairness. In contrast, with the individual objective, the red agent outperforms the green agent, possibly because the green agent is more frequently penalized by negative rewards due to the abundance of green coins. Meanwhile, the red agent learns without such penalties, accelerating its progress. However, FCGrad shows little variation across agents as α changes, indicating achieved fairness. Notably, FCGrad outperforms the baselines in terms of **Min** performance. As shown in Fig. 3, both the green and red agents converge to nearly identical returns, showing that fairness is effectively achieved.

4.3 Cleanup

The Cleanup environment consists of $N = 4$ agents, apples, and waste. Each agent receives a reward of 1 for collecting an apple. Apples grow in an orchard, but their growth depends on the



(a) α-fairness returns (Cleanup)

(b) Individual returns (Cleanup)

(c) α-fairness returns (Harvest)

(d) Individual returns (Harvest)

Figure 4: α-fairness returns and individual returns in the cleanup and harvest environments.

amount of waste present in the environment. Waste accumulates at a constant rate, and beyond a certain threshold, apple growth ceases entirely. Therefore, in order to sustain apple regrowth, some agents must sacrifice their immediate reward by cleaning up the waste. This creates a social dilemma, as the necessary act of cleaning benefits the group but does not provide direct individual reward, thereby generating a tension between self-interest and cooperative behavior. In contrast to the original configuration, where agents are randomly spawned across the map, we fix the spawn positions of agents: some (Agents 2 and 3 in our case) are placed near the apple orchard, while others (Agents 0 and 1) are positioned closer to the waste area. This spatial asymmetry further amplifies the conflict between fairness and efficiency. Note that, unlike the Unfair Coin, Cleanup introduces an intertemporal perspective, involving a trade-off between short-term individual interest and long-term collective interest [10].

Results. Fig. 4 (a) and (b) show the α-fairness performance and individual rewards during training in the Cleanup environment. FCGrad outperforms the baselines in terms of both **GeoMean** and **Min**, which reflect not only total return but also fairness. In addition, FCGrad achieves comparable performance to Col in terms of **Mean**, which is the optimization target of Col. As shown in Fig. 4 (b), under Col, Agent 3 learns to monopolize apple collection, while Agent 0 is trained to sacrifice by primarily cleaning waste. In contrast, FCGrad leads all four agents to obtain reasonably similar returns—demonstrating more fair behavior and achieving the best result in terms of **Min**. Since using the collective reward is essential in this environment, methods that rely heavily on individual rewards, such as Ind and IA, fail to learn effectively. In addition, AgA fails to properly balance between individual and collective objectives, also struggle to learn successfully.

4.4 Harvest

The Harvest environment features $N = 4$ agents and apples distributed across orchard patches. Each agent receives a reward of 1 per apple, but regrowth is stochastic and depends on nearby apples within a fixed radius. Over-harvesting depletes resources, risking environmental collapse, and thus agents must coordinate implicitly to sustain long-term returns. This creates a social dilemma between

short-term individual gain and long-term collective benefit. We also introduce spatial asymmetry: Agents 0 and 1 spawn near apples, while Agents 2 and 3 spawn farther away, making collection easier for the former. Similar to the Cleanup, Harvest also poses intertemporal challenges for both cooperation and fairness.

Results. Fig. 4 (c) and (d) show the α-fairness returns and individual agent returns during training. FCGrad outperforms the baselines across the considered α values. With the Col, Agents 0 and 1 achieve higher returns than Agents 2 and 3, indicating that they focus solely on collecting apples while accounting for the intertemporal dilemma, but not addressing the resulting unfairness toward Agents 2 and 3. In contrast, FCGrad leads all four agents to achieve similar returns, implying that Agents 0 and 1 take into account the outcomes of Agents 2 and 3. Similar to the results in Cleanup, methods that rely heavily on individual rewards, such as Ind and IA, perform poorly, though they achieve marginal learning. AgA performs better than the individual-reward-based methods, but still underperforms compared to FCGrad.

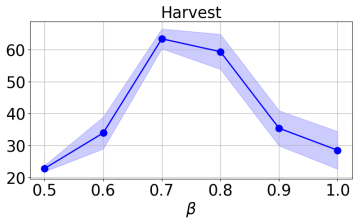


Figure 5: GeoMean of FCGrad with respect to β in the harvest environment.

4.5 Additional Analysis: Ablation and Fairness Metrics

Weighting factor: β determines the balance between the collective and individual objectives when no conflict is detected. It plays a particularly important role in tasks that require high-level cooperation. For example, in Cleanup, ignoring the collective objective makes it difficult for agents to discover how to improve their individual rewards. We observed this phenomenon in the previous section—solely maximizing individual rewards does not perform well. We present the **GeoMean** performance of FCGrad in the Harvest environment in Fig. 5, which shows that a β value between 0.7 and 0.8 yields the best performance. Thus, β reflects the required degree of cooperation over self-interest.

	Coin		**Cleanup**		**Harvest**	
Alg	Gini	Jain	Gini	Jain	Gini	Jain
Col	0.474	0.526	0.558	0.432	0.182	0.882
Ind	0.509	0.498	0.522	0.515	0.136	0.936
IA	0.122	0.942	0.536	0.497	**0.087**	**0.973**
Weighted	0.048	0.991	**0.266**	**0.801**	0.146	0.928
PCGrad	**0.039**	**0.994**	0.469	0.572	0.101	**0.965**
AgA	0.238	0.749	0.331	0.723	0.123	0.948
FCGrad	**0.010**	**0.999**	**0.223**	**0.835**	**0.093**	0.959

Table 1: Addtional fairness evaluation using Gini coefficient and Jain's index. Lower Gini and higher Jain indicate greater fairness. Top-2 most fair scores in each column are highlighted in bold.

Additional Fairness metrics: We additionally evaluate fairness using the Gini coefficient [2] and Jain's index [11]. The Gini coefficient is defined as $\text{Gini}(r_1, \cdots, r_N) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}|r_i - r_j|}{2N\sum_{i=1}^{N} r_i}$ and Jain's index is defined as $\text{Jain}(r_1, \cdots, r_N) = \frac{\left(\sum_{i=1}^{N} r_i\right)^2}{N\sum_{i=1}^{N} r_i^2}$, where both metrics range between 0 and 1 and lower Gini and higher Jain values indicate better fairness. Table 1 presents the results, showing that FCGrad generally achieves superior fairness.

5 Conclusion

In this work, we address the long-standing challenge of achieving both cooperation and fairness in mixed-motive multi-agent RL. We propose FCGrad, a conflict-aware gradient adjustment method that explicitly resolves gradient-level conflicts between individual and collective objectives. FCGrad dynamically adjusts the update direction based on which objective is more disadvantaged by projecting one gradient onto the normal plane of the other. We theoretically prove that this mechanism guarantees monotonic improvement and convergence of both objectives to the same value. Consequently, individual objectives across agents also converge, ensuring fairness. Extensive experiments in the Unfair Coin environment and sequential social dilemma settings, Cleanup and Harvest, demonstrate that FCGrad not only improves overall performance but also achieves superior fairness, as measured by α-fairness return metrics.

Limitation In practice, the recurrence of gradient conflicts, required for our theoretical guarantee, may not hold, as it can be influenced by the weighting factor in non-conflict cases. Understanding this interplay is a promising direction for future work.

Broader Impact Our work promotes fairness in learned behaviors, potentially preventing emergent inequalities in decentralized systems. We believe it has a positive societal impact.

6 Acknowledgement

This work was supported by the ONR MURI grant N00014-25-1-2116.

References

[1] Jasmine Jerry Aloor, Siddharth Nagar Nayak, Sydney Dolan, and Hamsa Balakrishnan. Cooperation and fairness in multi-agent reinforcement learning. *Journal on Autonomous Transportation Systems*, 2(2):1–25, 2024.

[2] Herbert A David. Miscellanea: Gini's mean difference rediscovered. *Biometrika*, 55(3):573–575, 1968.

[3] Christian Schroeder De Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip HS Torr, Mingfei Sun, and Shimon Whiteson. Is independent learning all you need in the starcraft multi-agent challenge? *arXiv preprint arXiv:2011.09533*, 2020.

[4] Rosemary Emery-Montemerlo, Geoff Gordon, Jeff Schneider, and Sebastian Thrun. Approximate solutions for partially observable stochastic games with common payoffs. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004.*, pages 136–143. IEEE, 2004.

[5] Dirk Engelmann and Martin Strobel. Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American economic review*, 94(4):857–869, 2004.

[6] Niko A Grupen, Bart Selman, and Daniel D Lee. Cooperative multi-agent fairness and equivariant policies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9350–9359, 2022.

[7] Zihao Guo, Richard Willis, Shuqing Shi, Tristan Tomilin, Joel Z Leibo, and Yali Du. Socialjax: An evaluation suite for multi-agent reinforcement learning in sequential social dilemmas. *arXiv preprint arXiv:2503.14576*, 2025.

[8] Songyang Han, Sanbao Su, Sihong He, Shuo Han, Haizhao Yang, Shaofeng Zou, and Fei Miao. What is the solution for state-adversarial multi-agent reinforcement learning? *Transactions on Machine Learning Research*.

[9] Andreas Haupt, Phillip Christoffersen, Mehul Damani, and Dylan Hadfield-Menell. Formal contracts mitigate social dilemmas in multi-agent reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 38(2):1–38, 2024.

[10] Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, et al. Inequity aversion improves cooperation in intertemporal social dilemmas. *Advances in neural information processing systems*, 31, 2018.

[11] Rajendra K Jain, Dah-Ming W Chiu, William R Hawe, et al. A quantitative measure of fairness and discrimination. *Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA*, 21(1), 1984.

[12] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*, pages 3040–3049. PMLR, 2019.

[13] 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. PMLR, 2022.

[14] Woojun Kim, Whiyoung Jung, Myungsik Cho, and Youngchul Sung. A variational approach to mutual information-based coordination for multi-agent reinforcement learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 40–48, 2023.

[15] Woojun Kim and Youngchul Sung. An adaptive entropy-regularization framework for multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 16829–16852. PMLR, 2023.

[16] Woojun Kim and Youngchul Sung. Parameter sharing with network pruning for scalable multi-agent deep reinforcement learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 1942–1950, 2023.

[17] Fanqi Kong, Yizhe Huang, Song-Chun Zhu, Siyuan Qi, and Xue Feng. Learning to balance altruism and self-interest based on empathy in mixed-motive games. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[18] JZ Leibo, VF Zambaldi, M Lanctot, J Marecki, and T Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *AAMAS*, volume 16, pages 464–473. ACM, 2017.

[19] Adam Lerer and Alexander Peysakhovich. Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *arXiv preprint arXiv:1707.01068*, 2017.

[20] Yang Li, Wenhao Zhang, Jianhong Wang, Shao Zhang, Yali Du, Ying Wen, and Wei Pan. Aligning individual and collective objectives in multi-agent cooperation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[21] Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

[22] Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in Neural Information Processing Systems*, 34:18878–18890, 2021.

[23] Andrei Lupu and Doina Precup. Gifting in multi-agent reinforcement learning. In *Proceedings of the 19th International Conference on autonomous agents and multiagent systems*, pages 789–797, 2020.

[24] Kevin R McKee, Ian Gemp, Brian McWilliams, Edgar A Duéñez-Guzmán, Edward Hughes, and Joel Z Leibo. Social diversity and social preferences in mixed-motive reinforcement learning. *arXiv preprint arXiv:2002.02325*, 2020.

[25] Jeonghoon Mo and Jean Walrand. Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on networking*, 8(5):556–567, 2000.

[26] Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In *International Conference on Machine Learning*, pages 16428–16446. PMLR, 2022.

[27] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

[28] Dmitry Senushkin, Nikolay Patakin, Arseny Kuznetsov, and Anton Konushin. Independent component alignment for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20083–20093, 2023.

[29] Martin Smit and Fernando P Santos. Learning fair cooperation in mixed-motive games with indirect reciprocity. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 220–228, 2024.

[30] Jane X Wang, Edward Hughes, Chrisantha Fernando, Wojciech M Czarnecki, Edgar A Duéñez-Guzmán, and Joel Z Leibo. Evolving intrinsic motivations for altruistic behavior. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 683–692, 2019.

[31] Lantao Yu, Jiaming Song, and Stefano Ermon. Multi-agent adversarial inverse reinforcement learning. In *International Conference on Machine Learning*, pages 7194–7201. PMLR, 2019.

[32] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836, 2020.

[33] John L Zhou, Weizhe Hong, and Jonathan Kao. Reciprocal reward influence encourages cooperation from self-interested agents. *Advances in Neural Information Processing Systems*, 37:59491–59512, 2024.

[34] Matthieu Zimmer, Claire Glanois, Umer Siddique, and Paul Weng. Learning fair policies in decentralized cooperative multi-agent reinforcement learning. In *International conference on machine learning*, pages 12967–12978. PMLR, 2021.

A Discussion on Fairness

In this appendix, we clarify the modeling assumption underlying our notion of fairness and contrast it with alternative perspectives such as reward-redistribution-based fairness. Our formulation intentionally focuses on a different regime: fairness is grounded purely in task-defined extrinsic rewards that directly reflect each agent's actual behavior, rather than assuming the availability of contract or currency-like mechanisms (e.g., reward exchanges or gifting) for compensating agents.

A.1 Extrinsic-Reward-Based Fairness vs. Reward Redistribution

Prior approaches such as gifting or incentive mechanisms (e.g., [9, 17, 23]) allow agents to redistribute rewards among one another—often interpreted as a currency-like signal or contract that enables division of labor. Under such assumptions, reward transfers are considered real, tangible returns and can be used to compensate agents for sacrificial roles (e.g., pollution cleaning without harvesting apples).

In contrast, our work adopts a more *primitive* perspective of fairness: we consider only *task-defined extrinsic rewards* that arise directly from environment state–action outcomes (e.g., rewards from collecting apples in the Cleanup environment). We intentionally do not assume the existence of an auxiliary payment mechanism, such as money or transferable reward tokens, that is external to the environment dynamics. From this viewpoint, if one agent continuously cleans while others only harvest apples, such an outcome is deemed unfair unless the cleaner also receives direct extrinsic returns. This modeling choice focuses on fairness that reflects *actual participation in the task*, rather than contractual compensation.

A.2 Pareto Optimality and the Role of α-Fairness

We emphasize that our objective is not to compute or approximate a game-theoretic equilibrium (e.g., Nash or correlated equilibrium), but rather to learn *Pareto-optimal* outcomes. In particular, α-fairness is used only as an evaluation metric, not as a training objective. By varying α, one can evaluate different trade-offs between pure efficiency ($\alpha = 0$), multiplicative balance ($\alpha = 1$), and max-min fairness ($\alpha \to \infty$). FCGrad yields outcomes that lie on the Pareto frontier across these trade-offs: in terms of collective return it performs comparably to the best baselines, while in terms of $\alpha = 1$ or $\alpha = \infty$ it significantly improves fairness without degrading performance.

B Theoretical Results

Lemma B.1 *Let $J : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable and L-smooth function. Let $g_1 = \nabla_\theta J(\theta)$ be the gradient of J at point θ, and let $g_2 \in \mathbb{R}^d$ be any vector satisfying $\langle g_1, g_2 \rangle > 0$. Then, for small step size $\eta < \frac{2\langle g_1, g_2 \rangle}{L\|g_2\|^2}$, the update $\theta \leftarrow \theta + \eta g_2$ yields a strict improvement:*

$$J(\theta + \eta g_2) > J(\theta).$$

Proof. Since J is L-smooth, for any $\theta \in \mathbb{R}^d$, update direction $g_2 \in \mathbb{R}^d$, and step size $\eta > 0$, the following inequality holds:

$$J(\theta + \eta g_2) \geq J(\theta) + \eta \langle \nabla_\theta J(\theta), g_2 \rangle - \frac{L}{2}\eta^2 \|g_2\|^2.$$

Let $g_1 = \nabla_\theta J(\theta)$. Then:

$$J(\theta + \eta g_2) \geq J(\theta) + \eta \langle g_1, g_2 \rangle - \frac{L}{2}\eta^2 \|g_2\|^2.$$

Define the right-hand side as a function of η:

$$\Delta(\eta) := \eta \langle g_1, g_2 \rangle - \frac{L}{2}\eta^2 \|g_2\|^2.$$

Since $\langle g_1, g_2 \rangle > 0$, this is a concave quadratic function that is positive for small enough η. Specifically, the inequality $\Delta(\eta) > 0$ holds when:

$$\eta < \frac{2\langle g_1, g_2 \rangle}{L\|g_2\|^2}.$$

Therefore, for any $\eta \in \left(0,\ \frac{2\langle g_1, g_2 \rangle}{L\|g_2\|^2}\right)$, we have:

$$J(\theta + \eta g_2) > J(\theta).$$

Theorem B.2 *Assume $V_{ind}(\theta)$ and $V_{col}(\theta)$ are differentiable and L-smooth. Let the update direction g be defined as in Equation 1. Then, for a sufficiently small step size $\eta > 0$, the update $\theta \leftarrow \theta + \eta g$ yields monotonically non-decreasing improvements in both $V_{col}(\theta)$ and $V_{int}(\theta)$.*

We consider three cases:

Case 1: (Non-conflict) $g_{\text{ind}} \cdot g_{\text{col}} \geq 0$. Then $g = \beta g_{\text{ind}} + (1-\beta) g_{\text{col}}$. Since $g_{\text{ind}}, g_{\text{col}}$ are ascent directions for $V_{\text{ind}}, V_{\text{col}}$, respectively, their convex combination also satisfies:

$$g_{\text{ind}} \cdot g = \beta \|g_{\text{ind}}\|^2 + (1-\beta) g_{\text{ind}} \cdot g_{\text{col}} > 0 \tag{5}$$

$$g_{\text{col}} \cdot g = \beta g_{\text{col}} \cdot g_{\text{ind}} + (1-\beta) \|g_{\text{col}}\|^2 > 0 \tag{6}$$

Since $g_{\text{ind}} \cdot g$ and $g_{\text{ind}} \cdot g$ are positive, according to Lemma 3.1, g yields a strict improvement in both V_{ind} and V_{col}.

Case 2: (Conflict) $g_{\text{ind}} \cdot g_{\text{col}} < 0$ and $V_{\text{ind}}(\theta) < V_{\text{col}}(\theta)$. We then use: $g = g_{\text{ind}} - \frac{g_{col} \cdot g_{\text{ind}}}{\|g_{col}\|^2} g_{col}$. Now,

$$g_{\text{ind}} \cdot g = g_{\text{ind}} \cdot g_{\text{ind}} - \frac{(g_{\text{ind}} \cdot g_{\text{col}})}{\|g_{\text{col}}\|^2}(g_{\text{ind}} \cdot g_{\text{col}}) = \frac{\|g_{\text{ind}}\|^2 \|g_{\text{col}}\|^2 - (g_{\text{ind}} \cdot g_{\text{col}})^2}{\|g_{\text{col}}\|^2} > 0$$

$$g_{\text{col}} \cdot g = g_{\text{col}} \cdot g_{\text{ind}} - \frac{(g_{\text{ind}} \cdot g_{\text{col}})}{\|g_{\text{col}}\|^2}\langle g_{\text{col}}, g_{\text{col}} \rangle = 0 \tag{7}$$

Since $g_{\text{ind}} \cdot g$ is positive, according to Lemma 3.1, g yields a strict improvement in V_{ind}. In addition, since $g_{\text{col}} \cdot g$ is zero, g does not decrease V_{col}.

Case 3: (Conflict) $g_{\text{ind}} \cdot g_{\text{col}} < 0$ and $V_{\text{ind}}(\theta) > V_{\text{col}}(\theta)$. Symmetric to Case 2: g yields a strict improvement in both V_{col} and does not decrease V_{ind}.

Thus, in all cases, g induces monotonically non-decreasing improvements in V_{ind} and V_{col}.

Lemma B.3 (Single conflict step) *When the conflict happens (i.e., $(g_{ind} \cdot g_{col}) < 0$), then for sufficiently small step size, $0 \le \eta_t \le \|\delta_t\|/L$, we have*

$$L_{t+1} - L_t \le -\frac{\eta_t}{2} |\delta_t| \, \|d_t\|^2. \tag{8}$$

Proof. When $\delta_t < 0$ (i.e. $V_{\text{col}} > V_{\text{ind}}$), we use $g = g_{\text{ind}} - \frac{g_{\text{ind}} \cdot g_{\text{col}}}{\|g_{\text{col}}\|^2} g_{\text{col}}$. Since L is L-smooth function, the following holds

$$L_{t+1} - L_t \le \eta_t (\nabla L_t \cdot g) + \frac{L}{2} \eta_t^2 \|g\|^2 \tag{9}$$

Here,

$$(\nabla L_t \cdot g) = \delta_t (g_{\text{ind}} - g_{\text{col}}) \cdot g = \delta_t (g_{\text{ind}} \cdot g - g_{\text{col}} \cdot g) = \delta_t (g_{\text{ind}} \cdot g) = \delta_t \|g\|^2 \tag{10}$$

Thus, we have

$$L_{t+1} - L_t \le \eta_t \delta_t \|g\|^2 + \frac{L}{2} \eta_t^2 \|g\|^2 \| \le \eta_t \delta_t \|g\|^2 + \frac{\|\delta_t\| \eta_t}{2} \|g\|^2 = -\frac{\eta_t}{2} \|\delta_t\| \|g\|^2 \tag{11}$$

Lemma B.4 (Single non-conflict step) *When the conflict does not happen, (i.e., $(g_{ind} \cdot g_{col}) \ge 0$), the proposed gradient is used. We assume that the step size meets the Robbins-Monro conditions (i.e. $\sum_{t=0}^{\infty} \eta_t = \infty$, $\sum_{t=0}^{\infty} \eta_t^2 < \infty$.) Then, the following holds:*

$$\sum_{t \in \mathcal{N}} \|L_{t+1} - L_t\| < \infty \tag{12}$$

where $\mathcal{N}$ is the set of all non-conflict indices.

Proof. $g = \beta g_{\text{ind}} + (1 - \beta) g_{\text{col}}$. Let us define $G := \sup_t \big(\|g_{1,t}\| + \|g_{2,t}\| \big)$ $(< \infty)$.

Since L is L-smooth, we have

$$L_{t+1} - L_t \le \eta_t \langle \nabla_\theta L_t, \, g \rangle + \frac{L}{2} \eta_t^2 \|g\|^2. \tag{13}$$

Since $\nabla_\theta L_t = \delta_t (g_{\text{ind}} - g_{\text{col}})$,

$$\begin{aligned}
\|\langle \nabla_\theta L_t, \, g \rangle\| &= \|\delta_t \langle g_{\text{ind}} - g_{\text{col}}, \beta g_{\text{ind}} + (1 - \beta) g_{\text{col}} \rangle\| && (14) \\
&\le |\delta_t| \Big[\beta \|g_{\text{ind}}\| \|g_{\text{ind}} - g_{\text{col}}\| + (1 - \beta) \|g_{\text{col}}\| \|g_{\text{ind}} - g_{\text{col}}\| \Big] \quad \text{(Cauchy-Schwarz)} && (15) \\
&\le |\delta_t| \Big[\beta \|g_{\text{ind}}\| + (1 - \beta) \|g_{\text{col}}\| \Big] 2G \le 2G^2 |\delta_t|. && (16)
\end{aligned}$$

Based on the assumption of the step size η ($\eta_t \le |\delta_t|/L$), we have

$$|\eta_t \langle \nabla_\theta L_t, d_t \rangle| \;\le\; 2G^2 \, |\delta_t| \, \eta_t \;\le\; 2G^2 L \, \eta_t^2. \tag{C}$$

Since $\|g\| = \|\beta g_{\text{ind}} + g_{\text{col}}\| \le \beta \|g_{\text{ind}}\| + (1 - \beta) \|g_{\text{col}}\| \le G$, the following holds.

$$\frac{L}{2} \eta_t^2 \|g\|^2 \le \frac{L}{2} \eta_t^2 G^2 \tag{17}$$

Combined above, we have

$$\|L_{t+1} - L_t\| \le (2G^2 L + \frac{L}{2} G^2) \eta_t^2 = \frac{5}{2} G^2 L \eta_t^2 \tag{18}$$

Define $C_0 := \frac{5}{2} G^2 L$ to obtain

$$|L_{t+1} - L_t| \;\le\; C_0 \, \eta_t^2.$$

Because $\sum_{t=0}^{\infty} \eta_t^2 < \infty$ (Robbins–Monro assumption),

$$\sum_{t \in \mathcal{N}} |L_{t+1} - L_t| \;\le\; C_0 \sum_{t \in \mathcal{N}} \eta_t^2 \;\le\; C_0 \sum_{t=0}^{\infty} \eta_t^2 \;<\; \infty.$$

Theorem B.5 *Let V_{ind} and V_{col} be L-smooth. Assume the step size satisfies the Robbins–Monro conditions: $0 < \eta_t \leq |\delta_t|/L$ with $\sum_t \eta_t = \infty$ and $\sum_t \eta_t^2 < \infty$. Also assume conflict recurrence, meaning that for any $\epsilon > 0$ and any t, if $|\delta_t| \geq \epsilon$, then there exists $t' \geq t$ such that $(g_{ind,t'} \cdot g_{col,t'}) < 0$. Then, the value gap converges to zero:*

$$\lim_{t \to \infty} |V_{ind}(\theta_t) - V_{col}(\theta_t)| = 0. \tag{19}$$

Proof. Denote conflict indices by $\mathcal{C}$ and non-conflict by $\mathcal{N}$. Lemma A.3 and Lemma A.4 give for every horizon T

$$L_T \;\leq\; L_0 - \frac{1}{2} \sum_{t \in \mathcal{C},\, t<T} \eta_t \, |\delta_t| \, \|d_t\|^2 + \; C_0 \sum_{t \in \mathcal{N},\, t<T} \eta_t^2. \tag{20}$$

According to the assumption of the Robbins-Monro, the following holds:

$$\sum_{t \in \mathcal{C}} \eta_t \, |\delta_t| \, \|d_t\|^2 < \infty. \tag{21}$$

For any conflict step the projection property and bounded gradients imply $\|g_t\| \geq \sigma > 0$ with $\sigma := \frac{1}{2} \min(\|g_{\text{ind},t}\|, \|g_{\text{col},t}\|)$. Thus, we have

$$\sum_{t \in \mathcal{C}} \eta_t \, |\delta_t| \leq \sigma^{-2} \sum_{t \in \mathcal{C}} \eta_t |\delta_t| \|g_t\|^2 < \infty. \tag{22}$$

Here, we use contradiction. Assume $\limsup_{t \to \infty} |\delta_t| = \varepsilon_0 > 0$. Set $\varepsilon := \varepsilon_0/2$. By the assumption, there exists an *infinite* set $\mathcal{C}_\varepsilon = \{\, t \in \mathcal{C} \mid |\delta_t| \geq \varepsilon \}$. Then for every $t \in \mathcal{C}_\varepsilon$, $\eta_t \, |\delta_t| \geq \eta_t \, \varepsilon$. Because $\sum_t \eta_t = \infty$, $\sum_{t \in \mathcal{C}_\varepsilon} \eta_t \, \varepsilon = \infty$, contradicting the finiteness of Eq. 22. Therefore, $\limsup_{t \to \infty} |\delta_t| = 0$.

C Implementation Details

All experiments were run on a local server equipped with an AMD EPYC 7713 64-Core CPU and five NVIDIA RTX 6000 Ada Generation GPUs. Each rollout consisted of 64–256 parallel environments depending on the task, and training time per run ranged from 2 to 8 hours. The official implementation of FCGrad is available at: `https://github.com/wjkim1202/fcgrad`.

C.1 Unfair Coin

Each agent has a CNN-based actor-critic network. The observation is processed through three convolutional layers with kernel sizes of 5×5, 3×3, and 3×3, each with 32 channels and ReLU activations, followed by a fully connected layer with 64 units. The actor head outputs a categorical distribution over discrete actions, while the critic consists of two separate heads estimating the individual and collective value functions.

We train the networks using the Adam optimizer with a learning rate of 1×10^{-4}, linearly annealed over time. PPO is used with a clipping threshold of 0.2 and two update epochs per iteration, using 500 minibatches. We collect trajectories from 256 parallel environments, each running for 1000 steps per rollout. The discount factor is set to $\gamma = 0.99$ and the GAE parameter to $\lambda = 0.95$. The entropy and value loss coefficients are set to 0.1, respectively. Gradients are clipped to a maximum global norm of 0.5.

C.2 Cleanup

Each agent is equipped with a convolutional actor-critical network. The observation is processed through three convolutional layers with kernel sizes of 5×5, 3×3, and 3×3, each with 32 channels and ReLU activations, followed by a fully connected layer with 64 units. The actor outputs a categorical distribution over discrete actions, and the critic consists of two heads that estimate the individual and collective value functions, respectively.

Training is performed using PPO with a clipping threshold of 0.2 and two update epochs per iteration. A total of 500 minibatches are used per update, with data collected from 64 parallel environments running 1000 steps per rollout. The discount factor is set to $\gamma = 0.99$, and the GAE parameter is set to $\lambda = 0.95$. We use the Adam optimizer with an initial learning rate of 5×10^{-4}, which is linearly annealed during training. The value loss coefficient and entropy coefficient are both set to 0.01, and the value function loss is weighted by 0.5. Gradients are clipped with a maximum global norm of 0.5.

C.3 Harvest

Each agent is equipped with a convolutional actor-critical network. The observation is processed through three convolutional layers with kernel sizes of 5×5, 3×3, and 3×3, each with 32 channels and ReLU activations, followed by a fully connected layer with 64 units. The actor outputs a categorical distribution over discrete actions, and the critic consists of two heads that estimate the individual and collective value functions, respectively.

Training is performed using PPO with a clipping threshold of 0.2 and two update epochs per iteration. A total of 500 minibatches are used per update, with data collected from 64 parallel environments running 1000 steps per rollout. The discount factor is set to $\gamma = 0.99$, and the GAE parameter is set to $\lambda = 0.95$. We use the Adam optimizer with an initial learning rate of 5×10^{-4}, which is linearly annealed during training. The entropy and value function loss coefficients are set to 0.01 and 0.5, respectively. Gradients are clipped to a maximum global norm of 0.5.

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 and introduction clearly state the core contributions and are consistent with both theoretical and empirical results.

   Guidelines:

   - The answer NA 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 NA 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: We stated the limitation in the conclusion.

   Guidelines:

   - The answer NA 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: We included the proof in the Appendix.

Guidelines:

- The answer NA 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: We provided the details in the Appendix and the main paper.

   Guidelines:

   - The answer NA 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: We cited the corresponding paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.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://nips.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, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provided the implementation details in the Appendix.

   Guidelines:

   - The answer NA 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: We provided the mean and variance of individual returns.

   Guidelines:

   - The answer NA 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: We stated this in the Appendix.

   Guidelines:

   - The answer NA 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: We follow the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA 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: We stated the broader impacts in the conclusion.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA 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: We have cited the paper.

    Guidelines:

    - The answer NA 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: [NA]

Justification: [NA]

Guidelines:

- The answer NA 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: [NA]

Justification: [NA]

Guidelines:

- The answer NA 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: [NA]

Justification: [NA]

Guidelines:

- The answer NA 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 rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA 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 (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.