IMPLICIT CREDIT LEARNING FOR SPARSE-REWARD MULTI-AGENT REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

When a cooperative team receives feedback only at the end of an episode, a single outcome must guide many agents' earlier actions. Learning a dense team reward helps, but its decomposition does not automatically give each agent a useful policy update. We introduce IMAP, an online method that fits a shared Q/V reward model by comparing completed episodes according to their observed terminal outcomes. Local Q/V residuals provide agent-specific rewards; the same normalized weights combine them for a centralized critic and scale local advantages for decentralized PPO actors. We characterize when these local updates follow the *learned* team objective. With full returns and frozen reward heads, a sharp bound relates their gradient error to variation in future Q/V compatibility; an exact identity identifies additional continuation terms for general GAE. In experiments with terminal-only feedback, IMAP's local-actor configuration attains the highest reported mean among the compared methods on all nine SMACv2 scenarios and four MAMuJoCo tasks, and on two of three SMACLite scenarios. Controlled relay interventions confirm the predicted changes in actor credit while holding the learned team reward fixed. The gradient result concerns the fitted objective; benchmark returns measure performance on the original task outcomes.

1 INTRODUCTION

In many cooperative tasks, agents act for a long time before the team receives a single outcome: a battle is won or lost, or a coordinated motion earns an episode return. That outcome identifies success, but gives no direct indication of which agent's action mattered or when. The problem is therefore both temporal and multi-agent credit assignment. It matters especially under centralized training with decentralized execution: training may use team information, but each deployed actor must choose actions from its own observations.

A learned dense reward can distribute terminal feedback across time, yet a team-level advantage still supplies every actor with the same reward signal. Decomposing that reward can provide local signals, but decomposition alone is insufficient. An agent's action may change the future states and rewards of other agents, so optimizing only its own component can depart from the learned team objective. Prior shared Q/V reward fitting uses residual comparisons and offline weighted policy extraction (Bui et al., 2025). The missing connection here is an online, terminal-outcome-driven local PPO update, together with a characterization of when its local credit tracks the team objective.

IMAP makes that connection with one reward model and one mixture. Completed episodes are ranked by their observed terminal outcomes to fit shared Q/V heads; no external preference labels are needed. Each head yields an inverse-Bellman residual that acts as a local reward. Normalized weights combine these residuals into the reward for a centralized critic and scale the corresponding advantages for decentralized PPO actors. Thus the critic and actors learn from the same fitted team score, while the actors receive distinct local credit. The method works with the observed win/loss signal or a cumulative return revealed only at episode end.

The central analytical question is whether distinct actor signals still point in the direction of the *learned* team return. The mixture makes global and weighted local advantages equal as numerical sums, but equality of those sums does not guarantee equality of their policy gradients: an action can affect other agents' future credit. We therefore study the local direction under frozen reward heads and evaluate task performance separately against the observed terminal outcome.

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

Our contributions are:

- **An online local-actor algorithm for terminal feedback.** IMAP turns comparisons among completed episodes into Q/V residual rewards, then uses one normalized mixture for a global critic and weighted local PPO updates. This connects shared Q/V fitting to online decentralized policy improvement, whereas the prior Q/V approach uses offline weighted policy extraction (Bui et al., 2025).
- **A precise account of when local credit follows the team score.** An exact identity isolates the continuation terms at general GAE. For full returns, a sharp trajectory-Fisher bound reduces the error to future Q/V compatibility without an explicit agent-count factor; a separate read-once result lower-bounds ideal local signal in structured coordination. The gradient guarantee concerns the fitted objective, with conditions for transfer to the true outcome given in Appendix A.9.
- **Evidence for both performance and the credit mechanism.** Under terminal-only feedback, IMAP's local-actor configuration has the highest reported mean among the compared methods on all nine SMACv2 scenarios and four MAMuJoCo tasks, and on two of three SMACLite scenarios. Model and actor-readout comparisons probe the design choices, while frozen relay interventions test the predicted credit changes at unchanged mixed reward.

2 Related work

Multi-agent policy optimization. CTDE enables agents to learn from global information during training while acting independently at execution (Foerster et al., 2018; Lowe et al., 2017; Kraemer & Banerjee, 2016). Value decomposition supplies factorized representations of a team objective (Sunehag et al., 2017; Rashid et al., 2020; Son et al., 2019; Wang et al., 2021a). MAPPO and HAPPO develop cooperative PPO updates (Schulman et al., 2017; Yu et al., 2022; Kuba et al., 2021). DOP derives weighted local policy gradients from a decomposed critic (Wang et al., 2021b); IMAP analyzes the temporal credit induced by learned Q/V residual rewards. COMA constructs a counterfactual advantage using a centralized Q-function (Foerster et al., 2018). IMAP instead obtains its local signals from a reward model fitted to terminal outcomes. The present analysis states when these signals reproduce the learned-team gradient and bounds their discrepancy otherwise.

Learning from delayed rewards. Reward redistribution converts episodic feedback into denser signals (Arjona-Medina et al., 2019; Ren et al., 2022). AREL, STAS and TAR^2 learn temporal or agent contributions (Xiao et al., 2022; Chen et al., 2024; Kapoor et al., 2025). DIFFER decomposes rewards for individual replay, LAIES uses model-based causal intrinsic rewards, and MASER generates replay-derived subgoals (Hu et al., 2023; Liu et al., 2023; Jeon et al., 2022). The latter three emphasize replay or exploration; IMAP studies local actor credit from terminal-outcome-fitted Q/V rewards.

Inverse value learning and actor geometry. The inverse-Bellman likelihood and exponential value update follow IPL/XQL (Hejna & Sadigh, 2023; Garg et al., 2023), building on inverse Q-learning (Garg et al., 2021). O-MAPL already combines shared multi-agent Q/V fitting, residual comparisons and Extreme-V with offline weighted policy extraction (Bui et al., 2025). IMAP specializes to normalized scalar weights, online terminal feedback and local PPO. Advantage-support analysis expresses bias through omitted score–reward couplings (Zhao & Li, 2026); our temporal analysis reduces them to future component compatibility and gives sharp trajectory-Fisher control, including shared parameters (Kakade, 2001).

3 Cooperative learning from terminal outcomes

We focus on cooperative MARL, modeled as a multi-agent partially observable Markov decision process. The team has agents $\mathcal{N} = \{1, \dots, n\}$ and joint action space $\mathcal{A} = \prod_i \mathcal{A}_i$. At stage t, the environment draws the next state from $P(\cdot \mid \mathbf{s}_t, \mathbf{a}_t)$. Agent i acts from its observation or history $o_{i,t}$, and the joint policy factorizes as $\pi_\theta(\mathbf{a}_t \mid \mathbf{s}_t) = \prod_i \pi_{i,\theta}(a_{i,t} \mid o_{i,t})$. Here $\mathbf{s}_t$ is a sufficient information state for analysis; it can contain the joint history. The actors remain decentralized. Their parameters may be shared.

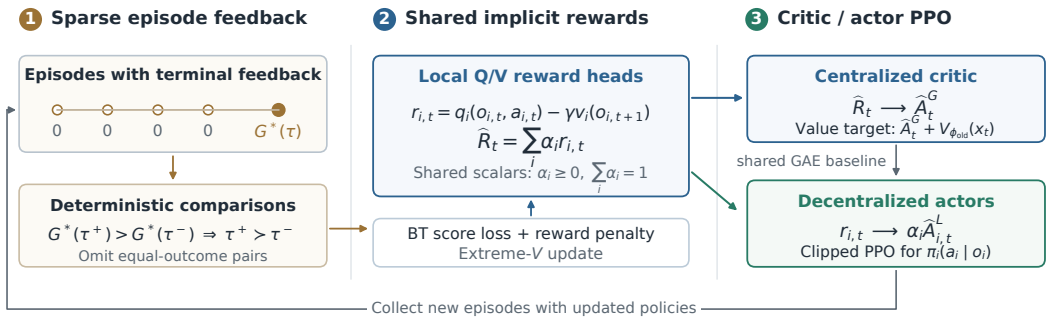

Figure 1: IMAP learns from sparse team outcomes. Completed episodes train a shared Q/V reward model; its local residuals supply a mixed reward for the centralized critic and agent-specific advantages for decentralized actors. Outcome comparisons are computed from the observed terminal returns.

A complete episode τ supplies one terminal team outcome $G^*(\tau)$. The learning objective is $J^*(\pi) = \mathbb{E}_\pi G^*(\tau)$. In the binary battle tasks, this outcome is the terminal win/loss signal. In the continuous-control tasks, the environment's cumulative episode return is withheld until termination. The learner never receives those intermediate task rewards during training. Evaluation uses the original team outcome in its original units.

The fitted model assigns each episode a discounted score $\widehat{F}(\tau)$, defined below from its per-step rewards. Its expectation $J_{\widehat{R}} = \mathbb{E}_\pi \widehat{F}$ is the *learned* team objective used in our credit analysis. This differs from the observed-outcome objective J^* unless additional calibration conditions hold; Appendix A.9 states those conditions.

4 Implicit Rewards and Local Policy Updates

Comparing terminal outcomes can fit a team-level trajectory score, but it does not tell a PPO actor which of its own actions to reinforce. Reading local terms from that score is also delicate: an action can change other agents' future rewards. IMAP's methodological contribution is a closed online loop. The current policy generates complete episodes; their observed outcomes label comparisons; fitted Q/V residuals supply dense rewards for the next PPO update. One normalized mixture defines the critic's team reward and weights the actors' local advantages. Unlike O-MAPL's offline preference learning and weighted behavior-cloning extraction (Bui et al., 2025), this loop updates the data-generating policy with local credit. Section 5 characterizes when that credit follows the learned team objective; Figure 1 depicts the loop.

4.1 A Shared Q/V Mixture for Team and Local Rewards

Additive value decomposition is established in cooperative MARL: VDN sums local values, DOP uses a weighted linear critic for decomposed policy gradients, and O-MAPL applies linear Q/V mixing to preference-based learning (Sunehag et al., 2017; Wang et al., 2021b; Bui et al., 2025). We use a normalized version because the critic and local PPO actors must read from the same fitted score. The local heads $q_i(o_i, a_i)$ and $v_i(o_i)$ have separate reward-model parameters. Their common mixer is

$$Q_{\text{tot}}(\mathbf{s}, \mathbf{a}) = \sum_i \alpha_i q_i(o_i, a_i), \qquad V_{\text{tot}}(\mathbf{s}) = \sum_i \alpha_i v_i(o_i), \tag{1}$$

$$\alpha_i = \frac{\text{softplus}(\omega_i)}{\sum_j \text{softplus}(\omega_j)}. \tag{2}$$

The nonnegative weights sum to one and are shared by Q and V. They are global scalars, with no additive bias. Thus the mixed reward is exactly the weighted sum of local Q/V residuals, while normalization preserves the common critic term in GAE (Equation (8)). Using these same α_i in local PPO ties each actor to its contribution to the fitted team score. This tractable class excludes

state-dependent mixing; our experiments do not compare mixer classes. The heads and weights are frozen during each actor update.

For a sampled transition $(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1})$, define

$$r_{i,t} = q_i(o_{i,t}, a_{i,t}) - \gamma v_i(o_{i,t+1}), \qquad \widehat{R}_t = \sum_i \alpha_i r_{i,t}. \tag{3}$$

Here γ is the discount factor. The next-observation value uses the successor reached in the rollout, and terminal values are zero. We call $r_{i,t}$ an *implicit* reward because it is computed from Q/V heads rather than predicted by a separate reward head. Averaging over transitions gives the expected-transition inverse-Bellman residual.

4.2 Learning dense rewards from terminal outcomes

The environment supplies no per-step training reward, but completed episodes can be ordered by their observed outcomes. For two episodes τ_1, τ_2, define $\tau_1 \succ \tau_2$ if $G^*(\tau_1) > G^*(\tau_2)$; discard ties. These labels come from online interaction, without an external preference source. Let $\widehat{F}(\tau) = \sum_t \gamma^t \widehat{R}_t$. The BT pair model and reward-fitting objective are

$$p(\tau^+ \succ \tau^-) = \mathrm{sigmoid}\left(\widehat{F}(\tau^+) - \widehat{F}(\tau^-)\right), \tag{4}$$

$$\mathcal{L}_Q = -\mathbb{E}_{\mathcal{P}} \log p(\tau^+ \succ \tau^-) + \lambda_R \mathbb{E}_{\mathcal{D}} \widehat{R}_t^2. \tag{5}$$

Here $\mathcal{P}$ denotes outcome-ordered episode pairs and $\mathcal{D}$ stored transitions; λ_R weights a squared-reward regularizer. The BT loss trains a trajectory score from sparse outcomes, while the regularizer controls its scale. Ordering discards continuous-return magnitudes, so good ranking does not establish calibration to G^*. Q-head and mixer updates use stored episodes with V-heads fixed.

The V-heads complete the residual reward: fitting episode rankings through Q alone would leave its next-state subtraction unspecified. Following Extreme-V (Garg et al., 2023), we encourage the mixed V value to track a soft value of the mixed Q-function:

$$\mathcal{L}_V = \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}} \left[\exp\left(\frac{Q_{\mathrm{tot}}(\mathbf{s},\mathbf{a}) - V_{\mathrm{tot}}(\mathbf{s})}{\beta} \right) - \frac{Q_{\mathrm{tot}}(\mathbf{s},\mathbf{a}) - V_{\mathrm{tot}}(\mathbf{s})}{\beta} - 1 \right]. \tag{6}$$

The temperature is $\beta > 0$. The V loss uses buffered actions; Section 5 instead measures each V-head's mismatch with its Q-head's current-policy mean at frozen checkpoints. Fitting the mixed V need not eliminate that mismatch. The PPO critic is trained separately.

4.3 Global evaluation with weighted local PPO

MAPPO gives every actor the same mixed-reward advantage (Yu et al., 2022). IMAP gives actors distinct residual-based advantages while a centralized critic $V_\phi(\mathbf{s}_t)$ evaluates the team reward. This may sharpen sparse-feedback credit; Section 6 compares the two readouts. Each actor $\pi_{i,\theta}$ uses only its local observation or history. Fix its advantage-computation snapshot $V_{\phi_{\mathrm{old}}}$ and write

$$\begin{aligned} B_t &= \gamma V_{\phi_{\mathrm{old}}}(\mathbf{s}_{t+1}) - V_{\phi_{\mathrm{old}}}(\mathbf{s}_t), \\ \delta_t^G &= \widehat{R}_t + B_t, \qquad \delta_{i,t}^L = r_{i,t} + B_t, \\ \widehat{A}_t^G &= \sum_{\ell\geq 0} (\gamma\lambda)^\ell \delta_{t+\ell}^G, \qquad \widehat{A}_{i,t}^L = \sum_{\ell\geq 0} (\gamma\lambda)^\ell \delta_{i,t+\ell}^L. \end{aligned} \tag{7}$$

Here λ is the GAE trace parameter. Both streams use the same terminal or rollout masks and bootstrap convention; on complete episodes they stop at the true terminal boundary. Because the mixer weights sum to one, the sampled advantages obey the exact identity

$$\widehat{A}_t^G = \sum_i \alpha_i \widehat{A}_{i,t}^L. \tag{8}$$

Mixing the local TD residuals gives the global residual, and the common GAE sum preserves that equality. The identity is an implementation check, but it does not imply that local and global actor gradients agree: each actor multiplies its own score by a different advantage. Section 5 analyzes the difference.

Algorithm 1 IMAP with terminal-outcome feedback

1: Initialize actors, critic, Q/V heads, normalized mixer, and both buffers.
2: **for** each iteration **do**
3: Collect rollouts and append completed episodes with their terminal outcomes.
4: Fit Q/mixer by (5) and V by (6) using available unequal-return pairs.
5: Freeze model/critic snapshots; compute both GAE streams with common masks.
6: Run PPO epochs using local actor targets and global critic targets.
7: Clear the on-policy rollout buffer.
8: **end for**

For each agent define the individual likelihood ratio $\rho_{i,t}(\theta) = \pi_{i,\theta}(a_{i,t} \mid o_{i,t})/\pi_{i,\theta_{\text{old}}}(a_{i,t} \mid o_{i,t})$. The IMAP actor objective is

$$\mathcal{L}_{\text{actor}}(\theta) = \mathbb{E}_t \sum_i \left[\alpha_i \min\Big\{\rho_{i,t}\widehat{A}^L_{i,t}, \text{clip}(\rho_{i,t}, 1-\epsilon_{\text{PPO}}, 1+\epsilon_{\text{PPO}})\widehat{A}^L_{i,t}\Big\} + \eta\mathcal{H}(\pi_{i,\theta})\right]. \quad (9)$$

Here ϵ_{PPO} is the clipping width, η is the entropy coefficient, and $\mathcal{H}$ denotes policy entropy. The ratio uses the policy snapshot θ_{old} that collected the rollout. The coefficient α_i is the component's weight in the fitted team reward, so it also scales that actor's PPO reward term, as in decomposed policy gradients (Wang et al., 2021b). Global controls use $\widehat{A}^G_t$ with a unit prefactor; both variants retain the same entropy term. Here $\mathbb{E}_t$ averages rollout actor samples. For a one-stage episode with frozen heads and critic, conditional action independence gives the expected old-policy reward-gradients

$$g_G = \mathbb{E}\sum_i z_i\Big(\sum_j \alpha_j q_j - V\Big) = \sum_i \alpha_i \mathbb{E}[z_i q_i] = g_L, \quad (10)$$

where $z_i = \nabla_\theta \log \pi_i$, V is the shared frozen baseline, and $g_L = \mathbb{E}\sum_i \alpha_i z_i(q_i - V)$. This equality concerns the expected old-policy reward gradient for one stage, not the full clipped PPO update. Appendix A.2 gives the shared-parameter proof. Section 5 characterizes the sequential discrepancy.

The critic uses the global target $\widehat{Y}_t = \widehat{A}^G_t + V_{\phi_{\text{old}}}(\mathbf{s}_t)$ and MAPPO's clipped value loss. Reward-model parameters and advantage targets remain fixed during PPO epochs; critic updates do not change these targets.

Common affine reward normalization and masks preserve the mixture identities. Appendix C specifies empirical reductions and normalization; Appendix A.7 gives bootstrap corrections.

5 When local credit follows the team objective

Equation (10) explains the local coefficient for one decision. Over an episode, an agent's action can alter the states in which other agents later act. Their future reward components can then change even when the current local term is correctly weighted. We first identify this continuation error at the GAE trace parameter and then bound it sharply for complete returns.

Setting. Consider an H-stage Markov game with discount $\gamma \in [0, 1]$ and a sufficient pre-action state $\mathbf{S}_h$, which may contain joint histories. Given $\mathbf{S}_h$, the agents sample actions independently from their local policies. Actor parameters may be shared. The initial law, transitions and observations are independent of these parameters. Assume finite state and action sets and differentiable policies positive on fixed feasible supports; Appendix A.10 covers general spaces. Freeze the local heads, the critic V_h, and scalar weights $\alpha_i \geq 0$, $\sum_i \alpha_i = 1$, with $v_{i,H} = V_H = 0$. Each Q-head uses only its own actor's action.

Define the learned-team objective and local actor vector by

$$J_{\widehat{R}}(\theta) = \mathbb{E}_{\pi_\theta} \sum_{h=0}^{H-1} \gamma^h \widehat{R}_h, \quad (11)$$

$$g_L^{(\lambda)} = \mathbb{E}\sum_{h=0}^{H-1}\sum_i \gamma^h \alpha_i z_{i,h}\widehat{A}^L_{i,h}, \qquad z_{i,h} = \nabla_\theta \log \pi_{i,\theta}(A_{i,h} \mid o_{i,h}). \quad (12)$$

Here $\nabla J_{\widehat{R}} := \nabla_\theta J_{\widehat{R}}(\theta)$ differentiates the policy with the reward model fixed. The local vector is the reward-part derivative at the old policy with objective-aligned time weights. Entropy, the empirical rollout reduction and later clipped PPO steps are separate.

The key quantity is how well each V-head matches the current policy's mean Q-head. Define the *compatibility error* as their difference:

$$e_{i,h}(\mathbf{s}) = \mathbb{E}_{a_i \sim \pi_i(\cdot|o_i)} q_{i,h}(o_i, a_i) - v_{i,h}(o_i), \qquad \varepsilon = \max_{i,h<H} \operatorname{span} e_{i,h}, \tag{13}$$

where $\operatorname{span} f = \max_{\mathbf{s}} f(\mathbf{s}) - \min_{\mathbf{s}} f(\mathbf{s})$. The bound below depends on variation across states, rather than on a constant offset between Q and V.

5.1 Exact discrepancy under general GAE

At finite trace length, the actor also depends on the gap between the centralized critic and its local V-head. Write $D_i = V - v_i$ and $\bar{e} = \sum_i \alpha_i e_i$. For a sequence of state functions f, define its current-policy continuation by $(\mathcal{R}_c f)_h = \mathbb{E}[\sum_{t=h}^{H-1} c^{t-h} f_t(\mathbf{S}_t) \mid \mathbf{S}_h]$, with value zero at H. Let $P_{i,h}^{a_i}$ average the next-state kernel over the other agents' actions, and let $P_{\pi,h}$ average over all actions. Define

$$F_i = \alpha_i \mathcal{R}_{\gamma\lambda}\big((1-\lambda)D_i + \lambda e_i\big) - \mathcal{R}_\gamma \bar{e}. \tag{14}$$

For every $\lambda \in [0,1]$, the exact actor discrepancy is

$$g_L^{(\lambda)} - \nabla J_{\widehat{R}} = \mathbb{E}\left[\sum_{h,i} \gamma^{h+1} z_{i,h}\big[(P_{i,h}^{A_{i,h}} - P_{\pi,h}) F_{i,h+1}\big](\mathbf{S}_h)\right]. \tag{15}$$

Equation (15) measures how the acting agent changes the expected future quantity F_i relative to the current policy. At training $\lambda = 0.95$, the local continuation contains $0.05 D_i + 0.95 e_i$, together with the mixed-compatibility correction in Equation (14). The identity permits shared actor parameters and arbitrary frozen heads. Appendix A.6 proves it; Appendix A.7 gives the finite-rollout endpoint correction.

5.2 Sharp control by future compatibility

For full returns, the critic–head continuation term disappears. To express the remaining error in a scale that accounts for the policy's parameterization, define the trajectory score and its Fisher matrix:

$$S = \nabla_\theta \log p_\theta(\tau) = \sum_{h,i} z_{i,h}, \qquad \mathcal{I}_\pi = \mathbb{E}[SS^\top]. \tag{16}$$

For $b \in \operatorname{range}(\mathcal{I}_\pi)$, let $\mathcal{I}_\pi^\dagger$ denote the Moore–Penrose pseudoinverse of the Fisher matrix in Equation (16), allowing singular $\mathcal{I}_\pi$. The norm $\|b\|_{\mathcal{I}_\pi^\dagger} = (b^\top \mathcal{I}_\pi^\dagger b)^{1/2}$ measures the largest error in first-order predicted gain over directions of unit trajectory-Fisher length.

Theorem 1 (Sharp local-credit bound)**.** *Under the preceding assumptions, full-return GAE ($\lambda = 1$) satisfies*

$$\boxed{\big\|g_L^{(1)} - \nabla J_{\widehat{R}}\big\|_{\mathcal{I}_\pi^\dagger} \le \frac{\varepsilon}{2} \sum_{t=1}^{H-1} \gamma^t.} \tag{17}$$

The discrepancy belongs to $\operatorname{range}(\mathcal{I}_\pi)$*. The constant* $1/2$ *is sharp uniformly over the allowed games and normalized mixers: for $H \ge 2$ and $\gamma > 0$, two agents attain equality with a nonnegative mixer, and strictly positive mixers approach equality.*

At $\lambda = 1$, Equation (14) reduces to $F_i = -\mathcal{R}_\gamma \sum_{j \ne i} \alpha_j e_j$. Thus the error comes from an actor's effect on other heads' future compatibility. The current-state terms cancel by conditional action independence; the shared critic acts as a baseline at $\lambda = 1$. The theorem bounds the local direction relative to the learned reward objective, not directly relative to terminal-outcome performance. It applies to full returns, whereas the reported PPO runs use $\lambda = 0.95$. Consequently, state-constant compatibility errors give exact agreement, even if the local Q/V values contain constant offsets. More precisely, the proof bounds the discrepancy by $\frac{1}{2}\sum_{t=1}^{H-1} \gamma^t \max_i \operatorname{span} e_{i,t}$. Only stages with varying compatibility contribute: a suffix with constant compatibility adds no error term. Appendix A.3 proves the bound and Appendix A.4 establishes sharpness.

5.3 AVAILABLE SIGNAL IN STRUCTURED COORDINATION

The preceding bound concerns credit accuracy, not the amount of local signal a task provides. We analyze the latter question in a restricted binary coordination class.

Consider independent binary decisions $A_i \sim \mathrm{Bernoulli}(p_i)$, followed by an action-free continuation and terminal outcome Y. Let Y be a nonconstant monotone AND/OR expression in which each decision occurs once: AND requires joint completion and OR allows alternative successful routes. Such an expression is called *read-once*. A minimal successful coalition is a minimal set of actors whose activation forces $Y = 1$. Define the ideal local values and their signal energy by

$$q_i^\star(a) = \mathbb{E}[Y \mid A_i = a], \qquad \mu = \mathbb{E}Y, \qquad \mathcal{C} = \sum_i \mathrm{Var}(q_i^\star(A_i)). \tag{18}$$

These are population quantities for the true outcome, distinct from fitted reward-model heads.

Theorem 2 (Sharp terminal-credit floor)**.** *In the read-once model above, suppose every minimal successful coalition has at most k actors, $0 < \rho \le 1/2$, $\rho \le p_i < 1$, and $0 < \mu \le 1/2$. Then*

$$\boxed{\mathcal{C} \ge a_k(\rho)(1-\mu)^2 \log \frac{1}{1-\mu}, \qquad a_k(\rho) = \frac{k\rho^{2k-1}(1-\rho)}{(1-\rho^k)^2 \log \frac{1}{1-\rho^k}}.} \tag{19}$$

Equality holds precisely for disjoint size-k successful coalitions, with every relevant probability equal to ρ.

Within this class, the floor depends on success probability, coalition width and exploration, with no additional team-size factor. It concerns the true outcome, not fitted reward heads. Appendix B proves the theorem and a single-cross-coalition extension; Appendix B.6 gives the calibrated transfer condition.

6 EXPERIMENTS

The experiments ask three questions: how well the complete learner performs with terminal-only feedback, whether the Q/V model and local actor readout each contribute, and whether the credit errors predicted by the analysis appear when team reward is held fixed. Benchmarks address the first two; frozen-model relay interventions address the third.

6.1 BENCHMARK SETUP AND COMPARISONS

SMACv2 covers nine scenarios across Protoss, Terran and Zerg, with shaped rewards removed and terminal $+1/-1$ win/loss feedback (Ellis et al., 2023). MAMuJoCo withholds per-step rewards and reveals the cumulative episode return at termination (de Witt et al., 2020). Its agent partitions are 3×1 for Hopper, 2×1 for Reacher, 6×1 for HalfCheetah and 2×1 for Swimmer, with one-hop observations and an episode limit of 1000.

Each method uses 10M steps and five seeds. Evaluation samples 32 stochastic-policy episodes every 50k steps; tables average the final five checkpoints per seed and report the seed mean and standard deviation. PPO uses $\gamma = 0.99$, $\lambda = 0.95$, clipping 0.2 and ten epochs. Architectures and budgets are in Appendix C.

SparseMAPPO and SparseHAPPO optimize terminal rewards directly. Online-IPL extracts policies from learned Q/V values by advantage-weighted behavior cloning. SL-MAPPO fits a direct reward head with the same terminal-outcome comparisons and uses global advantages. IMAP-GA uses the Q/V reward model with a global readout; IMAP-LA uses weighted local advantages and is the original IMAP-Rule configuration.

The SL-MAPPO/IMAP-GA comparison probes the Q/V construction against a direct reward head under the shared buffer, label-bearing update schedule and encoder design. The IMAP-GA/IMAP-LA comparison probes the actor readout. Each configuration collects its own on-policy data, so these are training-run comparisons rather than interventions on an identical fixed dataset; Appendix C specifies their scope.

Table 1: SMACv2 win rate (%). Mean ± standard deviation over five seeds after 10M environment steps. All methods receive only terminal win/loss feedback. IMAP-LA and IMAP-GA denote the reported local- and global-advantage configurations with the same Q/V reward-model design.

Scenario	SparseMAPPO	SparseHAPPO	Online-IPL	SL-MAPPO	IMAP-GA	IMAP-LA
protoss_5_vs_5	0.00 ± 0.00	11.61 ± 3.03	15.03 ± 3.80	26.97 ± 6.00	43.52 ± 2.21	$\mathbf{44.00 \pm 3.16}$
protoss_10_vs_10	0.00 ± 0.00	0.00 ± 0.00	9.35 ± 2.50	21.38 ± 3.58	31.71 ± 2.19	$\mathbf{33.64 \pm 2.78}$
protoss_10_vs_11	0.00 ± 0.00	0.00 ± 0.00	7.54 ± 2.33	15.67 ± 3.48	22.01 ± 3.75	$\mathbf{23.08 \pm 3.99}$
terran_5_vs_5	31.96 ± 3.26	35.46 ± 3.58	32.47 ± 4.42	38.11 ± 5.57	45.26 ± 2.15	$\mathbf{46.84 \pm 2.93}$
terran_10_vs_10	23.12 ± 4.24	23.47 ± 2.85	23.40 ± 2.66	28.22 ± 4.16	36.23 ± 3.26	$\mathbf{38.18 \pm 2.82}$
terran_10_vs_11	0.00 ± 0.00	0.00 ± 0.00	10.65 ± 1.78	14.72 ± 2.45	17.46 ± 4.25	$\mathbf{19.71 \pm 2.08}$
zerg_5_vs_5	18.45 ± 3.99	23.01 ± 3.92	28.23 ± 3.46	25.80 ± 3.74	38.48 ± 5.48	$\mathbf{42.14 \pm 3.44}$
zerg_10_vs_10	10.40 ± 2.54	13.26 ± 2.02	14.40 ± 2.26	20.98 ± 1.37	26.15 ± 1.51	$\mathbf{27.78 \pm 3.65}$
zerg_10_vs_11	4.63 ± 1.63	4.76 ± 1.60	4.78 ± 1.91	8.73 ± 2.72	14.42 ± 2.80	$\mathbf{17.24 \pm 1.77}$

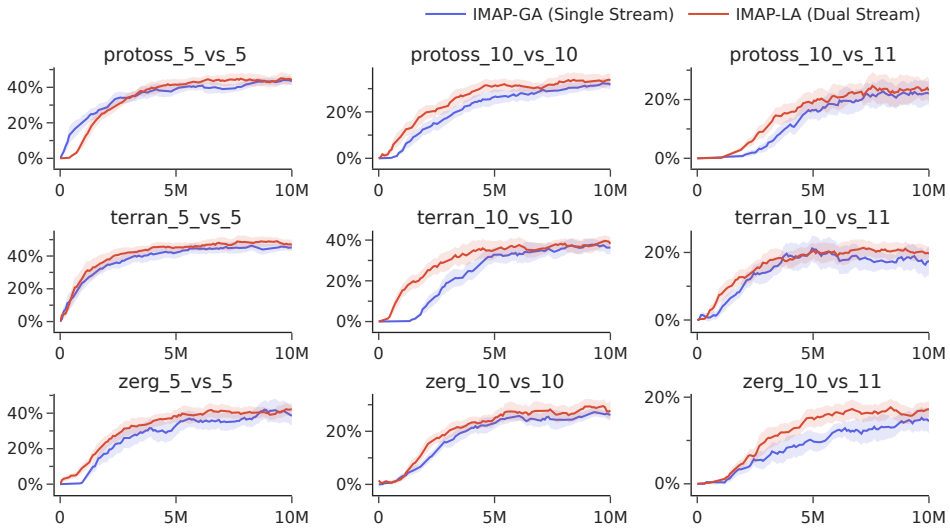

Figure 2: Global and weighted local actor readouts on SMACv2. Both use the same Q/V reward-model design and a global critic; curves show win rate over 10M steps.

6.2 Learning from terminal team feedback

IMAP-LA achieves the highest mean win rate on all nine SMACv2 scenarios. The Q/V construction improves the global-readout configuration over the direct reward head on every scenario: on Protoss 10 versus 10, the mean increases from 21.38% for SL-MAPPO to 31.71% for IMAP-GA. The local-actor configuration further improves all nine means, by 0.48–3.66 percentage points. Table 1 reports variability over seeds.

The same ordering of reported means holds across the four continuous-control tasks. On Hopper, the two global-readout configurations are close (296.45 and 296.48), while IMAP-LA reaches 361.22. On HalfCheetah, the mean rises from 2479.37 for SL-MAPPO to 3328.71 for IMAP-GA and 3853.58 for IMAP-LA. These comparisons support the combined design across discrete and continuous action spaces, while their seed variability and separately collected data limit causal interpretation of individual changes.

On identical held-out pairs, the Q/V model has better reported ranking accuracy and negative log-likelihood (NLL) than the direct head on all three measured tasks (Table 5). On Hopper, accuracy is 0.891 versus 0.812, and NLL is 0.517 versus 0.634. These three-seed descriptive means concern fitted team scores; they do not by themselves validate the local actor direction. The relay test addresses that direction.

SMACLite compares temporal-agent redistribution methods with cumulative native return revealed at termination (Michalski et al., 2023; Xiao et al., 2022; Chen et al., 2024; Kapoor et al., 2025).

Table 2: MAMuJoCo episode return. Mean ± standard deviation over five seeds after 10M environment steps. Each learner receives the cumulative environment return only at episode termination.

Method	Hopper-v2	Reacher-v2	HalfCheetah-v2	Swimmer-v2
SparseMAPPO	188.57 ± 18.34	-488.88 ± 76.07	1824.33 ± 181.46	16.29 ± 2.12
SparseHAPPO	215.41 ± 30.01	-497.58 ± 77.65	1926.69 ± 224.82	20.54 ± 2.89
Online-IPL	243.51 ± 9.67	-359.19 ± 17.31	1777.58 ± 341.44	23.10 ± 0.81
SL-MAPPO	296.45 ± 19.09	-358.15 ± 15.53	2479.37 ± 337.53	29.84 ± 1.21
IMAP-GA	296.48 ± 27.55	-316.83 ± 67.58	3328.71 ± 285.84	31.89 ± 1.18
IMAP-LA	$\mathbf{361.22 \pm 16.19}$	$\mathbf{-182.10 \pm 13.25}$	$\mathbf{3853.58 \pm 359.02}$	$\mathbf{33.98 \pm 1.04}$

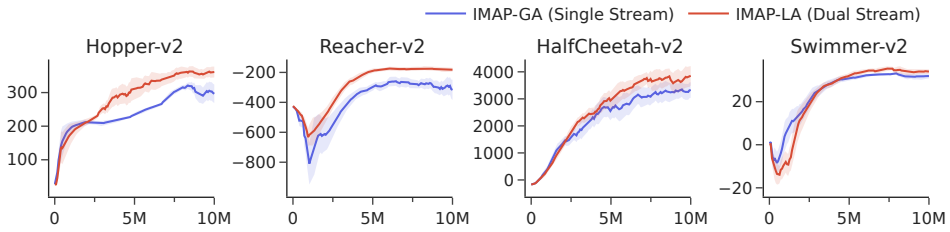

Figure 3: Global and weighted local actor readouts on MAMuJoCo, with the same Q/V reward-model design and global critic. IMAP-LA has higher final means after 10M steps.

IMAP-LA has the highest mean on two scenarios and exceeds SparseMAPPO and AREL throughout; TAR^2 has the highest mean on 3s5z (Appendix Table 3).

6.3 Which component errors change local credit?

We test the predicted error channels in a four-decision, two-agent terminal relay. At frozen checkpoints, Q-only and joint Q/V shifts preserve the mixed reward, policy and critic but change local credit through compatibility and critic–head continuation, respectively. At training $\lambda = 0.95$ and unit amplitude, the mean normalized actor-direction responses across ten models are 0.11496 and 0.00605. Their opposite zero-response endpoints and all 440 checkpoint bound checks meet the declared numerical tolerances (Appendices A.5–E; Appendix Figure 5). These tests diagnose frozen credit directions, not policy improvement; Appendix F separately examines the structured signal result.

The Q-only response vanishes at $\lambda = 0$, and the joint Q/V response at $\lambda = 1$. For the shared perturbations used here, their magnitudes at training λ differ by $\lambda/(1-\lambda) = 19$, as predicted by Equation (14). Thus an unchanged team score can support different local actor directions through the two continuation channels.

7 Conclusion

IMAP turns terminal outcomes into agent-specific online PPO updates. Comparisons between completed episodes fit shared Q/V heads; their residuals and normalized mixture give the centralized critic a team reward and decentralized actors weighted local advantages. With terminal-only feedback, IMAP-LA has the highest reported mean among the compared methods on nine SMACv2 and four MAMuJoCo tasks, and on two of three SMACLite scenarios. Direct-head and global-readout comparisons support the model and actor choices; frozen relay interventions confirm predicted credit responses at unchanged mixed reward.

An exact general-GAE identity and sharp full-return bound characterize local-credit error for the learned objective; a read-once theorem quantifies ideal task signal. Calibrating the fitted score to terminal outcomes and extending the analysis to finite-rollout clipped PPO and richer mixers are promising directions for future work.

Statement of Use of AI

The research questions, the main ideas, the methodology, and the experimental protocols of this paper were chosen by the authors, who divided the work among themselves, supervised every stage of the project, and made all decisions on claims and conclusions. Under the authors' direction, we used generative AI tools as assistants for the following tasks with required disclosure: refining the conceptual framing, stress-testing mathematical claims and proof arguments, and implementing and running the relay and numerical-verification experiments. In the theoretical component, generative AI was used in part as a "theory attacker": it assisted in implementing computational searches and large-scale simulations over extensive collections of generated examples in order to identify potential counterexamples, edge cases, or weaknesses in candidate theorem statements and proof arguments. These searches were used to help the authors examine and refine the theoretical claims. The failure of such searches to identify a counterexample does not constitute a proof, and no theorem in this paper is considered established on the basis of AI-assisted simulation, search, or numerical testing.

The authors independently verified all theorem statements and proofs, citations, experimental specifications, and AI-assisted text against their research artifacts. No experimental data or results were produced by generative AI tools themselves: every reported number comes from experiments that were actually executed, and translation and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools for literature discovery and for drafting and editing the manuscript, including checks of notation and of the consistency of reported numbers. The authors take full responsibility for every claim, proof, citation, and experimental result in this paper; no AI system is an author.

Reproducibility statement

Section 3 defines terminal feedback, and Section 4 and Algorithm 1 specify IMAP. The local-credit and structured-signal proofs are in Appendices A and B. Section 6 reports benchmark results; Appendix C gives training and evaluation settings (Table 4). The relay construction, checkpoint population, and numerical checks are specified in Appendices D–E; Appendix F describes the fitted-signal diagnostic. The supplementary material provides the source code for IMAP and the benchmark experiments, the relay and numerical-verification studies, and the figures, together with the recorded evaluation curves and the outputs of the numerical studies.

References

Jose A Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. RUDDER: Return decomposition for delayed rewards. *Advances in Neural Information Processing Systems*, 32, 2019.

The Viet Bui, Tien Anh Mai, and Thanh Hong Nguyen. O-MAPL: Offline multi-agent preference learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 5811–5841. PMLR, 2025. URL `https://proceedings.mlr.press/v267/bui25a.html`.

Sirui Chen, Zhaowei Zhang, Yaodong Yang, and Yali Du. STAS: spatial-temporal return decomposition for solving sparse rewards problems in multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17337–17345, 2024.

Christian Schroeder de Witt, Bei Peng, Pierre-Alexandre Kamienny, Philip Torr, Wendelin Böhmer, and Shimon Whiteson. Deep multi-agent reinforcement learning for decentralized continuous cooperative control. *arXiv preprint arXiv:2003.06709v1*, 2020.

Benjamin Ellis, Jonathan Cook, Skander Moalla, Mikayel Samvelyan, Mingfei Sun, Anuj Mahajan, Jakob Foerster, and Shimon Whiteson. SMACv2: An improved benchmark for cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 36: 37567–37593, 2023.

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

Divyansh Garg, Shuvam Chakraborty, Chris Cundy, Jiaming Song, and Stefano Ermon. IQ-Learn: Inverse soft-Q learning for imitation. *Advances in Neural Information Processing Systems*, 34: 4028–4039, 2021.

Divyansh Garg, Joey Hejna, Matthieu Geist, and Stefano Ermon. Extreme Q-learning: MaxEnt RL without entropy. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2301.02328.

Joey Hejna and Dorsa Sadigh. Inverse preference learning: Preference-based RL without a reward function. *Advances in Neural Information Processing Systems*, 36, 2023.

Xunhan Hu, Jian Zhao, Wengang Zhou, Ruili Feng, and Houqiang Li. DIFFER: Decomposing individual reward for fair experience replay in multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/edac78c3e300629acfe6cbe9ca88fb84-Abstract-Conference.html.

Jeewon Jeon, Woojun Kim, Whiyoung Jung, and Youngchul Sung. MASER: Multi-agent reinforcement learning with subgoals generated from experience replay buffer. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10041–10052. PMLR, 2022. URL https://proceedings.mlr.press/v162/jeon22a.html.

Sham M. Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, volume 14, 2001. URL https://proceedings.neurips.cc/paper/2001/hash/4b86abe48d358ecf194c56c69108433e-Abstract.html.

Aditya Kapoor, Kale ab Tessera, Mayank Baranwal, Harshad Khadilkar, Jan Peters, Stefano Albrecht, and Mingfei Sun. Redistributing rewards across time and agents for multi-agent reinforcement learning, 2025. URL https://arxiv.org/abs/2502.04864.

Landon Kraemer and Bikramjit Banerjee. Multi-agent reinforcement learning as a rehearsal for decentralized planning. *Neurocomputing*, 190:82–94, 2016.

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. *arXiv preprint arXiv:2109.11251*, 2021.

Boyin Liu, Zhiqiang Pu, Yi Pan, Jianqiang Yi, Yanyan Liang, and D. Zhang. Lazy agents: A new perspective on solving sparse reward problem in multi-agent reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21937–21950. PMLR, 2023. URL https://proceedings.mlr.press/v202/liu23ac.html.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

Adam Michalski, Filippos Christianos, and Stefano V. Albrecht. SMAClite: A lightweight environment for multi-agent reinforcement learning, 2023. URL https://arxiv.org/abs/2305.05566.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

Ryan O'Donnell and Li-Yang Tan. A composition theorem for the Fourier Entropy-Influence conjecture, 2013.

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. *The Journal of Machine Learning Research*, 21(1):7234–7284, 2020.

Zhizhou Ren, Ruihan Guo, Yuan Zhou, and Jian Peng. Learning long-term reward redistribution via randomized return decomposition. In *International Conference on Learning Representations*, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

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

Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. QPLEX: Duplex dueling multi-agent Q-learning. In *International Conference on Learning Representations*, 2021a.

Yihan Wang, Beining Han, Tonghan Wang, Heng Dong, and Chongjie Zhang. DOP: Off-policy multi-agent decomposed policy gradients. In *International Conference on Learning Representations*, 2021b. URL https://arxiv.org/abs/2007.12322.

Baicen Xiao, Bhaskar Ramasubramanian, and Radha Poovendran. Agent-temporal attention for reward redistribution in episodic multi-agent reinforcement learning. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 1391–1399, 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. *Advances in Neural Information Processing Systems*, 35:24611–24624, 2022.

Zijian Zhao and Sen Li. Aggregate in the advantage, not the ratio: A canonical-form analysis of cooperative multi-agent policy optimization. arXiv:2607.17924, 2026. URL https://arxiv.org/abs/2607.17924.

A PROOFS OF THE LOCAL-CREDIT RESULTS

A.1 SETTING AND QUANTITIES

Consider a finite-horizon Markov game with stages $h = 0, \dots, H-1$, $H \geq 1$, and a sufficient controlled information state $\mathbf{S}_h$ containing all actor conditioning information. The joint history is one admissible choice. Agent i observes $o_{i,h}(\mathbf{S}_h)$ and samples its action from $\pi_{i,\theta}(\cdot \mid o_{i,h}(\mathbf{S}_h))$. Conditional on $\mathbf{S}_h$, actions are independent throughout a neighborhood of the current parameter:

$$\pi_\theta(\mathbf{a} \mid \mathbf{s}) = \prod_{i=1}^{n} \pi_{i,\theta}(a_i \mid o_{i,h}(\mathbf{s})).$$

All agents may share the parameter vector $\theta \in \mathbb{R}^d$. The initial distribution, transition kernels and observation maps are independent of θ. We first give the proof for finite state and action sets, with differentiable policies positive on fixed feasible action sets in a neighborhood of the current parameter. A forced no-op is a singleton action set and has zero score. The extension to general state and action spaces is stated at the end.

Fix $\gamma, \lambda \in [0, 1]$. Freeze real-valued local functions $q_{i,h}$, $v_{i,h}$, a shared critic V_h, and scalar weights $\alpha_i \geq 0$ with $\sum_i \alpha_i = 1$. Weights do not vary with the stage, state or action. Each $q_{i,h}(o_i, a_i)$ depends on its own action only. Terminal values satisfy $v_{i,H} = V_H = 0$. All expectations and scores below are at the current policy, and model quantities remain frozen when differentiating the learned return. The reward and local residual are

$$r_{i,h} = q_{i,h}(o_{i,h}, A_{i,h}) - \gamma v_{i,h+1}(o_{i,h+1}), \tag{20}$$

$$\widehat{R}_h = \sum_i \alpha_i r_{i,h}, \qquad \delta_{i,h} = r_{i,h} + \gamma V_{h+1}(\mathbf{S}_{h+1}) - V_h(\mathbf{S}_h). \tag{21}$$

We suppress observation arguments when viewing local functions as functions of $\mathbf{s}$. Define complete-episode advantages and learned returns by

$$\widehat{A}^L_{i,h} = \sum_{t=h}^{H-1} (\gamma\lambda)^{t-h} \delta_{i,t}, \qquad \widehat{G}_h = \sum_{t=h}^{H-1} \gamma^{t-h} \widehat{R}_t, \qquad J_{\widehat{R}}(\theta) = \mathbb{E}_\theta \widehat{G}_0.$$

Let $\widehat{A}^G_h$ use the same GAE sum with residual $\widehat{R}_h + \gamma V_{h+1} - V_h$. Normalization of the mixer gives the pathwise identity

$$\widehat{A}^G_h = \sum_i \alpha_i \widehat{A}^L_{i,h}. \tag{22}$$

Indeed, mixing each local residual gives the global residual, and the constant weights commute with the common finite GAE sum.

For general GAE, use the unnormalized discounted time weights in

$$z_{i,h} = \nabla_\theta \log \pi_{i,\theta}(A_{i,h} \mid o_{i,h}(\mathbf{S}_h)), \qquad g^{(\lambda)}_L = \mathbb{E} \sum_{h=0}^{H-1} \sum_i \gamma^h \alpha_i z_{i,h} \widehat{A}^L_{i,h}.$$

The derivative acts on the policy only.

Write

$$e_{i,h} = \mathbb{E}_{a_i \sim \pi_i} q_{i,h}(o_i, a_i) - v_{i,h}, \qquad D_{i,h} = V_h - v_{i,h}, \qquad \bar{e}_h = \sum_i \alpha_i e_{i,h}, \qquad \bar{v}_h = \sum_i \alpha_i v_{i,h}.$$

Let $P^{\mathbf{a}}_h$ be the next-state kernel conditional on the joint action, $P^{a_i}_{i,h}$ average the other agents' actions, and $P_{\pi,h}$ average all actions under the current policy. For a state function f, $(Pf)(\mathbf{s})$ denotes its expectation under the relevant next-state kernel. Define the finite continuation operator

$$(\mathcal{R}_c f)_h(\mathbf{s}) = \sum_{\ell=0}^{H-h-1} c^\ell \mathbb{E}[f_{h+\ell}(\mathbf{S}_{h+\ell}) \mid \mathbf{S}_h = \mathbf{s}], \qquad (\mathcal{R}_c f)_H = 0. \tag{23}$$

An empty sum is zero, and a zeroth power is one, including when its base is zero. Thus $\mathcal{R}_c f$ is the unique backward solution of $u_h = f_h + c P_{\pi,h} u_{h+1}$ with $u_H = 0$.

A.2 CONSISTENCY OF THE GLOBAL AND LOCAL ACTOR COEFFICIENTS

The coefficient α_i in the local actor follows from the same mixture that defines the learned team objective. The following calculation records this relationship; it does not assert that the two stochastic updates or their finite PPO trajectories coincide. Use the assumptions and frozen quantities of Appendix A.1, and define the unit-prefactor global vector

$$g_G^{(\lambda)} = \mathbb{E} \sum_{h=0}^{H-1} \sum_i \gamma^h z_{i,h} \widehat{A}_h^G. \tag{24}$$

The expectations are at the old policy. The advantages are fixed targets, without a nonlinear or sample-dependent advantage transformation.

One-stage equality. For $H = 1$, every $\lambda \in [0, 1]$ gives

$$g_G^{(\lambda)} = g_L^{(\lambda)} = \nabla J_{\widehat{R}} = \sum_i \alpha_i \mathbb{E}[z_{i,0} q_{i,0}]. \tag{25}$$

In contrast, applying an additional actor prefactor α_i to the global advantage gives

$$\widetilde{g}_G := \mathbb{E} \sum_i \alpha_i z_{i,0} \widehat{A}_0^G = \sum_i \alpha_i^2 \mathbb{E}[z_{i,0} q_{i,0}], \tag{26}$$

which generally changes the reward gradient.

Proof. With zero terminal values, $r_{i,0} = q_{i,0}$ and

$$\widehat{A}_{i,0}^L = q_{i,0} - V_0(\mathbf{S}_0), \qquad \widehat{A}_0^G = \sum_j \alpha_j q_{j,0} - V_0(\mathbf{S}_0).$$

Policy normalization implies $\mathbb{E}[z_{i,0} \mid \mathbf{S}_0] = 0$. For $j \neq i$, conditional action independence and the own-action restriction on q_j give

$$\mathbb{E}[z_{i,0} q_{j,0} \mid \mathbf{S}_0] = \mathbb{E}[z_{i,0} \mid \mathbf{S}_0] \, \mathbb{E}[q_{j,0} \mid \mathbf{S}_0] = 0.$$

The shared critic is also a baseline, regardless of its accuracy. Substitution into the two actor vectors leaves precisely the right-hand side of Equation (25). Since the initial law and frozen heads do not depend on actor parameters, likelihood-ratio differentiation of $J_{\widehat{R}} = \mathbb{E} \sum_j \alpha_j q_{j,0}$ gives the same expression. Inserting the extra prefactor in the global vector gives Equation (26). These conditional cancellations hold coordinatewise even when actor parameters are shared. □

Full-return sequential extension. For $\lambda = 1$ on a complete episode, $\widehat{A}_h^G = \widehat{G}_h - V_h(\mathbf{S}_h)$, where $\widehat{G}_h = \sum_{t=h}^{H-1} \gamma^{t-h} \widehat{R}_t$. The policy-gradient identity and baseline cancellation therefore give $g_G^{(1)} = \nabla J_{\widehat{R}}$ for any frozen critic. Theorem 1 and its stagewise refinement in Appendix A.3 yield

$$\|g_L^{(1)} - g_G^{(1)}\|_{\mathcal{I}_\pi^\dagger} \leq \frac{1}{2} \sum_{t=1}^{H-1} \gamma^t \max_j \operatorname{span} e_{j,t}. \tag{27}$$

In particular, the two expected reward gradients agree when every $e_{j,t}$ is constant across stage-t states for $t \geq 1$; those constants need not be zero or equal across components. The discounted time weights γ^h in both vectors are those of $J_{\widehat{R}} = \mathbb{E} \sum_h \gamma^h \widehat{R}_h$.

Entropy and implementation scope. Both actor objectives use the same entropy coefficient outside the reward prefactor. At a common old policy and with the same state-sampling measure and reduction, their entropy derivatives are identical. Hence, whenever the expected reward vectors above agree, adding the same entropy term preserves that equality. This observation does not compare the random reward-gradient magnitudes or their subsequent balance with entropy.

The statements concern expected derivatives at the old policy, where positive-width PPO clipping is locally inactive. They do not establish equal minibatch gradients, variance reduction, or equality after finite steps and repeated clipped updates. For $H > 1$ and $\lambda < 1$, the continuation terms in

Appendix A.6 must be retained; Equation (27) does not apply directly to the training setting $\lambda = 0.95$. Rollout truncation requires the endpoint terms of Appendix A.7. A fixed common rescaling of both reward vectors rescales these equalities and bounds, whereas empirical agent-specific or sample-dependent advantage normalization and different time reductions require separate analysis. Finally, an additional-weight global variant remains a useful empirical control, but Equation (26) distinguishes it from the global gradient of the original mixed objective.

A.3 Proof of Theorem 1

Proof. Write $L_H = \sum_{t=1}^{H-1} \gamma^t$, with $L_1 = 0$. Let $G_j = \sum_{t=0}^{H-1} \gamma^t r_{j,t}$, $S_j = \sum_h z_{j,h}$ and $S_{-j} = \sum_{h,i\neq j} z_{i,h}$. At $\lambda = 1$ and $V_H = 0$, the shared-critic terms telescope:

$$\widehat{A}^L_{i,h} = \sum_{t=h}^{H-1} \gamma^{t-h} r_{i,t} - V_h(\mathbf{S}_h).$$

Policy normalization gives zero conditional mean for $z_{i,h}$ given the pre-action history. Hence its product with $V_h(\mathbf{S}_h)$ has zero expectation. For $t < h$, $r_{i,t}$ is measurable before the stage-h action, so $\mathbb{E}[z_{i,h} r_{i,t}] = 0$. We may therefore replace the discounted return-to-go in the actor sum by the complete local return. Likelihood-ratio differentiation of the finite trajectory law gives the team-return gradient:

$$g^{(1)}_L = \sum_j \alpha_j \mathbb{E}[S_j G_j], \qquad \nabla J_{\widehat{R}} = \sum_j \alpha_j \mathbb{E}[S G_j],$$

and their difference is $-\sum_j \alpha_j \mathbb{E}[S_{-j} G_j]$. This cancellation uses a frozen critic; its accuracy is not required.

The Q/V reward telescopes pathwise:

$$G_j = v_{j,0}(\mathbf{S}_0) + \sum_{t=0}^{H-1} \gamma^t \big(q_{j,t}(o_{j,t}, A_{j,t}) - v_{j,t}(o_{j,t})\big).$$

To compute its product with a score $z_{i,h}$ for $i \neq j$, consider three cases. If $h < t$, condition on the full pre-action history at t: the score is already measurable and the bracket averages to $e_{j,t}(\mathbf{S}_t)$. If $h = t$, conditional action independence and zero mean of the own-agent score make the product zero; the product with $e_{j,t}(\mathbf{S}_t)$ is also zero. If $h > t$, the bracket is measurable before the later action and the later score has conditional mean zero. The same holds for $e_{j,t}(\mathbf{S}_t)$. The initial value and $t = 0$ compatibility terms also have zero product with S_{-j}. Consequently

$$b := g^{(1)}_L - \nabla J_{\widehat{R}} = -\sum_j \alpha_j \mathbb{E}[S_{-j} T_j], \qquad T_j = \sum_{t=1}^{H-1} \gamma^t e_{j,t}(\mathbf{S}_t). \tag{28}$$

Here e_j is the function evaluated at the current policy and then fixed. No differentiation through its definition is taken in Equation (28).

Distinct score increments are orthogonal in second moment. Across stages this follows by conditioning on the later pre-action history. Within a stage it follows from conditional action independence and zero conditional means. Therefore, even with shared parameter coordinates,

$$\mathbb{E}[S_{-j} S_{-j}^\top] = \sum_{h,i\neq j} \mathbb{E}[z_{i,h} z_{i,h}^\top] \preceq \sum_{h,i} \mathbb{E}[z_{i,h} z_{i,h}^\top] = \mathcal{I}_\pi. \tag{29}$$

The range of T_j has length at most $L_H \varepsilon$. A random variable in an interval of length r has variance at most $r^2/4$: subtract its midpoint and use that the mean minimizes squared error over constants. Since $\mathbb{E} S_{-j} = 0$, Cauchy–Schwarz gives, for every vector u,

$$\begin{aligned} |u^\top b| &\le \sum_j \alpha_j \sqrt{\mathbb{E}[(u^\top S_{-j})^2] \operatorname{Var}(T_j)} \\ &\le \sqrt{u^\top \mathcal{I}_\pi u}\, \frac{L_H \varepsilon}{2}. \end{aligned}$$

In particular, every vector in the kernel of $\mathcal{I}_\pi$ is orthogonal to b, proving the range assertion. Taking the supremum over $u^\top \mathcal{I}_\pi u \le 1$ proves Equation (17). □

Exact agreement and effective duration. If each compatibility error is constant across states at each stage, $\varepsilon = 0$ and the two expected gradients agree. More finely, the same proof gives

$$\|b\|_{\mathcal{I}_\pi^\dagger} \le \frac{1}{2} \sum_{t=1}^{H-1} \gamma^t \max_j \operatorname{span} e_{j,t}.$$

Only stages with nonconstant compatibility error contribute to this certificate. An inert suffix with state-constant errors adds no term. There is no factor depending on the agent count, parameter dimension or a transition-contraction coefficient. This describes the total expected actor vector, including all stages and shared parameters.

A.4 Sharpness of the constant

Fix $H \ge 2$, $\gamma > 0$ and $\varepsilon > 0$. Agent 1 makes a single binary decision at stage zero, with probability $p = \operatorname{sigmoid}(\theta)$ of the positive branch. The next state records the branch, and all remaining actions of both agents are forced no-ops. Agent 2 observes the branch. Take all values and the shared critic to be zero, $q_1 = 0$, $q_{2,0} = 0$, and at every remaining stage set

$$q_{2,t} = e_{2,t} = \begin{cases} \varepsilon/2, & \text{positive branch,} \\ -\varepsilon/2, & \text{negative branch.} \end{cases}$$

With $\alpha = (0, 1)$, the local actor vector is zero. The team return is $J_{\widehat{R}} = L_H \varepsilon (p - 1/2)$. At $p = 1/2$,

$$|\nabla J_{\widehat{R}}| = L_H \varepsilon / 4, \qquad \mathcal{I}_\pi = p(1-p) = 1/4,$$

so Equation (17) is an equality. With the strictly positive mixer $(\eta, 1-\eta)$, the left side is $(1-\eta) L_H \varepsilon / 2$, approaching the bound as $\eta \downarrow 0$. Thus no smaller constant works uniformly over strictly positive normalized mixers either. Sharpness is uniform over mixers, not attainment for every fixed positive mixer.

The environmental reward can be restricted to the terminal binary branch outcome. The learned return is a positive affine transform of that outcome when $\gamma > 0$.

A.5 Pathwise local GAE and rollout boundaries

For a fixed sampled rollout from stage h to its boundary H, write $D_{i,t} = V_t - v_{i,t}$ and $c = \gamma\lambda$. The endpoint values are the declared frozen bootstrap values, which can be nonzero at a rollout cutoff. Define finite GAE with the same rewards, critic and endpoint masks as in Equation (7). Then

$$\begin{aligned} \widehat{A}^L_{i,h} = & \sum_{t=h}^{H-1} c^{t-h}(q_{i,t} - v_{i,t}) - D_{i,h} \\ & + \gamma(1-\lambda) \sum_{t=h}^{H-2} c^{t-h} D_{i,t+1} + \gamma c^{H-h-1} D_{i,H}. \end{aligned} \tag{30}$$

All identities are samplewise; no expectation, action-factorization or calibration assumption is needed. Empty sums equal zero, and powers with exponent zero equal one, including at $c = 0$. At a true terminal boundary, $v_{i,H} = V_H = 0$, giving

$$\begin{aligned} \widehat{A}^L_{i,h} = & \sum_{t=h}^{H-1} (\gamma\lambda)^{t-h}(q_{i,t} - v_{i,t}) - D_{i,h} \\ & + \gamma(1-\lambda) \sum_{t=h}^{H-2} (\gamma\lambda)^{t-h} D_{i,t+1}. \end{aligned} \tag{31}$$

Proof. The implicit reward and shared critic residual give

$$\delta^L_{i,t} = q_{i,t} - \gamma v_{i,t+1} + \gamma V_{t+1} - V_t = (q_{i,t} - v_{i,t}) - D_{i,t} + \gamma D_{i,t+1}.$$

Multiply by c^{t-h} and sum from $t = h$ through $H - 1$. The coefficient of $D_{i,h}$ is -1. For any interior index $u \in \{h + 1, \dots, H - 1\}$, its coefficient is

$$\gamma c^{u-h-1} - c^{u-h} = \gamma(1 - \lambda)c^{u-h-1}.$$

The endpoint coefficient is γc^{H-h-1}. Collecting these terms proves Equation (30), including $\lambda = 0$ and $\gamma = 0$, without division by c. □

For a trajectory that terminates before a batch rollout ends, apply the identity up to its true terminal boundary, or equivalently use an absorbing zero-valued suffix and the same terminal masks throughout. A truncation with bootstrapping retains the displayed endpoint term; treating it as a true terminal transition would change the readout. The population-gradient correction for the declared bootstrapped objective is given separately in Appendix A.7.

At $\lambda = 1$ the interior D terms vanish, so the difference from the discounted sum of local Q/V gaps consists only of the current baseline and, if present, the boundary term. At $\lambda < 1$, the coefficient $(1 - \lambda)$ is exact, but the full accumulated correction need not be small without a bound on future D and the effective rollout length. The current-state term has zero expected old-policy score product; it can still affect finite-sample variance and later clipped PPO updates. Thus the identity does not assert automatic variance reduction.

Fixed-score responses. Fix locally measurable state potentials $\phi_{i,h}$ with $\sum_i \alpha_i \phi_{i,h} = 0$ pointwise and $\phi_{i,H} = 0$. They and the mixer are frozen during actor differentiation. Under a joint Q/V shift of amplitude ζ, the local Q/V gap is unchanged and D_i changes to $D_i - \zeta\phi_i$. Under a Q-only shift, the local Q/V gap changes by $\zeta\phi_i$ and D_i is unchanged. The mixed Q and V, hence mixed rewards and global critic targets, are unchanged in both cases.

Set $c = \gamma\lambda$ and define

$$K_\phi = \mathbb{E} \sum_{h,i} \gamma^h \alpha_i z_{i,h} \sum_{t=h}^{H-2} c^{t-h} \phi_{i,t+1}(\mathbf{S}_{t+1}).$$

The current-state potential has zero expected score product: $\mathbb{E}[z_{i,h}\phi_{i,h}(\mathbf{S}_h)] = 0$ by conditioning on the pre-action state. Applying Equation (31) to the joint shift therefore gives

$$\Delta g_{QV} = -\zeta\gamma(1 - \lambda)K_\phi.$$

For the Q-only shift, direct summation of its GAE change gives $\zeta \sum_{t=h}^{H-1} c^{t-h}\phi_{i,t}(\mathbf{S}_t)$. Its current-state term cancels in expectation; reindexing the remainder gives

$$\Delta g_Q = \zeta\gamma\lambda K_\phi.$$

Combining the response formulas yields

$$(1 - \lambda)\Delta g_Q + \lambda\Delta g_{QV} = 0. \tag{32}$$

The relation includes the endpoints $\lambda = 0, 1$ without division. They permit shared actor parameters. The response identity also holds with a common frozen pre-action-measurable time weight in place of γ^h, since the same current-state cancellation and reindexing apply within each stage's advantage. It concerns expected reward-gradients at the old policy, before a nonlinear advantage transform or subsequent clipped PPO updates. A common fixed scalar normalization preserves the relation.

For the Fisher responses in Figure 5, the policy, Fisher matrix, potential amplitude and original-head scale are identical between the two interventions. Homogeneity of the norm thus gives $\|\Delta g_Q\| = 19\|\Delta g_{QV}\|$ at $\lambda = 0.95$, with the corresponding endpoint zeros. The controlled ratio describes sensitivity at fixed mixed reward; it is not a comparison of policy returns.

A.6 Exact continuation identity for general GAE

Lemma 3 (Continuation discrepancy)**.** *Define*

$$F_i = \alpha_i \mathcal{R}_{\gamma\lambda}\big((1 - \lambda)D_i + \lambda e_i\big) - \mathcal{R}_\gamma \bar{e}. \tag{33}$$

Under the preceding assumptions,

$$\boxed{g_L^{(\lambda)} - \nabla J_{\widehat{R}} = \mathbb{E}\sum_{h=0}^{H-1}\sum_i \gamma^{h+1} z_{i,h}\big[(P_{i,h}^{A_{i,h}} - P_{\pi,h})F_{i,h+1}\big](\mathbf{S}_h).} \tag{34}$$

This holds with shared policy parameters and with zero mixer weights.

Proof. First establish the learned-return policy gradient. Differentiating the finite trajectory sum, only policy factors depend on θ, so

$$\nabla J_{\widehat{R}} = \mathbb{E}\left[\widehat{G}_0 \sum_{h,i} z_{i,h}\right].$$

Let $\mathcal{H}_h$ be the full pre-action history. Policy normalization and conditioning only on $\mathbf{S}_h$ give $\mathbb{E}[z_{i,h} \mid \mathcal{H}_h] = \mathbb{E}[z_{i,h} \mid \mathbf{S}_h] = 0$. Rewards preceding stage h are $\mathcal{H}_h$-measurable, so their products with that stage's score have zero expectation. The remaining terms are

$$\nabla J_{\widehat{R}} = \mathbb{E}\sum_{h,i} \gamma^h z_{i,h}\widehat{G}_h. \tag{35}$$

This derivation sums each agent's score even when the scores use the same parameter coordinates.

We next compute the two conditional continuation values. Let $M_{i,h}(\mathbf{s}) = \mathbb{E}[\widehat{A}_{i,h}^L \mid \mathbf{S}_h = \mathbf{s}]$ and $M_{i,H} = 0$. Equation (21) is $\delta_{i,h} = q_{i,h} - V_h + \gamma D_{i,h+1}$. Conditioning the GAE recursion on the current state gives

$$M_{i,h} = e_{i,h} - D_{i,h} + \gamma P_{\pi,h} D_{i,h+1} + \gamma\lambda P_{\pi,h} M_{i,h+1}.$$

Set $W_{i,h} = D_{i,h} + \lambda M_{i,h}$. Multiplying the preceding recursion by λ and adding $D_{i,h}$ gives

$$W_{i,h} = (1-\lambda)D_{i,h} + \lambda e_{i,h} + \gamma\lambda P_{\pi,h} W_{i,h+1}, \qquad W_{i,H} = 0.$$

There is no division by λ. By backward substitution,

$$W_i = \mathcal{R}_{\gamma\lambda}\big((1-\lambda)D_i + \lambda e_i\big). \tag{36}$$

Consequently, conditional on the current joint action,

$$\mathbb{E}[\widehat{A}_{i,h}^L \mid \mathbf{s}, \mathbf{a}] = q_{i,h}(o_i, a_i) - V_h(\mathbf{s}) + \gamma(P_h^{\mathbf{a}} W_{i,h+1})(\mathbf{s}). \tag{37}$$

For the team return, let $U_h(\mathbf{s}) = \mathbb{E}[\widehat{G}_h \mid \mathbf{S}_h = \mathbf{s}]$. Its Bellman recursion is

$$U_h = \bar{v}_h + \bar{e}_h + \gamma P_{\pi,h}(U_{h+1} - \bar{v}_{h+1}), \qquad U_H = \bar{v}_H = 0.$$

Thus $U - \bar{v} = \mathcal{R}_\gamma \bar{e}$, and, writing $\bar{q}_h(\mathbf{s}, \mathbf{a}) = \sum_j \alpha_j q_{j,h}(o_j, a_j)$,

$$\mathbb{E}[\widehat{G}_h \mid \mathbf{s}, \mathbf{a}] = \bar{q}_h(\mathbf{s}, \mathbf{a}) + \gamma\big(P_h^{\mathbf{a}}(\mathcal{R}_\gamma \bar{e})_{h+1}\big)(\mathbf{s}). \tag{38}$$

Subtract Equation (38) from α_i times Equation (37). The result is

$$-\sum_{j\neq i} \alpha_j q_{j,h}(o_j, a_j) - \alpha_i V_h(\mathbf{s}) + \gamma(P_h^{\mathbf{a}} F_{i,h+1})(\mathbf{s}). \tag{39}$$

The shared-critic term is independent of a_i. For $j \neq i$, conditional factorization and the own-action dependence of q_j give

$$\mathbb{E}[z_{i,h} q_{j,h}(o_j, A_{j,h}) \mid \mathbf{S}_h = \mathbf{s}] = \mathbb{E}[z_{i,h} \mid \mathbf{S}_h = \mathbf{s}]\,\mathbb{E}[q_{j,h}(o_j, A_{j,h}) \mid \mathbf{S}_h = \mathbf{s}] = 0.$$

This cancellation remains valid with shared parameters: it uses conditional action independence, not disjoint parameter coordinates. Combining Equation (35) with Equation (39) therefore yields

$$g_L^{(\lambda)} - \nabla J_{\widehat{R}} = \mathbb{E}\sum_{h,i} \gamma^{h+1} z_{i,h}(P_h^{\mathbf{A}_h} F_{i,h+1})(\mathbf{S}_h).$$

Averaging other agents' actions replaces $P_h^{\mathbf{A}_h}$ by $P_{i,h}^{A_{i,h}}$. Finally, conditional on $\mathbf{S}_h$, $(P_{\pi,h}F_{i,h+1})(\mathbf{S}_h)$ does not depend on the sampled $A_{i,h}$, so its product with $z_{i,h}$ has zero conditional mean. Subtracting it proves Equation (34). □

The retained GAE setting. The reference training value $\lambda = 0.95$ gives

$$F_i^{(0.95)} = \alpha_i \mathcal{R}_{0.95\gamma}(0.05 D_i + 0.95 e_i) - \mathcal{R}_\gamma \bar{e}. \tag{40}$$

Substitution in Equation (34) gives the exact discrepancy of $g_L^{(0.95)}$ under the stated time weights and episode boundary. This is the continuation used when evaluating the training readout. At $\lambda = 1$, the same formula reduces to $F_i = -\mathcal{R}_\gamma \sum_{j \neq i} \alpha_j e_j$, which underlies Equation (28).

A.7 Finite-rollout endpoints

Fix a rollout boundary H and a frozen terminal target $b_H(\mathbf{S}_H)$. The comparison objective is the bootstrapped return $J_{\widehat{R},b} = \mathbb{E}[\sum_{t=0}^{H-1} \gamma^t \widehat{R}_t + \gamma^H b_H(\mathbf{S}_H)]$. The local GAE still sums the H observed residuals, with the declared endpoint values $v_{i,H}$ and V_H. Let $(P_{\pi,h:H} f)(\mathbf{s}) = \mathbb{E}[f(\mathbf{S}_H) \mid \mathbf{S}_h = \mathbf{s}]$ and $\bar{v}_H = \sum_j \alpha_j v_{j,H}$. Replace F_i in Lemma 3 by

$$\begin{aligned} \widetilde{F}_{i,h} = F_{i,h} &+ \alpha_i (\gamma\lambda)^{H-h} P_{\pi,h:H}(V_H - v_{i,H}) \\ &+ \gamma^{H-h} P_{\pi,h:H}(\bar{v}_H - b_H). \end{aligned} \tag{41}$$

Then Equation (34) holds with $\widetilde{F}_i$ and $\nabla J_{\widehat{R},b}$. To verify it, the local recursion now has $W_{i,H} = V_H - v_{i,H}$, and the team recursion has $(U - \bar{v})_H = b_H - \bar{v}_H$. Backward substitution adds exactly the two terms in Equation (41); the same score cancellation completes the identity. For a critic bootstrap use $b_H = V_H$; for the truncated reward sum use $b_H = 0$. Zero terminal values recover the complete-episode result. The bootstrap objective is explicitly the declared frozen-endpoint target.

A.8 Technical bounds for general GAE

Throughout this subsection, $\|\cdot\|$ denotes the Euclidean norm. For a real-valued function on the stage-h state set, write $\operatorname{span} f = \max_{\mathbf{s}} f(\mathbf{s}) - \min_{\mathbf{s}} f(\mathbf{s})$, and let $\mathrm{TV}(p, q) = \frac{1}{2} \sum_{\mathbf{s}} |p(\mathbf{s}) - q(\mathbf{s})|$. Define

$$\eta_{i,h}(\mathbf{s}, a_i) = \mathrm{TV}(P_{i,h}^{a_i}(\cdot \mid \mathbf{s}), P_{\pi,h}(\cdot \mid \mathbf{s})), \tag{42}$$

$$\mathcal{C}_\pi = \mathbb{E} \sum_{h,i} \gamma^{h+1} \|z_{i,h}\| \eta_{i,h}(\mathbf{S}_h, A_{i,h}). \tag{43}$$

For any probability vectors p, q and function f,

$$|(p - q) f| \leq \mathrm{TV}(p, q) \operatorname{span} f. \tag{44}$$

To see the constant exactly, subtract the midpoint of the range of f. The sum of absolute signed masses is $2\,\mathrm{TV}(p, q)$ and the centered function has supremum norm at most $\operatorname{span} f / 2$.

Corollary 4 (Action influence and continuation variation)**.** *The continuation discrepancy satisfies*

$$\|g_L^{(\lambda)} - \nabla J_{\widehat{R}}\| \leq \mathbb{E} \sum_{h,i} \gamma^{h+1} \|z_{i,h}\| \eta_{i,h}(\mathbf{S}_h, A_{i,h}) \operatorname{span} F_{i,h+1} \tag{45}$$

$$\leq \mathcal{C}_\pi \max_{i,\, 1 \leq h \leq H} \operatorname{span} F_{i,h}. \tag{46}$$

Proof. Apply the triangle inequality to Equation (34), then apply Equation (44) inside the expectation. All coefficients are nonnegative. Taking out the maximum continuation span gives the second inequality. The terminal function $F_{i,H}$ is zero, so the maximum is well-defined also at $H = 1$. □

Technical refinement in local Q/V quantities. The compact bound can be evaluated through stagewise compatibility and value variation. Let $\kappa \in [0, 1]$ satisfy, at every stage,

$$\sup_{\mathbf{s}, \mathbf{s}'} \mathrm{TV}(P_{\pi,h}(\cdot \mid \mathbf{s}), P_{\pi,h}(\cdot \mid \mathbf{s}')) \leq \kappa.$$

Set

$$\varepsilon_Q = \max_{h<H,i} \operatorname{span} e_{i,h}, \qquad \varepsilon_V = \max_{h<H,i} \operatorname{span} D_{i,h}, \qquad K_c(H) = \sum_{\ell=0}^{H-2} c^\ell,$$

where $K_c(1) = 0$.

Lemma 5 (Compatibility and GAE bound)**.** *Under the preceding assumptions,*

$$\|g_L^{(\lambda)} - \nabla J_{\widehat{R}}\| \le \mathcal{C}_\pi \big[K_{\gamma\lambda\kappa}(H)((1-\lambda)\varepsilon_V + \lambda\varepsilon_Q) + K_{\gamma\kappa}(H)\varepsilon_Q\big]. \tag{47}$$

Proof. Equation (44), applied to two rows of $P_{\pi,h}$, implies $\operatorname{span}(P_{\pi,h} f) \le \kappa \operatorname{span} f$. A product of ℓ such kernels contracts spans by at most κ^ℓ, by repeated application. Hence, for $c \ge 0$ and $h < H$,

$$\operatorname{span}(\mathcal{R}_c f)_h \le \sum_{\ell=0}^{H-h-1} (c\kappa)^\ell \operatorname{span} f_{h+\ell}. \tag{48}$$

Nonnegative normalized weights give $\operatorname{span} \bar{e}_h \le \sum_j \alpha_j \operatorname{span} e_{j,h} \le \varepsilon_Q$. Likewise,

$$\operatorname{span}\big((1-\lambda)D_{i,h} + \lambda e_{i,h}\big) \le (1-\lambda)\varepsilon_V + \lambda\varepsilon_Q.$$

Using Equation (33) and Equation (48), for $1 \le h < H$ we obtain

$$\operatorname{span} F_{i,h} \le \alpha_i \sum_{\ell=0}^{H-h-1} (\gamma\lambda\kappa)^\ell ((1-\lambda)\varepsilon_V + \lambda\varepsilon_Q) + \sum_{\ell=0}^{H-h-1} (\gamma\kappa)^\ell \varepsilon_Q.$$

Because $\alpha_i \le 1$ and $h \ge 1$, each sum is at most its corresponding $K_c(H)$. The terminal span is zero. Substitution in Equation (46) proves the result. No strict contraction is required: $\kappa = 1$ is allowed because all sums are finite. □

A sharper quantity for exact diagnostics. The proof also gives a computable stage- and agent-specific certificate. With $b_{i,h} = (1-\lambda)D_{i,h} + \lambda e_{i,h}$, set

$$B_{i,h} = \alpha_i \sum_{\ell=0}^{H-h-1} (\gamma\lambda\kappa)^\ell \operatorname{span} b_{i,h+\ell} + \sum_{\ell=0}^{H-h-1} (\gamma\kappa)^\ell \operatorname{span} \bar{e}_{h+\ell}, \qquad B_{i,H} = 0.$$

Then $\operatorname{span} F_{i,h} \le B_{i,h}$. Substituting $B_{i,h+1}$ in Equation (45) retains the actual mixer weights and remaining horizon. For each parameter coordinate k, the same argument replaces $\|z_{i,h}\|$ by $|z_{i,h,k}|$ and bounds $|(g_L^{(\lambda)} - \nabla J_{\widehat{R}})_k|$. These certificates follow from the same proof and do not require extra model-fitting assumptions.

Corollary 6 (Exact agreement)**.** *If* $[(P_{i,h}^{a_i} - P_{\pi,h})F_{i,h+1}](\mathbf{s}) = 0$ *for every reached* $(h, \mathbf{s}, i, a_i)$ *with positive policy probability, then* $g_L^{(\lambda)} = \nabla J_{\widehat{R}}$*. In particular, agreement holds if* $\lambda = 1$ *and every* $e_{i,h}$ *is constant across stage-*h *states. It also holds when all actions leave the next-state distribution unchanged, when* $H = 1$*, or when* $\gamma = 0$*.*

Proof. The first condition sets every summand in Equation (34) to zero. At $\lambda = 1$ and $\varepsilon_Q = 0$, Equation (47) is zero, whatever ε_V. If $P_{i,h}^{a_i} = P_{\pi,h}$ for all reached actions, the transition contrasts vanish. When $H = 1$, only $F_{i,1} = 0$ enters; when $\gamma = 0$, every γ^{h+1} is zero. □

At $\lambda = 1$, Equation (33) simplifies further to

$$F_i = -\mathcal{R}_\gamma \sum_{j \ne i} \alpha_j e_j.$$

Thus the acting agent's own compatibility term cancels from its continuation discrepancy; the remaining terms come from the other components. In particular, a one-agent model has exact agreement at $\lambda = 1$ for arbitrary local compatibility error. The first condition is sufficient. The sum can also vanish through score projection or cancellation between its terms. The value discrepancies e_i and D_i therefore need not vanish to obtain exact expected-gradient agreement.

A.9 TECHNICAL TRANSFER TO TERMINAL RETURN

The following lemma uses a Euclidean credit-error certificate, such as Equation (46) or Equation (47), at the chosen λ. Let $J^*(\vartheta) = \mathbb{E}_{\pi_\vartheta} G^*$ denote the task's expected team return. Fix $a > 0$ and $b \in \mathbb{R}$ for the frozen reward model. For the current parameter θ and the actual updated parameter $\theta + u$, suppose

$$|J_{\widehat{R}}(\vartheta) - aJ^*(\vartheta) - b| \le \varepsilon_{\text{out}}, \qquad \vartheta \in \{\theta, \theta + u\}. \tag{49}$$

A uniform condition on any class containing these two policies implies this endpoint condition. Neither the intermediate segment nor unrelated policies need outcome calibration for the result below.

Lemma 7 (Terminal-gain certificate)**.** *Assume the endpoint condition in Equation* (49)*. Suppose $J_{\widehat{R}}$ is differentiable along the segment from θ to $\theta + u$, and its gradient is L-Lipschitz on that segment. For any valid certificate $\varepsilon_{\text{credit}} \ge \|g_L^{(\lambda)} - \nabla J_{\widehat{R}}(\theta)\|$,*

$$\begin{aligned} a\big(J^*(\theta + u) - J^*(\theta)\big) \ge \big(g_L^{(\lambda)}\big)^\top u - \varepsilon_{\text{credit}}\|u\| \\ - \frac{L}{2}\|u\|^2 - 2\varepsilon_{\text{out}}. \end{aligned} \tag{50}$$

The displacement u is the actual parameter change and may result from any number of optimizer steps.

Proof. By the fundamental theorem of calculus along the segment,

$$\begin{aligned} J_{\widehat{R}}(\theta + u) - J_{\widehat{R}}(\theta) &= \nabla J_{\widehat{R}}(\theta)^\top u + \int_0^1 [\nabla J_{\widehat{R}}(\theta + tu) - \nabla J_{\widehat{R}}(\theta)]^\top u \, dt \\ &\ge \nabla J_{\widehat{R}}(\theta)^\top u - \int_0^1 Lt\|u\|^2 \, dt \\ &\ge \big(g_L^{(\lambda)}\big)^\top u - \varepsilon_{\text{credit}}\|u\| - \frac{L}{2}\|u\|^2. \end{aligned}$$

The last line uses Cauchy–Schwarz. Writing $\rho(\vartheta) = J_{\widehat{R}}(\vartheta) - aJ^*(\vartheta) - b$, we have

$$\begin{aligned} a\big(J^*(\theta + u) - J^*(\theta)\big) &= J_{\widehat{R}}(\theta + u) - J_{\widehat{R}}(\theta) - \rho(\theta + u) + \rho(\theta) \\ &\ge J_{\widehat{R}}(\theta + u) - J_{\widehat{R}}(\theta) - 2\varepsilon_{\text{out}}. \end{aligned}$$

Combining the inequalities proves Equation (50). □

A positive right-hand side certifies a positive terminal-return gain because $a > 0$. Endpoint-specific residual bounds can replace $2\varepsilon_{\text{out}}$ by their sum. If u is random, the statement holds for each realized update satisfying the stated conditions; no independence between u and the training data is used.

A.10 GENERAL SPACES AND EPISODE BOUNDARIES

The same proofs apply to general measurable state and action spaces if sums are replaced by integrals, functions are measurable and the following regularity conditions hold. Policy distributions must admit differentiable densities on a parameter-independent feasible support, satisfy $\mathbb{E}[z_{i,h} \mid \mathbf{S}_h] = 0$, and allow differentiation under the trajectory expectation. A sufficient condition is bounded frozen q_i, v_i, V together with a common trajectory reference measure and an integrable local envelope for the norm of the trajectory-density derivative. Require also the integrability of the score-weighted return and advantage terms, and $\mathbb{E}\sum_{h,i} \|z_{i,h}\|^2 < \infty$ for the Fisher matrix. Spans become supremum minus infimum, and total variation is the usual total variation of probability measures. The midpoint argument, Markov recursions and finite backward sums remain unchanged. Finite spans and finite $\mathcal{C}_\pi$ give finite certificates. History states may yield $\kappa = 1$, which is already allowed.

All derivatives hold the learned reward model and advantage-computation critic fixed. Complete episodes use the terminal conditions stated above. A zero-boundary rollout cutoff instead defines a truncated-return target; full-episode and bootstrapped targets retain their corresponding boundary terms.

B SIGNAL FLOORS FOR STRUCTURED TERMINAL OUTCOMES

Boolean influence analysis supplies the first-level energy and classical tribes family (O'Donnell, 2014). Biased read-once composition also appears in Fourier-entropy/total-influence analysis (O'Donnell & Tan, 2013). Our signal result lower-bounds first-level energy with a sharp certificate-width coefficient.

Proof roadmap. The scalar core, sparse-AND and all-mean composition inequalities lead to the read-once theorem. The width-two correction handles intermediate means outside the sparse regime. The one-coalition theorem is a separate extension; the final corollary connects the floor to calibrated actor directions.

This appendix proves Theorem 2 and an extension with one additional overlapping coalition. Throughout, A_i are independent Bernoulli actions in a one-decision task, $\rho \le p_i < 1$, and $0 < \rho \le 1/2$. The outcome Y is a nonconstant monotone Boolean function of these actions, with mean $\mu = \mathbb{E}Y$. A *minimal winning certificate* is an inclusion-minimal set of actors whose activation guarantees $Y = 1$; its maximum size is the certificate width k. A monotone AND/OR formula is *read-once* when each physical actor occurs at most once and no Boolean constants are included. Irrelevant actors may be omitted.

The population individual-credit energy is

$$\mathcal{C}(Y) = \sum_i \mathrm{Var}(\mathbb{E}[Y \mid A_i]) = \sum_i p_i(1-p_i)(\partial_{p_i}\mu)^2. \tag{51}$$

Indeed, conditioning on a Bernoulli coordinate gives $\partial_{p_i}\mu = \mathbb{E}[Y \mid A_i = 1] - \mathbb{E}[Y \mid A_i = 0]$. For $0 < u < 1$, write

$$\varphi(u) = (1-u)^2[-\log(1-u)], \qquad a_k(\rho) = \frac{k\rho^{2k-1}(1-\rho)}{(1-\rho^k)^2[-\log(1-\rho^k)]}. \tag{52}$$

All logarithms are natural. The stated full-support assumptions imply $0 < \mu < 1$ and make the equality cases below unambiguous.

For separate actor logits $\theta_i = \log[p_i/(1-p_i)]$, the Fisher matrix is $F = \mathrm{diag}(p_i(1-p_i))$ and $\partial_{\theta_i}\mu = p_i(1-p_i)\partial_{p_i}\mu$. Thus

$$(\nabla_\theta \mu)^\top F^{-1} \nabla_\theta \mu = \mathcal{C}(Y). \tag{53}$$

This identity concerns the true terminal outcome and independently parameterized logits. The proof below does not assume that fitted local reward heads equal the conditional expectations in (51).

B.1 EXACT COMPOSITION IDENTITIES

For a Boolean event U on independent physical actors, write its mean as u and its credit energy as C_U. If events U, V use disjoint actor sets, differentiation of their product mean gives

$$C_{U\wedge V} = v^2 C_U + u^2 C_V. \tag{54}$$

For independent events $Y_1, \dots, Y_m$, with means μ_j and energies C_j, their OR has mean v satisfying

$$1 - v = \prod_j (1-\mu_j), \qquad \frac{C_{\vee_j Y_j}}{(1-v)^2} = \sum_j \frac{C_j}{(1-\mu_j)^2}. \tag{55}$$

Indeed a derivative belonging to component j is multiplied by the failure probabilities of all other components. Summing its squared contribution proves the identity. Consequently, if $C_j \ge a\varphi(\mu_j)$ for a common a, then

$$\frac{C_{\vee_j Y_j}}{(1-v)^2} \ge a \sum_j [-\log(1-\mu_j)] = a[-\log(1-v)].$$

Equality holds exactly when every component bound is an equality. These identities require disjoint physical actor sets; they do not replace a shared actor by independent copies.

For a singleton $Y = A_i$, put $t = -\log(1-p_i)$. Then

$$\frac{p_i(1-p_i)}{\varphi(p_i)} = \frac{e^t - 1}{t} = \int_0^1 e^{st}\, ds \ge a_1(\rho), \tag{56}$$

with equality exactly when $p_i = \rho$, since the ratio increases strictly with p_i.

For m disjoint coalitions, each containing k actors at probability ρ, set $u = \rho^k$ and $q = 1 - \mu = (1-u)^m$. A single coalition has energy $k\rho^{2k-1}(1-\rho)$, so the OR identity gives

$$\frac{\mathcal{C}}{q^2} = \frac{mk\rho^{2k-1}(1-\rho)}{(1-u)^2} = a_k(\rho)[-\log q]. \tag{57}$$

This family establishes sharpness of the coefficient, including the sparse-success case whenever $\mu \le 1/2$.

B.2 Scalar inequalities

For $t > 0$, set

$$B(t) = (e^t - 1)^2[-\log(1 - e^{-t})], \qquad B(0) = 0.$$

Thus $B(-\log u) = \varphi(u)/u^2$.

For the elementary comparisons below, put $y = (x-1)/(x+1)$ and define

$$\ell_-(x) = 2\sum_{m=0}^{3} \frac{y^{2m+1}}{2m+1}, \qquad \ell_+(x) = \ell_-(x) + \frac{2y^9}{9(1-y^2)}.$$

The positive logarithm series and its geometric tail give

$$\ell_-(x) < \log x < \ell_+(x), \qquad x > 1. \tag{58}$$

Thus the comparisons below reduce to fixed-degree polynomial inequalities, without decimal approximations.

Lemma 8. *For $t > 0$, $B'(t) > 0$, $B''(t) > B'(t)$, and $\log B'(t)$ is strictly concave. Moreover,*

$$\frac{d}{dt}\left(\frac{tB'(t)}{B(t)}\right) > 0 \qquad \textit{for } t \ge \log 2.$$

Proof. The logarithm series gives

$$B(t) = e^t - \frac{3}{2} + \sum_{n=1}^{\infty} \frac{2e^{-nt}}{n(n+1)(n+2)}. \tag{59}$$

The series and its derivatives converge locally uniformly on $t > 0$. Hence

$$B''(t) - B'(t) = \sum_{n=1}^{\infty} \frac{2e^{-nt}}{n+2} > 0.$$

Also $B'(t) = e^t[1 - A(t)]$, where

$$A(t) = \sum_{m=2}^{\infty} \frac{2e^{-mt}}{m(m+1)}.$$

The coefficients of A sum to one, so $0 < A(t) < 1$ and $A''(t) > 0$. This proves $B' > 0$ and

$$(\log B')''(t) = -\frac{A''(t)}{1 - A(t)} - \frac{[A'(t)]^2}{[1 - A(t)]^2} < 0.$$

It remains to prove the restricted elasticity claim. Put $p = e^{-t} \le 1/2$, $L = -\log(1-p)$, and

$$\frac{B'(t)}{B(t)} = H(p) = \frac{2 - p/L}{1-p}, \qquad N(p) = 2 - p/L.$$

We first establish $N'(p)/N(p) \le 2/3$. Since $N' = [p/(1-p) - L]/L^2$, this is equivalent to

$$4L^2 + (3 - 2p)L - \frac{3p}{1-p} \ge 0.$$

The elementary bound $L \geq 2p/(2-p)$ follows by differentiating the difference. The preceding expression is increasing in positive L and is therefore at least

$$\frac{p^2(2-p-4p^2)}{(2-p)^2(1-p)} > 0, \qquad 0 < p \leq \tfrac{1}{2}.$$

Consequently

$$\frac{H'(p)}{H(p)} = \frac{N'(p)}{N(p)} + \frac{1}{1-p} \leq \frac{8}{3}.$$

For $E(t) = tB'(t)/B(t) = tH(e^{-t})$, the inequality $te^{-t} \leq 1/e$ now gives

$$E'(t) = H(p) - tpH'(p) \geq H(p)\left(1 - \frac{8}{3e}\right) > 0.$$

Here $e > 8/3$ follows, for example, from its positive power series. □

Lemma 9. *For fixed $0 < \rho < 1$, $a_j(\rho)$ decreases strictly with positive integer j.*

Proof. Let $T = -\log\rho$ and $K = (1-\rho)/\rho$. Then

$$a_j(\rho) = \frac{jK}{B(jT)}.$$

For all $t > 0$, put $p = e^{-t}$ and $L = -\log(1-p)$. Since $L > p$,

$$\frac{B'(t)}{B(t)} = \frac{2 - p/L}{1-p} > \frac{1}{1-p}.$$

Also $t = -\log p > 1 - p$, so $tB'(t)/B(t) > 1$. Thus $B(t)/t$ increases strictly, proving the assertion. This argument does not assert elasticity monotonicity below $\log 2$. □

Lemma 10 (Core composition)**.** *Let $c, l \geq 1$ be integers, $0 < \rho \leq 1/2$, $\rho^c \leq u \leq 1$, and $\rho^l \leq v < 1$. Then*

$$u^2\left[v^2 c(u^{-1/c} - 1) + a_l(\rho)\varphi(v)\right] \geq a_{c+l}(\rho)\varphi(uv). \tag{60}$$

Equality holds exactly when $u = \rho^c$ and $v = \rho^l$.

Proof. Write $T = -\log\rho \geq \log 2$, $K = e^T - 1$, $z = u^{-1/c} \in [1, e^T]$, and $s = -\log v \in (0, lT]$. Dividing by u^2v^2, the desired inequality is

$$F(z,s) := c(z-1) + a_l B(s) - a_{c+l}B(c\log z + s) \geq 0.$$

All coefficients here use the same ρ. By Lemma 8,

$$\partial_z^2 F = -\frac{a_{c+l}c}{z^2}\left[cB''(c\log z + s) - B'(c\log z + s)\right] < 0.$$

Thus it suffices to check the two endpoints in z. At $z = 1$, $F(1,s) = (a_l - a_{c+l})B(s) > 0$ by Lemma 9.

At $z = e^T$, write

$$G(s) = cK + a_l B(s) - a_{c+l}B(cT + s), \qquad R(s) = \frac{B'(cT+s)}{B'(s)}.$$

Strict concavity of $\log B'$ makes R strictly decreasing. For $s \leq lT$,

$$\frac{a_{c+l}}{a_l}R(s) \geq \frac{a_{c+l}}{a_l}R(lT) = \frac{E((c+l)T)}{E(lT)} > 1,$$

where $E(t) = tB'(t)/B(t)$ and both arguments are at least $\log 2$. Hence $G'(s) = B'(s)[a_l - a_{c+l}R(s)] < 0$. The coefficient formula gives $G(lT) = cK + lK - (c+l)K = 0$, so $G(s) \geq 0$. Concavity of F proves the inequality. The first endpoint is strictly positive, the second has its unique zero at $s = lT$, and F is strictly concave in z. Its only zero is therefore $(z,s) = (e^T, lT)$, giving the asserted equality criterion. □

B.3 AND COMPOSITION FOR A SPARSE PARENT

The next result permits both factors to be general independent component tasks, provided their individual signal bounds already hold. They need not be single conjunctions.

Lemma 11 (Sparse AND composition)**.** *For integers $c, l \geq 1$, $0 < \rho \leq 1/2$, and*

$$\rho^c \leq u < 1, \qquad \rho^l \leq v < 1, \qquad uv \leq \tfrac{1}{2},$$

one has

$$\boxed{a_c(\rho)v^2\varphi(u) + a_l(\rho)u^2\varphi(v) \geq a_{c+l}(\rho)\varphi(uv).}$$

Equality holds exactly when $u = \rho^c$ and $v = \rho^l$.

The proof uses strict subadditivity and a uniform separation between successive coefficients. Both follow from the same scalar function B.

Lemma 12. *Let $h(t) = B'(t)/B''(t)$. If $t, s > 0$ and $t + s \geq \log 2$, then $h(t) + h(s) > h(t + s)$.*

Proof. Differentiating the locally uniformly convergent series (59),

$$B''''(t) = e^t + \sum_{n\geq 1} \frac{2n^3 e^{-nt}}{(n+1)(n+2)} > 0.$$

Hence B''' increases strictly. Lemma 8 gives $0 < h < 1$ and $h' > 0$. For $d(t) = h(t) - t$,

$$d'(t) = -\frac{B'(t)B'''(t)}{[B''(t)]^2}.$$

Thus d can only increase and then decrease, so its minimum on any compact interval occurs at an endpoint. Since $B'(t) \to 0$ and $B''(t) \geq e^t$ as $t \downarrow 0$, its continuous extension has $d(0) = 0$.

Choose the split

$$L = \log 2, \qquad a = \log(6/5), \qquad b = \log(5/3), \qquad a + b = L,$$

and set $\delta = \min\{d(a), d(b)\}$. We establish the separations

$$B'''(b) > 0, \qquad \delta > 0, \qquad d(L) < 2\delta.$$

To see them directly, for $p = e^{-t}$ and $\ell = -\log(1 - p)$ write

$$H(p, z) = \frac{(1 - p)(2z - p)}{(4 - 2p)z - 3p}, \qquad h(t) = H(p, \ell).$$

This rational function decreases in z on the intervals used below. Set

$$\begin{aligned} D_- &= \min\{H(5/6, \ell_+(6)) - \ell_+(6/5), \\ &\qquad\qquad H(3/5, \ell_+(5/2)) - \ell_+(5/3)\}, \\ D_+ &= H(1/2, \ell_-(2)) - \ell_-(2). \end{aligned}$$

The explicit polynomials defining $\ell_\pm$ give $0 < D_+ < 2D_-$ after clearing their positive denominators. Equation (58) therefore yields $\delta > D_- > D_+/2 > d(L)/2$. Differentiation also gives

$$B'''(t) = \frac{(8 - 2p)\ell - p(7 - 5p)/(1 - p)}{p^2}.$$

At $p = 3/5$, replacing ℓ by $\ell_-(5/2)$ leaves a positive numerator, again a rational polynomial comparison. Hence $B'''(b) > 0$.

It follows that d decreases and $h' < 1$ on $[b, \infty)$. Also $d \geq 0$ on $[0, b]$ and $d \geq \delta$ on $[a, b]$. The lower logarithm bound gives $b > 1/2$. If $t, s \geq b$, then $h(t) + h(s) \geq 2h(b) > 1 > h(t + s)$. If $t \geq b \geq s > 0$, integration gives $h(t + s) - h(t) < s \leq h(s)$; the other ordering is symmetric. Otherwise $t, s < b$ and $t + s \geq L$ put both arguments in $[a, b]$, so

$$h(t) + h(s) \geq t + s + 2\delta > t + s + d(L) \geq t + s + d(t + s) = h(t + s).$$

This covers every case. □

Lemma 13. *There is a universal constant ϑ such that, for every integer $j \geq 2$ and $0 < \rho \leq 1/2$,*

$$\frac{a_{j-1}(\rho)}{a_j(\rho)} \geq \vartheta > \frac{3}{2}, \qquad 2\vartheta B\left(\frac{\log 2}{2}\right) > B(\log 2).$$

Proof. Write $L = \log 2$ and $T = -\log \rho$. Elasticity monotonicity in Lemma 8 shows that each coefficient ratio increases with $T \geq L$. Define

$$\vartheta = \min\left\{\frac{5}{3},\ \min_{2\leq j\leq 5} \frac{(j-1)B(jL)}{jB((j-1)L)}\right\}.$$

For $j \geq 6$, $B'/B > 1$ implies

$$\frac{a_{j-1}}{a_j} = \frac{j-1}{j}\frac{B(jT)}{B((j-1)T)} > 2\frac{j-1}{j} \geq \frac{5}{3}.$$

The desired ratio bound follows by definition for the remaining indices.

For the two strict separations, put $x_j = 2^j/(2^j-1)$. The identities

$$B(jL) = (2^j-1)^2 \log x_j, \qquad B(L/2) = (\sqrt{2}-1)^2 \log(2+\sqrt{2})$$

and (58) give

$$\begin{aligned}\vartheta \ &\geq \min\left\{\frac{5}{3},\ \min_{2\leq j\leq 5} \frac{(j-1)(2^j-1)^2\ell_-(x_j)}{j(2^{j-1}-1)^2\ell_+(x_{j-1})}\right\} \\ &> \max\left\{\frac{3}{2},\ \frac{\ell_+(2)}{2(\sqrt{2}-1)^2\ell_-(2+\sqrt{2})}\right\} > \frac{B(L)}{2B(L/2)}.\end{aligned}$$

The middle comparison is a fixed-degree polynomial inequality: clear positive denominators, reduce powers of $\sqrt{2}$ using its square, and compare the resulting rational terms. This proves both strict separations. □

Proof of Lemma 11. Put $k = c+l$, $T = -\log \rho$, $t = -\log u$, and $s = -\log v$. Dividing the desired inequality by u^2v^2 gives $F(t,s) \geq 0$, where

$$F(t,s) = a_c B(t) + a_l B(s) - a_k B(t+s).$$

Consider the compact closure

$$\mathcal{D} = \{(t,s) : 0 \leq t \leq cT,\ 0 \leq s \leq lT,\ t+s \geq \log 2\}.$$

The function is continuous there, with $B(0) = 0$. On $t = 0$, $F = (a_l - a_k)B(s) > 0$, and similarly for $s = 0$. On $t = cT$, Lemma 10 at its anchored core mean gives $F \geq 0$, because $a_c B(cT) = c(e^T - 1)$; equality requires $s = lT$. The face $s = lT$ is symmetric. The intersections of these faces with a zero coordinate are already covered by the strict zero-face bounds. On $t + s = \log 2$, convexity of B and Lemma 13 give

$$F \geq a_{k-1}[B(t) + B(s)] - a_k B(\log 2) \geq 2a_{k-1}B\left(\frac{\log 2}{2}\right) - a_k B(\log 2) > 0.$$

At an interior stationary point, set $z = t + s$ and

$$a_c B'(t) = a_l B'(s) = a_k B'(z) = M > 0, \qquad r = B''/B' = 1/h.$$

The Hessian is

$$M\begin{pmatrix} r(t) - r(z) & -r(z) \\ -r(z) & r(s) - r(z)\end{pmatrix},$$

whose determinant is

$$M^2 r(t)r(s)r(z)[h(z) - h(t) - h(s)] < 0$$

by Lemma 12. Thus no interior stationary point can be a local minimum. The nonnegative boundary values imply $F \geq 0$ throughout $\mathcal{D}$. An interior zero would also be a local minimum and is excluded. The boundary analysis leaves only $(t,s) = (cT, lT)$ for equality; substitution verifies this point. □

B.4 AN ALL-MEAN PROFILE FOR THE INDUCTION

We now remove the restriction on the means of intermediate AND nodes. The necessary correction affects only width two above success probability one half. Throughout this section put

$$L = \log 2, \qquad T = -\log \rho \geq L, \qquad K = e^T - 1, \qquad a_j = \frac{jK}{B(jT)},$$

and define

$$b_1 = a_1, \qquad b_2 = \min(a_2, a_1/2), \qquad b_j = a_j \quad (j \geq 3), \tag{61}$$

$$\kappa_j(u) = \begin{cases} b_2, & j = 2 \text{ and } u > 1/2, \\ a_j, & \text{otherwise.} \end{cases} \tag{62}$$

The dependence on the common exploration floor ρ is suppressed.

Why width two needs a correction. Let Y be the AND of two independent ORs, each containing m actors at probability ρ. With $\varepsilon = (1-\rho)^m$, direct use of the composition identities gives

$$\mu = (1-\varepsilon)^2, \qquad \mathcal{C} = \frac{2m\rho}{1-\rho}\varepsilon^2(1-\varepsilon)^2, \qquad \frac{\mathcal{C}}{\varphi(\mu)} \longrightarrow \frac{a_1(\rho)}{2}.$$

At $\rho = 1/2$, this limit is strictly below $a_2(1/2)$: the comparison is equivalent to $2^{14} < 3^9$. Thus the uncorrected all-mean assertion fails even for read-once width-two tasks. Conversely, ORs of sufficiently many disjoint width-two conjunctions have mean above one half and attain a_2 exactly. Any coefficient valid uniformly over nonsparse width-two tasks must therefore be at most $\min(a_2, a_1/2)$.

Lemma 14 (All-mean AND composition)**.** *For positive integers c, l, $0 < \rho \leq 1/2$, and $\rho^c \leq u < 1$, $\rho^l \leq v < 1$,*

$$\boxed{\kappa_c(u)v^2\varphi(u) + \kappa_l(v)u^2\varphi(v) \geq \kappa_{c+l}(uv)\varphi(uv).}$$

Equality holds exactly when $u = \rho^c$ and $v = \rho^l$.

B.4.1 QUADRATIC COMPARISON AND CORRECTED COEFFICIENTS

Lemma 15. *The function $B(t)/t^2$ decreases strictly on $0 < t \leq L$. Moreover,*

$$\frac{a_1}{a_3} > 3, \qquad \frac{B'(2L)}{B'(L)} < \frac{9}{4}.$$

Proof. Put $q = 1 - e^{-t} \leq 1/2$ and $\ell = -\log q$. The inequality $tB'(t)/B(t) < 2$ is equivalent to

$$2\ell(t-q) < t(1-q).$$

Choose $\xi = 2 - L^{-1}$. The elementary bounds $2/3 < L < 3/4$ give $1/2 < \xi < 2/3$. The function

$$D(q) = \frac{q}{1-\xi q} + \log(1-q)$$

vanishes at both endpoints of $[0, 1/2]$, and its derivative has the sign of $q[(2\xi - 1) - \xi^2 q]$. Its minimum is therefore at an endpoint. Consequently $t \leq q/(1-\xi q)$, and

$$2\ell(t-q) \leq 2\xi q(-\log q)t \leq \frac{2\xi}{e}t < \frac{1}{2}t \leq (1-q)t.$$

Here $e > 8/3$ follows from its positive power series.

Next, $a_1/a_3 = B(3T)/(3B(T))$ increases with $T \geq L$ by Lemma 8. At $T = L$,

$$B(3L) = 49\log(8/7) > 9L = 9B(L),$$

where $49\ell_-(8/7) > 9\ell_+(2)$ follows from the explicit polynomial bounds (58). This proves the coefficient claim. Finally $B'(L) = 4L - 1$ and $B'(2L) = 24\log(4/3) - 3$. The same bounds give $32\ell_+(4/3) - 12\ell_-(2) < 1$, hence $4B'(2L) < 9B'(L)$. □

Lemma 16 (Harmonic closure)**.** *For all positive integers c, l,*

$$\frac{1}{b_{c+l}} \geq \frac{1}{b_c} + \frac{1}{b_l}.$$

In particular, if $t, s > 0$ and $z = t + s \leq L$, then

$$b_c B(t) + b_l B(s) > b_{c+l} B(z). \tag{63}$$

Proof. Lemma 15 gives $a_1 > 3a_3$. Induction using Lemma 13 gives $a_1 > ja_j$ for every $j \geq 3$. Thus $b_j = \min(a_j, a_1/j)$ for all $j \geq 1$.

Put $d_j = 1/b_j$. Then $d_2 \geq 2d_1$, while

$$\frac{d_3}{d_2} = \frac{b_2}{a_3} = \min\left(\frac{a_2}{a_3}, \frac{a_1}{2a_3}\right) > \frac{3}{2}.$$

For $j \geq 4$, $d_j/d_{j-1} \geq \vartheta > 3/2$. If $c = l = 1$, the claim is the bound on d_2. If one index is one and the other is $m \geq 2$, then $d_{m+1} > (3/2)d_m \geq d_m + d_1$. If $c \geq l \geq 2$, then $d_{c+l} > (3/2)^l d_c > 2d_c \geq d_c + d_l$. These cases prove the reciprocal inequality.

By Lemma 15 and Cauchy–Schwarz,

$$\begin{aligned} b_c B(t) + b_l B(s) &> \frac{B(z)}{z^2}(b_c t^2 + b_l s^2) \\ &\geq \frac{B(z)}{d_c + d_l} \geq b_{c+l} B(z). \end{aligned}$$

The first inequality is strict because $0 < t, s < z \leq L$. □

B.4.2 Shifted and anchored boundary inequalities

Lemma 17. *For every $s \geq 0$,*

$$B(L + s) < \frac{3}{2} B(L) + 3B(s). \tag{64}$$

For $0 \leq s \leq L$, this also gives

$$B(L + s) < \frac{9}{4}[B(L) + B(s)]. \tag{65}$$

Proof. Use (59), $B(L) = L$, and $x = e^s \geq 1$. Retaining the first two positive terms gives

$$\frac{3}{2}L + 3B(s) - B(L + s) \geq x - 3 + \frac{3}{2}L + \frac{5}{6x} + \frac{11}{48x^2}.$$

The expression $x + 5/(6x) + 11/(48x^2)$ is strictly greater than two. Indeed, multiplying its difference from two by $48x^2$ and putting $y = x - 1 \geq 0$ gives

$$48x^3 - 96x^2 + 40x + 11 = 48(y - 1/12)^2 + 8/3 + 48y^3 > 0.$$

The displayed gap is thus greater than $-1 + (3/2)L > 0$. The series extends continuously to $s = 0$. If $s \leq L$, then $B(s) \leq B(L)$ and $(3/2)B(L) + 3B(s) \leq (9/4)[B(L) + B(s)]$, proving the second claim. □

Lemma 18 (Interface faces)**.** *For $l \geq 1$, $k = l + 2$, $T \geq L$, and $s \geq 0$,*

$$b_2 B(L) + b_l B(s) - a_k B(L + s) > 0. \tag{66}$$

Proof. For $l = 1$, $b_2/a_3 > 3/2$ and $a_1/a_3 > 3$, so (64) applies directly. For $l \geq 2$, both b_2/a_k and b_l/a_k exceed $(3/2)\vartheta > 9/4$. For the first ratio, combine $b_2/a_3 > 3/2$, $a_3/a_4 \geq \vartheta$, and coefficient monotonicity. For $l = 2$ the second ratio is identical; for $l \geq 3$, $a_l/a_{l+2} \geq \vartheta^2 > (3/2)\vartheta$. Thus (65) gives positivity on $0 \leq s \leq L$.

For $s \geq L$, strict log-concavity of B' and Lemma 15 give

$$\frac{B'(L + s)}{B'(s)} \leq \frac{B'(2L)}{B'(L)} < \frac{9}{4} < \frac{3}{2}\vartheta < \frac{b_l}{a_k}.$$

Hence the left side of (66) increases strictly on $[L, \infty)$ from an already positive value. □

Lemma 19 (Anchored upper faces). *For $l \geq 1$, $k = l + 2$, $T \geq L$, and $0 \leq t \leq L$,*

$$b_2 B(t) + a_l B(lT) - a_k B(lT + t) > 0. \tag{67}$$

Proof. Divide the expression by a_k, obtaining

$$H_T(t) = \frac{b_2}{a_k} B(t) + \frac{l}{k} B(kT) - B(lT + t).$$

The ratio b_2/a_k is the minimum of a_2/a_k and $a_1/(2a_k)$. Each increases strictly with T: its logarithmic derivative is a positive difference of elasticities divided by T, by Lemma 8. The derivative of the last two terms is $l[B'(kT) - B'(lT + t)] > 0$, since $2T > t$ and $B'' > 0$. Thus $H_T(t)$ increases with T, and its minimum occurs at $T = L$.

Temporarily denote coefficients at $T = L$ by a_j^*, b_2^*. The bound $\log(1+x) < x$ gives $a_k^* > k/(2^k - 1)$. Differentiating (59) gives $B'(r) > e^r - 1$ for $r > 0$: the coefficients in its subtracted exponential series sum to one. Therefore

$$a_k^* B'((k-1)L) > \frac{k(2^{k-1} - 1)}{2^k - 1}.$$

The rational factor increases with $k \geq 3$. At $k = 3$, exceeding $2 - (2L)^{-1}$ is equivalent to $L < 7/10$; this follows from $\ell_+(2) < 7/10$ and (58). Meanwhile,

$$b_2^* B'(L) \leq \frac{4L - 1}{2L} = 2 - \frac{1}{2L}.$$

Consequently $a_k^* B'((k-1)L) > b_2^* B'(L)$. The ratio $B'(lL + t)/B'(t)$ decreases with t, so

$$H_L'(t) = \frac{b_2^*}{a_k^*} B'(t) - B'(lL + t) < 0 \qquad (0 < t \leq L).$$

Consequently $H_L(t) \geq H_L(L) > 0$. The last strict inequality follows from Lemma 18 at $s = lL$, using $a_l \geq b_l$. Continuity includes $t = 0$. □

B.4.3 Proof of all-mean and composition

Proof of Lemma 14. Put $t = -\log u$, $s = -\log v$, and $z = t + s$. If $z < L$, every width-two coefficient involved is in its corrected regime, and (63) proves the result strictly. If $z \geq L$ and neither input uses the correction, the assertion is Lemma 11. This includes $c = l = 1$, for which the parent coefficient returns to a_2 at $z = L$.

It remains to consider a corrected input. By symmetry let $c = 2$ and $0 < t < L$. First suppose $l \neq 2$. On the compact region

$$0 \leq t \leq L, \qquad 0 \leq s \leq lT, \qquad t + s \geq L,$$

consider the continuous expression

$$F(t, s) = b_2 B(t) + a_l B(s) - a_{l+2} B(t + s).$$

All its boundary faces are strictly positive. On a zero-coordinate face, use coefficient monotonicity and $b_2 > a_3$. The face $t = L$ is covered by Lemma 18, with $a_l \geq b_l$; the face $s = lT$ is Lemma 19. On the positive part of $t + s = L$, use (63), $b_l \leq a_l$ and $b_{l+2} = a_{l+2}$. The endpoints of this diagonal are already covered by the zero faces.

At an interior stationary point the three first-derivative terms equal some $M > 0$. With $r = B''/B' = 1/h$, the Hessian determinant is

$$M^2 r(t) r(s) r(z)[h(z) - h(t) - h(s)] < 0$$

by Lemma 12, exactly as in the proof of Lemma 11. There is no interior local minimum. Compactness and the strictly positive boundary therefore give $F > 0$.

If $l = 2$, split the region at $s = L$. On its lower piece use b_2 for both inputs; on its upper piece use b_2 for the first and a_2 for the second. The stationary-point argument is unchanged. All exterior faces were just treated, and the new interface $s = L$ is positive by Lemma 18, with the factors interchanged.

Using a_2 instead of b_2 can only increase that boundary value. Thus both pieces are strictly positive. At an input mean exactly one half, (62) uses a_2; the boundary arguments deliberately allowed the smaller b_2, so this jump leaves no missing case.

The small-z and corrected-input regions are strict. In the remaining region, Lemma 11 gives equality precisely at $(t, s) = (cT, lT)$. These inputs use the uncorrected coefficients and do attain equality, completing both directions. □

B.4.4 THE READ-ONCE THEOREM

A monotone AND/OR formula is *read-once* if each physical actor appears in at most one leaf of the expression. This permits intersecting successful coalitions after expansion, and arbitrary alternation of AND and OR. It does not permit copying an actor to make two component tasks independent.

Theorem 20 (Sharp signal bound for read-once tasks)**.** *Let Y be a finite nonconstant read-once monotone AND/OR formula, with no Boolean constants, on independent Bernoulli actors satisfying $\rho \le p_i < 1$ and $0 < \rho \le 1/2$. Let k be the maximum size of a minimal winning certificate. At every $0 < \mu = \mathbb{E}Y < 1$,*

$$\boxed{\mathcal{C}(Y) \ge \kappa_k(\mu)\varphi(\mu).}$$

In particular, if $\mu \le 1/2$, then $\mathcal{C}(Y) \ge a_k(\rho)\varphi(\mu)$ with the coefficient in Theorem 2. In that sparse regime, equality holds if and only if the minimal winning coalitions are disjoint, each has size k, and every relevant actor has probability ρ.

Proof. The singleton bound is (56), valid at every probability $p \ge \rho$. At an AND node, child actors are disjoint and their certificate widths add. A child of width at most c has a winning certificate with probability at least ρ^c, so its mean is at least that value. The exact identity (54), followed by Lemma 14, propagates the profile bound through the node. Higher-arity ANDs may be grouped into binary nodes because that lemma applies at all intermediate means.

For an OR node, the width k is the maximum child width, and (55) gives

$$\frac{C_{\vee_j Y_j}}{(1-\mu)^2} = \sum_j \frac{C_{Y_j}}{(1-\mu_j)^2}, \qquad -\log(1-\mu) = \sum_j [-\log(1-\mu_j)].$$

Every child coefficient is at least the parent coefficient. For $k \ge 3$, this uses $b_2 > a_3 \ge a_k$ and ordinary coefficient monotonicity. For $k = 2$ and $\mu \le 1/2$, every child mean is at most μ, so no child uses the correction. For $k = 2$ and $\mu > 1/2$, the parent coefficient b_2 is the minimum possible child coefficient. The case $k = 1$ is immediate. The two identities therefore propagate the bound through OR. Induction proves the all-mean assertion.

Now assume $\mu \le 1/2$ and equality. At an equality AND node, Lemma 14 forces the child means to be ρ^c and ρ^l. For any monotone event of width at most c, the presence of one certificate already gives mean at least ρ^c. Equality in this lower bound forces a single certificate of exactly c actors, all with probability ρ: a shorter certificate or a larger probability increases its probability, and an additional minimal certificate adds a positive-probability success assignment under the full-support product law. Thus an equality AND node is a single conjunction at the exploration floor.

At an equality OR node, its child means are also at most one half. Strict coefficient monotonicity forces every child width to equal the parent width, and (55) forces every child bound to be an equality. Recursive application gives exactly disjoint size-k conjunctions with all relevant probabilities equal to ρ. Conversely, equation (57) gives equality for that family. □

Proof of Theorem 2. Let $j \le k$ be the actual maximum certificate width. For $\mu \le 1/2$, the profile theorem gives $\mathcal{C} \ge a_j(\rho)\varphi(\mu)$. Lemma 9 gives $a_j(\rho) \ge a_k(\rho)$, strictly when $j < k$. The preceding theorem therefore gives the claimed bound; equality requires $j = k$ and its stated disjoint-coalition equality case. □

B.5 ONE OVERLAPPING COALITION, WITH OPTIONAL FRESH ACTORS

A single insertion into the disjoint equality family permits an explicit form of actor reuse. The added coalition may also contain actors outside the original task.

Theorem 21 (One additional coalition)**.** *Start with $m \geq 1$ disjoint coalitions of $k \geq 1$ independent actors each, all acting with probability ρ, where $0 < \rho \leq 1/2$. Add one distinct coalition of k actors. It selects r_j actors from each of h old blocks, with*

$$1 \leq r_j < k, \qquad R = \sum_{j=1}^{h} r_j \leq k,$$

and uses $k - R$ fresh independent actors, also with probability ρ. The case $h = R = 0$ is allowed. Let Y indicate completion of at least one old or new coalition. At every $0 < \mu = \mathbb{E}Y < 1$,

$$\boxed{\mathcal{C} \geq a_k(\rho)(1-\mu)^2[-\log(1-\mu)].}$$

Equality holds exactly when $R = 0$. Every permitted overlap, $R > 0$, gives strict inequality.

The condition $r_j < k$ excludes reproducing an old coalition. At $R = k$ all added actors come from the old system; at $R = 0$ the new coalition is independent of it.

Proof. Put

$$u = \rho^k, \qquad h_0 = -\log(1-u), \qquad e_0 = \frac{\rho^{2k-1}(1-\rho)}{(1-u)^2}, \qquad a_k = \frac{ke_0}{h_0}.$$

For failure probability $q = 1 - \mu$, write

$$\mathcal{E} = \frac{\mathcal{C}}{q^2} = \sum_i \rho(1-\rho)(-\partial_{p_i} \log q)^2, \qquad \mathcal{L} = -\log q.$$

Derivatives are taken in the independent physical probabilities and then evaluated at $p_i = \rho$. By (57), the old disjoint system satisfies $\mathcal{E}_{\text{old}} = a_k \mathcal{L}_{\text{old}}$. Conditional on its failure, the new coalition is complete with probability

$$\delta = ux, \qquad x = \prod_{j=1}^{h} \frac{1-\rho^{k-r_j}}{1-u} \in (0,1].$$

The selected old actors contribute ρ^R, while the fresh actors contribute ρ^{k-R}. Define

$$\alpha = R/k, \qquad D = 1 - ux, \qquad W = \frac{ux}{D}.$$

Then $\Delta\mathcal{L} = -\log(1-ux)$.

Every old actor has old log-failure derivative $g = \rho^{k-1}/(1-u)$. In a hit block, let a and b be the products of its selected and unselected probabilities. At general physical probabilities, this block contributes $a(1-b)/(1-ab)$ to δ. Thus

$$\partial_{p_i} \log \delta = \begin{cases} \dfrac{1}{p_i(1-ab)}, & i \text{ selected}, \\ -\dfrac{b(1-a)}{p_i(1-b)(1-ab)}, & i \text{ unselected}. \end{cases}$$

Since $q_{\text{new}} = q_{\text{old}}(1-\delta)$, differentiating before setting $p_i = \rho$ gives the new log-failure derivatives

$$-\partial_{p_i} \log q_{\text{new}} = \begin{cases} g(1+W/u), & i \text{ selected from an old block}, \\ g(1-Wf_j), & i \text{ unselected in hit block } j, \\ g, & i \text{ in an untouched old block}, \\ W/\rho, & i \text{ fresh}, \end{cases} \qquad f_j = \frac{\rho^{-r_j}-1}{1-\rho^{k-r_j}}.$$

A fresh actor has old derivative zero. Define

$$\mathcal{A} = \sum_j (k-r_j) f_j, \qquad \mathcal{B} = \sum_j (k-r_j) f_j^2 \geq 0.$$

There are R selected old actors and $k-R$ fresh actors, and $\rho(1-\rho)g^2 = e_0$. Summing the squared derivative changes over all physical actors yields the exact identity

$$\frac{\Delta\mathcal{E}}{e_0} = 2W\left(\frac{R}{u} - \mathcal{A}\right) + W^2\left[\frac{R}{u^2} + \mathcal{B} + \frac{(k-R)(1-u)^2}{u^2}\right]. \tag{68}$$

For integers $r, s \geq 1$ with $r + s = k$, geometric sums give

$$1 - \rho^r \leq r(1-\rho), \qquad s\rho^{s-1}(1-\rho) \leq 1 - \rho^s,$$

and hence

$$u\, s\frac{\rho^{-r} - 1}{1 - \rho^s} = \frac{s\rho^s(1-\rho^r)}{1-\rho^s} \leq r\rho.$$

Summing over hit blocks proves $u\mathcal{A} \leq \rho R$.

A second geometric comparison gives $x \geq 1 - \alpha$. For $r + s = k$ with $r, s \geq 1$, put

$$g_r = 1 - \frac{1-\rho^s}{1-u} = \frac{\sum_{a=s}^{k-1}\rho^a}{\sum_{a=0}^{k-1}\rho^a} \leq \frac{r}{k}.$$

The numerator consists of the r smallest terms of the decreasing sequence $1, \rho, \ldots, \rho^{k-1}$. Since $0 \leq g_r < 1$,

$$x = \prod_j (1 - g_{r_j}) \geq 1 - \sum_j g_{r_j} \geq 1 - R/k = 1 - \alpha. \tag{69}$$

Use $\rho \leq 1/2$, $\mathcal{B} \geq 0$, and $W = ux/D$ in (68) to obtain

$$\frac{\Delta\mathcal{E}}{e_0} \geq k\left[\frac{\alpha x}{D} + \frac{x^2\{\alpha + (1-\alpha)(1-u)^2\}}{D^2}\right]. \tag{70}$$

After factoring kx/D^2, the numerator inside this expression is

$$x(1-u)^2 + \alpha[1 + u(1-u)x].$$

Its coefficient of α is positive. Substituting $\alpha \geq 1 - x$ from (69) and simplifying yields

$$\frac{\Delta\mathcal{E}}{e_0} \geq kx\left[1 + \frac{ux(1-x)}{(1-ux)^2}\right] \geq kx. \tag{71}$$

Convexity of $z \mapsto -\log(1-z)$, whose value at zero is zero, gives

$$\Delta\mathcal{L} = -\log(1-ux) \leq x[-\log(1-u)] = xh_0.$$

Combining this with (71) proves $\Delta\mathcal{E} \geq a_k\Delta\mathcal{L}$. Adding the old equality and multiplying by q_{new}^2 establishes the asserted signal bound.

If $R > 0$, at least one g_{r_j} is strictly positive and every factor in x is positive. Hence $0 < x < 1$, and the first lower bound in (71) is strictly greater than kx. If $R = 0$, the new coalition is independent of every old coalition and the disjoint-block computation gives equality. This proves both directions. □

The argument also gives an explicit increase of the normalized deficit:

$$\Delta(\mathcal{E} - a_k\mathcal{L}) \geq ke_0\left[\frac{ux^2(1-x)}{(1-ux)^2} + x - \frac{-\log(1-ux)}{h_0}\right].$$

Both terms in brackets are nonnegative, and the first is positive for every permitted overlap. This establishes stability of the disjoint equality family under one such insertion, including instances without a read-once representation. It does not justify starting from an already overlapping system, performing repeated insertions, or changing to heterogeneous actor probabilities.

Table 3: SMACLite terminal-feedback returns (five-seed mean ± SD).

Scenario	SparseMAPPO	AREL	STAS	TAR^2	IMAP-LA
2s_vs_1sc	3.43 ± 0.93	9.40 ± 0.13	9.95 ± 0.02	13.24 ± 0.96	$\mathbf{13.64 \pm 0.85}$
3s_vs_5z	4.42 ± 0.05	7.61 ± 0.50	9.22 ± 0.89	9.60 ± 0.85	$\mathbf{9.98 \pm 0.81}$
3s5z	8.45 ± 2.60	19.15 ± 0.68	19.82 ± 0.19	$\mathbf{19.91 \pm 0.16}$	19.75 ± 0.26

An instance without a read-once representation. Take $k = m = 2$, old coalitions $\{1, 2\}, \{3, 4\}$ and new coalition $\{2, 3\}$. The resulting success event is

$$A_1A_2 \vee A_2A_3 \vee A_3A_4.$$

Its minimal true certificate $\{2, 3\}$ and minimal false certificate $\{A_2 = 0, A_3 = 0\}$ intersect in two actors. In a monotone read-once formula, every minimal true certificate meets every minimal false certificate in exactly one actor. This follows by induction: at an OR node the true certificate chooses one child while the false certificate includes one from each child; at an AND node the roles reverse. Thus this instance has no read-once representation in the physical actors. At $\rho = 1/2$ it has $\mu = 1/2$ and $\mathcal{C} = 5/32$, so it also lies in the sparse-success domain of Theorem 2.

The independent-composition arguments require disjoint physical actor sets; identifying independent copies of the same actor changes the task. The read-once theorem and the single-insertion theorem therefore do not establish a bound for arbitrary overlapping coalition families.

B.6 From signal strength to a calibrated local direction

Specialize the setting of Theorem 1 to the product-action model in Section 5.3. Each actor makes its Bernoulli decision once, with a separate logit coordinate; subsequent steps have forced actions and parameter-independent dynamics. The terminal outcome is the stated function of the initial decisions. The complete trajectory score then contains only the initial action scores, so $\mathcal{I}_\pi = F = \mathrm{diag}(p_i(1 - p_i))$ is positive definite. In particular, the error bound and the signal identity (53) use the same metric.

Corollary 22 (Calibrated natural-direction certificate)**.** *Let the model satisfy Theorems 1 and 2 in this common coordinate system. Write* $B_{\mathrm{cred}} = (\varepsilon/2) \sum_{t=1}^{H-1} \gamma^t$. *For a frozen scale* $a > 0$*, assume*

$$\|\nabla J_{\widehat{R}} - a\nabla\mu\|_{F^{-1}} \le \delta_{\mathrm{cal}}.$$

If

$$B_{\mathrm{cred}} + \delta_{\mathrm{cal}} < a(1 - \mu)\sqrt{a_k(\rho) \log \frac{1}{1 - \mu}}, \tag{72}$$

then $\nabla\mu^\top F^{-1} g_L^{(1)} > 0$.

Proof. The triangle inequality and Theorem 1 give $\|g_L^{(1)} - a\nabla\mu\|_{F^{-1}} \le B_{\mathrm{cred}} + \delta_{\mathrm{cal}}$. Hence

$$\nabla\mu^\top F^{-1} g_L^{(1)} \ge \|\nabla\mu\|_{F^{-1}} \left[a\|\nabla\mu\|_{F^{-1}} - B_{\mathrm{cred}} - \delta_{\mathrm{cal}}\right] > 0.$$

The last inequality uses (53), Theorem 2, and (72). □

The corollary concerns the first-order natural direction $F^{-1} g_L^{(1)}$ at a frozen checkpoint. Its gradient-calibration assumption is distinct from accurate pair ordering or endpoint return calibration. It does not certify a finite clipped-PPO displacement. Arbitrary actor parameter sharing can project away part of this independently parameterized signal, so the floor is not asserted after such a projection. The shared-parameter credit-error theorem remains valid in its own stated setting.

C Implementation and benchmark protocol

All benchmarks use terminal-outcome feedback. SMACv2 actors and local reward heads share parameters across homogeneous agents, with agent identity included in the observations. They

use recurrent history encoders. MAMuJoCo actors and reward heads are separate across agents; actors and critics use two hidden layers of width 256. Reward-model encoders follow the domain's actor/critic encoder design but have separate parameters and optimizers. Discrete actions are one-hot encoded before the Q head, and continuous actions are concatenated with the observation embedding.

Outcome-comparison buffer. The buffer persists across PPO iterations. Binary pairs come from the win and loss groups; continuous-return pairs come from groups separated by the buffer's running median. Only completed episodes enter the comparisons. Equal-return pairs are discarded, and fitting is skipped when no unequal-return pair is available.

Table 4: Benchmark training settings. The relay uses the separate settings below.

Setting	SMACv2	MAMuJoCo
Actor / critic learning rate	0.0005	0.0003
Discount / GAE	0.99 / 0.95	0.99 / 0.95
PPO clipping / epochs	0.2 / 10	0.2 / 10
PPO minibatch size	256	512
Entropy coefficient	0.01	0.01
Parallel environments	8	8
Rollout length	400	200
Training steps	10M	10M
Evaluation episodes	32	32
Reward-model learning rate	0.0003	0.0003
Reward-model pair batch	256	512
IMAP label-bearing Q updates / iteration	5	5
IMAP additional V updates / iteration	5	5
Comparison FIFO capacity	8192 pairs	8192 pairs
Extreme-V temperature	1.0	1.0
Reward-square coefficient	0.001	0.001
Reward-model gradient clipping	0.5	0.5

Terminal feedback and rollout boundaries. SMACv2 disables intermediate damage, kill and health rewards. MAMuJoCo retains its task dynamics and cumulative evaluation return, while the learner receives only the complete-episode scalar. Complete episodes generate comparison labels even when an actor rollout ends earlier. Actor GAE uses the actual termination mask and its rollout bootstrap convention; a reset observation is not a successor inside the completed episode. The complete-episode theory diagnostics use zero terminal local values and critic values.

Reward and advantage snapshots. The normalized scalar mixer is shared by Q and V. A shared affine transform of the form $(r_i - \mu_R)/\sigma_R$, applied identically to the global reward, preserves $\sum_i \alpha_i r_i = \widehat{R}$. Each diagnostic stores the raw Q/V outputs, the scalar weights, the normalization parameters and the critic snapshot. Both GAE streams use these same fixed quantities and masks. The residual identity is verified before comparing gradients. If an implementation applies an additional reduction or advantage transform, its resulting vector is evaluated with that transform; the theoretical vector is computed with the stated discounted time weights.

Reward scaling. Certificates use the same frozen heads and reward scale as the compared vectors. Under a shared positive rescaling of q_i, v_i, V by $1/\sigma_R$, reward-part gradients, compatibility spans and their certificates scale by $1/\sigma_R$. State-constant reward shifts at fixed stages do not alter these spans. If a shift or mask varies by state, compatibility must instead be recomputed from the transformed heads. The numerical studies here use raw residuals.

For the benchmark controls, the shared affine reward transform is applied before advantage construction. The objectives use the actor prefactors in Equation (9) and a unit-prefactor global control.

What each ablation changes. SL-MAPPO and IMAP-GA share the global actor readout, terminal comparisons and label-bearing update schedule. SL-MAPPO predicts rewards directly; IMAP-GA uses Q/V residuals and additionally optimizes V, so total optimizer work differs. IMAP-LA retains the Q/V reward model with local actor advantages. Online-IPL extracts policies from learned

Table 5: Held-out reward-model diagnostics. Both frozen models score the same 2000 pairs from a shared pool of 256 trajectories per task and seed after 10M training steps. Values are descriptive means over three seeds.

	Ranking accuracy ↑		BT NLL ↓	
Task	IMAP-GA	SL-MAPPO	IMAP-GA	SL-MAPPO
protoss_5_vs_5	0.857	0.841	0.618	0.632
terran_5_vs_5	0.851	0.828	0.605	0.654
Hopper-v2	0.891	0.812	0.517	0.634

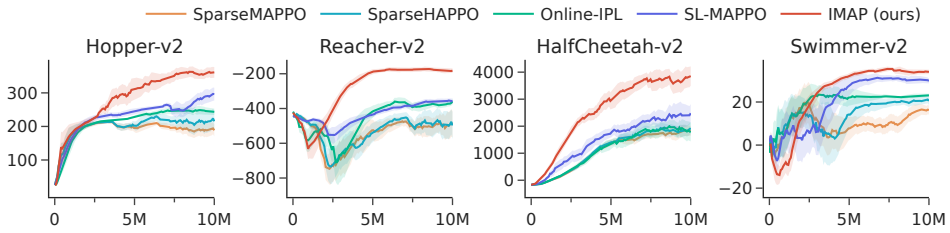

Figure 4: MAMuJoCo comparison with the benchmark baselines. IMAP (ours) denotes the terminal-outcome IMAP-Rule/LA configuration; this figure supplements the main-text GA/LA ablation.

Q/V values by advantage-weighted behavior cloning. Its extraction family is related to O-MAPL, whose published evaluation uses a separate offline protocol (Bui et al., 2025). SparseMAPPO and SparseHAPPO optimize terminal rewards directly.

The unit-prefactor global and α_i-weighted local readouts preserve the same one-stage population gradient (Appendix A.2). The local coefficient retains its component's weight in the mixed objective; multiplying the global readout by α_i again would square this coefficient. This explains the readout convention, including its common entropy term. Sequential GAE, finite samples, advantage transforms and repeated clipped updates can still produce different optimization behavior. The experiments compare complete configurations, without attributing each gain to an isolated mechanism.

Additional held-out fitting measurements. The held-out reward-model diagnostic freezes IMAP-GA and SL-MAPPO after 10M training steps. Per task and seed, both models score the same 2000 pairs drawn from a common held-out pool of 256 trajectories, with trajectories from both policies included. Table 5 reports descriptive means over three seeds. Pairs sharing trajectories are not separate experimental replicates. This diagnostic concerns score ordering and likelihood; the policy-gradient and outcome-calibration measurements are defined separately.

D Relay protocol and numerical credit checks

The relay is a finite diagnostic environment for the IMAP construction in Section 4. Four independent uniform cue bits determine four correct binary actions. Agents observe stage, current cue, surviving-success status and active role. Alternating ownership assigns successive decisions to different agents; concentrated ownership assigns all four to the first agent. Both conditions retain the same two-agent roster and parameter allocation. Failure changes the progress flag but leaves the decision schedule intact. Terminal reward is one for four correct choices and zero otherwise.

Networks and fitting. Actors allocate 32 Bernoulli logits indexed by agent, decision stage, progress and cue. The shared critic allocates 276 scalar values over a common state registry, including unused suffix states in shorter episodes. Observations have 12 coordinates: scaled time, five stage indicators, cue, progress, two role indicators and two identity indicators. Each local Q network has widths $15 \to 32 \to 32 \to 1$ and each local V network $12 \to 32 \to 32 \to 1$, with SiLU activations. Including two mixer logits, the reward model has 6214 parameters. Hidden layers use Xavier-uniform initialization;

final output heads and mixer logits start at zero. Forced no-ops have singleton action supports, and terminal local and critic values are zero.

The model uses IMAP's terminal-outcome comparison loss and mixed Extreme-V update, with temperature one and reward-square coefficient 0.001. Per 64-episode PPO batch, five fitting cycles each use 128 comparison pairs. The comparison FIFO capacity is 8192. Both Q/mixer and V have their own optimizer state. Each 4096-episode run has 320 Q/mixer and 320 V updates. PPO uses ten epochs, learning rate 0.0005, clipping 0.2, entropy coefficient 0.01, value coefficient 0.5, $\gamma = 1$ and $\lambda = 0.95$. Reward-model learning rate is also 0.0005. All gradient norms are clipped at 0.5.

Complete checkpoint population. The numerical collection contains eleven Q/V configurations: GA and LA at delays 0, 16 and 64 with alternating ownership; GA and LA with concentrated ownership at delay 16; and GA with fixed fitting datasets containing 1, 4 or 16 observed successes. Each configuration has seeds 0–9. Four saved checkpoints per run, at batches 16, 32, 48 and 64, give $11 \times 10 \times 4 = 440$ cases. All cases are included regardless of policy success, compatibility error or bound ratio. The fixed-data cases use 512 fitting episodes from a common behavior-policy pool per seed, with nested success/failure samples, then freeze the reward model during policy collection. They enter here as fitted models for numerical verification, not as learning-performance comparisons.

Frozen targets and reductions. At each checkpoint, the reward model is frozen after fitting, while the actor and critic are frozen before that batch's PPO update. Local/global rewards use the same model, weights, masks and critic snapshot. No per-agent reward rescaling or advantage centering is used. The actor loss averages the four active decision positions per episode; critic loss averages within episodes and then across episodes. At $\gamma = 1$, the implemented actor mean is one quarter of the theorem's unnormalized sum. The reported Fisher-error and continuation checks use the latter. Stored episodes and repeated pair samples are not treated as independent outcome labels.

Exact path calculations. Every frozen policy is evaluated on all 256 cue/action paths. Independent return differentiation and a fixed-target surrogate give the team and local actor vectors. The trajectory Fisher uses the complete undiscounted score and its numerically resolved range. All checks use fixed scale-aware tolerances; near-zero bounds are handled separately from ratios. No relay checkpoint is excluded. Panel (c) uses each checkpoint's original error and bound, targeting the learned objective $J_{\widehat{R}}$.

Controlled decomposition maps. The heatmaps use all ten final reference IMAP-LA checkpoints: alternating ownership, delay 16, seeds 0–9, after 4096 collected episodes. Each snapshot fixes the post-fitting reward model and the actor/critic immediately before its final PPO update. No model or policy is refitted during the intervention. The original 440-checkpoint collection remains a separate coverage check; the heatmaps use the fixed reference condition so that delay, ownership and model-fitting regime do not vary within a map.

For each seed, let

$$s_{\mathrm{QV}} = \max\left\{ \max_{t,i} \operatorname{span}_{\mathbf{s},a_i} q_{i,t}(o_i(\mathbf{s}), a_i),\ \max_{t,i} \operatorname{span}_{\mathbf{s}} v_{i,t}(o_i(\mathbf{s})) \right\},$$

where each span uses the reachable states and feasible actions at its stage. All ten scales are positive. Generate independent standard-normal entries $\psi_i(\mathbf{s})$ on the relay's public state registry using the fixed per-model seeds in the accompanying numerical record. Center them as $\widetilde{\phi}_i(\mathbf{s}) = \psi_i(\mathbf{s}) - \sum_j \alpha_j \psi_j(\mathbf{s})$, set terminal entries to zero, and define

$$\phi_i(\mathbf{s}) = s_{\mathrm{QV}} \frac{\widetilde{\phi}_i(\mathbf{s})}{\max_{t,j} \operatorname{span}_{\mathbf{s}} \widetilde{\phi}_{j,t}(\mathbf{s})}.$$

Every observed denominator is positive. The same potentials and fixed scale are used in both intervention channels and at every grid cell within a seed. The public state is present in each agent's observation, so adding these action-independent functions preserves the local-head information restriction.

Use $\lambda \in \{0, 0.05, \ldots, 1\}$ and $\zeta \in \{-2, -1.75, \ldots, 2\}$. For each of the two channels, ten seeds give $10 \times 21 \times 17$ exact evaluations, or 7140 in total. Each heatmap cell is the arithmetic mean of ten

Table 6: Variation across the ten fixed reference models at training $\lambda = 0.95$ and unit amplitude $\zeta = 1$. Values use the normalized response in Equation (73); SD and range are across seeds, not grid cells.

Output shift	Mean	Sample SD	Seed range
Joint Q/V	0.00605	0.00275	[0.00236, 0.01026]
Q-only	0.11496	0.05218	[0.04488, 0.19488]

responses normalized before averaging. Both maps use one unsaturated linear color scale; cells are displayed without interpolation. No seed, sign, endpoint or high-response cell is excluded. The complete per-seed grids retain their ranges as well as their means. The response is a magnitude, so its color does not encode an improvement direction.

Both channels leave Q_{tot} and V_{tot} unchanged. Consequently the mixed reward, its trajectory score and the reward-fitting losses on fixed data are unchanged, although individual actor directions can move. These are interventions on frozen outputs, not claims that a particular network refit realizes the altered functions. For the joint Q/V shift,

$$\Delta e_i = 0, \qquad \Delta D_i = -\zeta\phi_i;$$

for the Q-only shift,

$$\Delta e_i = \zeta\phi_i, \qquad \Delta D_i = 0.$$

Substituting in Equation (33) shows that the forcing changes are respectively $-(1-\lambda)\zeta\phi_i$ and $\lambda\zeta\phi_i$, while $\bar{e}$ remains fixed. This explains the opposite zero-response endpoints in Figure 5. At a fixed λ, the two response vectors are collinear for this shared-potential construction. The factor of 19 between their magnitudes at $\lambda = 0.95$ follows from the coefficients; it is a consistency check, not an empirical advantage of one update.

Independent surrogate differentiation agrees with the score and continuation formulas at the declared numerical tolerances. Both interventions preserve the mixed Q/V functions and reward, their zero-response endpoints are recovered, and no tested full-return or stagewise bound is violated.

The complete simultaneous-channel grid also agrees with the sum of the separate responses. Since these shared-potential channels are collinear, this is an algebraic control rather than an independent mechanism.

For a GAE parameter λ, the normalized actor response used in Figure 5 and Table 6 is

$$\Delta_\lambda(\zeta) = \frac{\|g_L^{(\lambda)}(\zeta) - g_L^{(\lambda)}(0)\|_{\mathcal{I}_\pi^\dagger}}{s_{\mathrm{QV}}}. \tag{73}$$

Delay control. A fixed-policy terminal-reward control confirms full-return gradient invariance across the three delays. In fitted models, however, the reward-square penalty averages over all transitions, including the suffix. Different suffixes therefore change this averaging as well as GAE propagation; they do not isolate exploration difficulty.

Model fitting uses float32 arithmetic; all exact path diagnostics use float64. Stored model, policy and critic snapshots specify the frozen functions used in each check.

E Numerical verification of the theory

These checks use frozen finite models and are separate from benchmark performance. They cover coupled transitions, shared categorical policy parameters, observation aliases and explicit attaining constructions. Independent enumeration and recursive differentiation agree at the declared numerical tolerances; no tested identity, Fisher-range condition or sharp bound is violated. Near-zero bounds are checked directly, without interpreting unstable error/bound ratios.

Figure 6 verifies numerical attainment in the analytic family and approach to the same constant with strictly positive mixers. The proof establishes sharpness; these calculations check its realization.

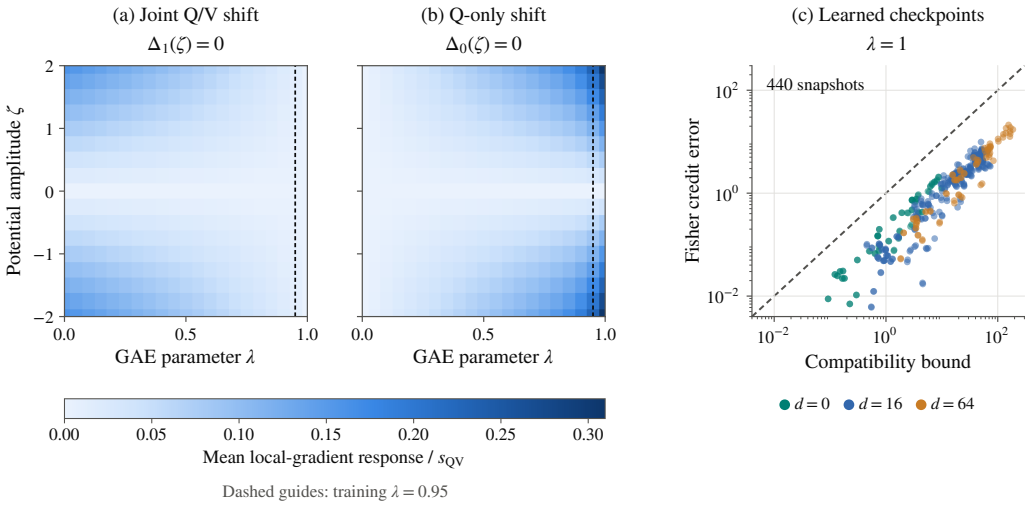

Figure 5: Local-credit sensitivity at unchanged mixed reward. (a,b) Mean response $\Delta_\lambda(\zeta)$ over ten frozen models; shared color scale and dashed training $\lambda = 0.95$. Joint Q/V shifts preserve compatibility; Q-only shifts change it. (c) Full-return Fisher error versus the Theorem 1 bound for all 440 checkpoints; maximum error/bound ratio 0.2620, with diagonal denoting equality. All panels use the frozen learned-return objective.

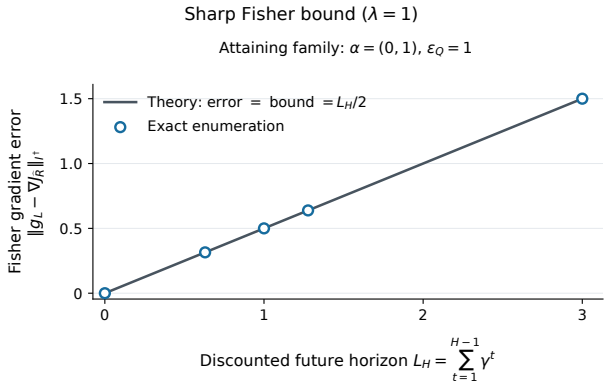

Figure 6: The attaining two-agent family at $\lambda = 1$, $\alpha = (0, 1)$ and $\varepsilon_Q = \varepsilon = 1$. Analytic error and bound coincide at $L_H/2$. Six saved finite-model checks occupy five distinct coordinates; strictly positive mixers approach the same constant.

Reproducibility record. The numerical record in the supplementary material specifies the fixed seeds, spectral cutoffs, scale-aware tolerances, complete audit populations, fitting-loss ranges and per-case output locations used in Appendices D–F. All original models and numerical records are retained.

F Instantiating the signal-to-error condition

This finite-population diagnostic evaluates both sides of (72) on fitted Q/V heads in the setting of Theorem 2. Actors are fixed throughout; no policy learning or PPO improvement is measured.

Tasks and fixed data. Four actors independently choose binary actions once. The next state publicly records their joint action, all subsequent actions are forced no-ops, and the binary outcome

Table 7: Complete product-action checkpoint grid. Counts refer to fixed task/seed/policy cases, not independent statistical replicates. Certification uses (72) at the stated tolerance.

Fitting cycles	Cases	Positive alignment	Oracle-scale certified	Return-scale certified
0	60	0	0	0
32	60	60	0	0
128	60	60	0	0
512	60	60	6	0

is delivered at termination. With $H = 2$ and $\gamma = 1$, the four tasks are

$$A_0A_1A_2A_3, \quad A_0A_1 \vee A_2A_3, \quad (A_0 \vee A_1)(A_2 \vee A_3), \quad A_0\big(A_1 \vee A_2A_3\big),$$

with maximum certificate widths $4, 2, 2, 3$, respectively. Each training dataset contains all 16 joint-action trajectories under the complete uniform-behavior population. Only their terminal outcomes supply labels. Every unequal-outcome pair is used, oriented winner first.

We fit all four tasks with seeds 0–4, giving 20 reward models. Each actor has a Q network $15 \to 32 \to 32 \to 1$ and a V network $12 \to 32 \to 32 \to 1$, with SiLU hidden layers, plus four shared mixer logits; the total is 12428 parameters. The 12 observation coordinates contain a three-way stage indicator, four public action bits (zero before the decision), two identity bits, a constant, and two unused zeros. Q additionally receives a three-way action encoding, including the no-op. Hidden weights use Xavier-uniform initialization; biases, output layers and mixer logits start at zero.

Each fit performs 512 Q-then-V cycles with the objectives (5)–(6), temperature one, reward-square coefficient 0.001, learning rate 0.0005 and gradient clipping 0.5. Q/mixer and V use separate Adam states with $(0.9, 0.999)$ moments and 10^{-8} denominator constant. V is frozen in the Q update; Q and mixer are frozen in the V update. The reward-square penalty averages the complete positive/negative pair population with its trajectory multiplicities. The V loss uses all 16 trajectories and both nonterminal transitions uniformly. This specifies the fitting distribution independently of the diagnostic policies.

Complete checkpoint and policy grid. Checkpoints at cycles $0, 32, 128, 512$ give 80 snapshots, including initialization. Each is evaluated at three fixed product policies: $p_i = 0.2$ for all actors, $p_i = 0.4$ for all actors, and $p = (0.25, 0.35, 0.45, 0.55)$. Set $\rho = \min_i p_i$. All twelve task/policy combinations satisfy $0 < \mu \le 1/2$, with $\mu \in [0.0016, 0.4096]$. Every one of the 240 cases is retained.

Full trajectory enumeration computes the true gradient $g_Y = \nabla\mu$, learned-score gradient $g_F = \nabla J_{\widehat{R}}$, ideal energy $\mathcal{C} = g_Y^\top F^{-1} g_Y$, and its lower bound $\mathcal{C}_{\mathrm{floor}}$ from (19). The full-return local vector is $(g_L)_i = \alpha_i \mathbb{E}[(A_i - p_i)G_i]$, where $G_i = \sum_h r_{i,h}$. We use a zero shared critic; full-return gradients are unchanged by the stage-zero baseline. Since the initial state is fixed and stage one has forced actions,

$$\varepsilon = \max_i \operatorname{span}_{\mathbf{A}}[q_{i,1}(\mathbf{A}, \mathrm{noop}) - v_{i,1}(\mathbf{A})], \qquad B_{\mathrm{cred}} = \varepsilon/2.$$

Calibration is measured as a diagnostic. The primary scale is the explicitly oracle gradient projection

$$a_g = \frac{g_F^\top F^{-1} g_Y}{g_Y^\top F^{-1} g_Y}, \qquad \delta_g = \|g_F - a_g g_Y\|_{F^{-1}}.$$

A second diagnostic uses the return least-squares scale $a_{\mathrm{out}} = \mathrm{Cov}(\widehat{F}, Y)/\mathrm{Var}(Y)$, with intercept $\mathbb{E}\widehat{F} - a_{\mathrm{out}}\mu$ and its corresponding gradient residual. A nonpositive scale is inadmissible. Both use the true population outcome, and a_g additionally uses its gradient. Neither is learned from pair ordering or proposed as an operational calibration estimator.

A case is certified only when $a\sqrt{\mathcal{C}_{\mathrm{floor}}} - B_{\mathrm{cred}} - \delta_{\mathrm{cal}}$ exceeds $\varepsilon_{\mathrm{num}}[1 + |a\sqrt{\mathcal{C}_{\mathrm{floor}}}| + |B_{\mathrm{cred}} + \delta_{\mathrm{cal}}|]$. Here $\varepsilon_{\mathrm{num}}$ is the fixed arithmetic tolerance specified in the accompanying numerical record. We also evaluate the signed alignment $g_Y^\top F^{-1} g_L$ directly.

All 60 initialization cases have zero learned directions and inadmissible zero scales; undefined normalized quantities remain recorded as such. All 180 nonzero fitted cases have positive natural

Table 8: Budget components at cycle 512, over all five seeds and three policies per task. Entries are medians [minimum, maximum], not confidence intervals. Certificates use the pointwise sum; counts repeat the final row of Table 7.

Task	B_{cred}/s_g	δ_g/s_g	Certified / 15
AND4	$1.155\ [0.543, 2.466]$	$0.059\ [0.034, 0.639]$	4
OR-of-AND	$1.344\ [0.774, 1.758]$	$0.018\ [0.008, 0.470]$	2
AND-of-OR	$2.586\ [1.746, 4.128]$	$0.028\ [0.017, 0.515]$	0
Nested	$1.921\ [1.032, 2.447]$	$0.379\ [0.120, 0.500]$	0

Table 9: Every prescribed intervention at the final fitting checkpoint. Each row contains 60 cases. The six certificates at $\zeta = 0$ repeat the six final cases in Table 7. Return-scale certification is zero at every amplitude.

ζ	Positive alignment	Negative alignment	Oracle-scale certified
-2	60	0	0
-1	60	0	0
-0.5	60	0	0
0	60	0	6
0.5	60	0	0
1	57	3	0
2	1	59	0
4	0	60	0

alignment, but only six are certified under a_g: four AND4 cases and two OR-of-AND cases, all at cycle 512. No case is certified under a_{out}. Thus the criterion is sufficient and conservative. These results show limited nonvacuity on learned heads under an oracle calibration premise; they do not establish a calibration procedure or benchmark-scale benefit.

Which term limits certification? Table 8 decomposes the normalized budget for all 60 final base cases. Write $s_g = a_g\sqrt{\mathcal{C}_{\mathrm{floor}}}$; each case passes only if its own two ratios sum to less than one, subject to the stated numerical tolerance. The reported medians and ranges summarize each component separately and must not be added as if they came from the same case.

The credit-error term alone exceeds the unit budget in every AND-of-OR and nested case. Thus, even zero calibration residual would not certify those cases with the same credit bound and signal floor. The two simpler tasks include cases below the unit credit budget, where the calibration residual must still be accounted for. This describes the existing sufficient test, not a necessity result or an additional experiment.

A fixed-score intervention grid. At every final snapshot and policy, use $\zeta \in \{-2, -1, -0.5, 0, 0.5, 1, 2, 4\}$. At the public stage-one state define

$$\phi_i(\mathbf{A}) = -(A_i - p_i) + \sum_j \alpha_j (A_j - p_j).$$

Its mixture is pointwise zero. Normalize its maximum component span to the unshifted heads' scale s_{QV}, set it to zero at the other stages, and shift $q_{i,1}$ by $\zeta\phi_i$ with V fixed. The potentials are frozen before differentiation. The policy, critic, mixer and neural parameters are unchanged; this is an intervention on outputs, not a refit or policy update. The eight amplitudes were fixed before execution, and all 480 cases are retained.

The interventions preserve the mixed score, reward and fitting objectives at the declared numerical tolerances. Nevertheless, 122 local directions become negatively aligned with the true gradient; none is certified. This distinguishes score fit from local-credit accuracy without implying that ordinary fitting selects these interventions.

Fitting and independent checks. Final pair ordering is exact on three tasks; on the AND-of-OR task, accuracy is at least 0.968 across the retained seeds. These are fixed-budget fits, not convergence

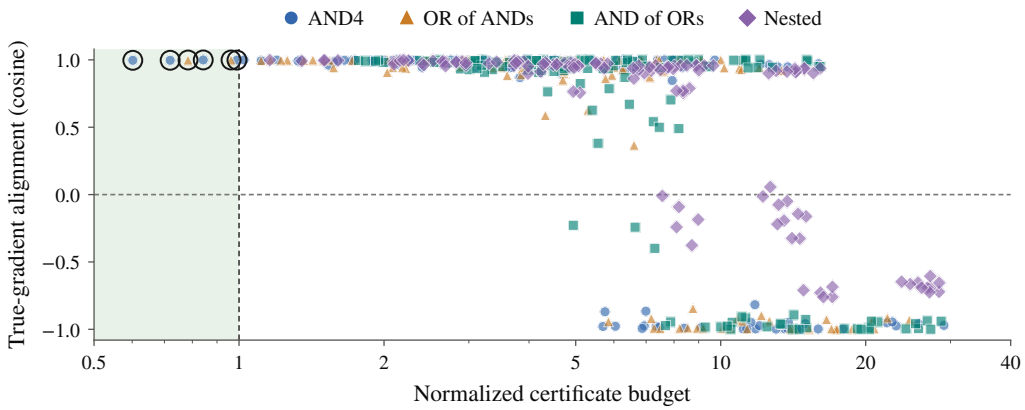

Figure 7: All 480 frozen product-action interventions. The horizontal coordinate is $(B_{\mathrm{cred}} + \delta_g)/(a_g\sqrt{C_{\mathrm{floor}}})$; the vertical coordinate is the cosine between g_Y and g_L in F^{-1} geometry. Outlined points are certified and all have positive alignment. Many points outside the certificate region are also positive: failure of the sufficient test does not imply failure of the direction. The calibration scale is oracle, and the six outlined points are the unshifted controls already counted in Table 7.

claims. Final per-task BT and Extreme-V loss ranges are in the numerical record; its linked complete data also retain every initialization case.

A separate implementation recomputes the neural outputs, gradients, spans, losses and certificates. The results agree at the declared tolerances, with no signal-floor or credit-bound violation and no certified nonpositive alignment. All snapshots and failed certificates remain in the complete record.