MEASURING COUNTERFACTUAL CREDIT IN A REPLAYABLE WILDFIRE-SUPPRESSION BENCHMARK

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting a team's return and assigning credit to individual actions are distinct tasks. FLARE, a stochastic wildfire-suppression benchmark, uses coupled-draw replay to measure immediate placement effects for training and long-horizon action contrasts for grading critics. In an eight-seed MAPPO comparison, simulator credit reduces interquartile mean burned area from 26.6% with shared reward to 20.9%. A fixed non-replay local reward scores 24.7%; simulator credit wins every paired seed against both. Informed learned credit scores 27.0% and meets a prospectively fixed ± 1-point equivalence criterion against shared reward. Four informed critics paired with responsive main-comparison actors ($0/256$ saturated audited decisions) attain factual-return $R^2 = 0.911$–0.954, yet raw credit RMSE is 8.81–15.59 times zero's. Matching critic credit to the finite replay reference's RMS still leaves squared error at 1.44–1.82 times zero's; ten earlier-rule critics show 1.80–2.24. Held-out MSE fitting instead attenuates current predictions by over 97%, approaching zero-credit error. Current critics also give heterogeneous agents nearly identical action rankings (median Spearman 0.980–0.988), with largest-magnitude credits on below-1%-probability actions in $219/256$ rows. These measured structures complement the finite-reference calibration diagnostic. Coupled replay separates feedback performance, numerical scale and action-level structure while quantifying replay precision.

1 INTRODUCTION

Cooperative multi-agent reinforcement learning gives a team one score, although individual decisions can help, duplicate a teammate's effort, or have no effect. Credit assignment seeks feedback that separates these contributions from teammates' actions and environmental randomness (Castellini et al., 2025). An agent's *counterfactual credit* compares the team outcome with a specified alternative to that agent's action or its effect.

We ask whether changing feedback improves learning and whether critics that track team returns supply calibrated, selective action credit. Within multi-agent proximal policy optimization (MAPPO), shared reward, one-step simulator credit and a fixed local reward for placements near active fire use the same actor/update settings, generalized advantage estimation (GAE) and state critic $V(x_t)$ (Yu et al., 2022). A complementary audit grades long-horizon learned credit from centralized Q-critics saved during training, following COMA (Foerster et al., 2018). Informed critics add positions, resources, timers and tactical inputs.

FLARE supplies a stochastic wildfire-suppression task with hand crews, dozers, aerial units, and rules coupling their actions. Its two replay operations answer different questions (Figure 1). Simulator credit measures the immediate effect of deleting an accepted placement (the cells treated by one agent's suppression action) and trains the policy. The replay reference estimates long-horizon, policy-centered action differences and grades learned credit against its own target.

In the main experiment, simulator credit reduces interquartile mean burned area from 26.6% with shared reward and 24.7% with a fixed non-replay local reward to 20.9%, winning all eight paired training seeds against both. Informed learned credit scores 27.0% and meets a prospectively fixed ± 1 percentage-point practical-equivalence criterion against shared reward. An offline control finds that independent random draws make the training credit about thirty times noisier.

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.

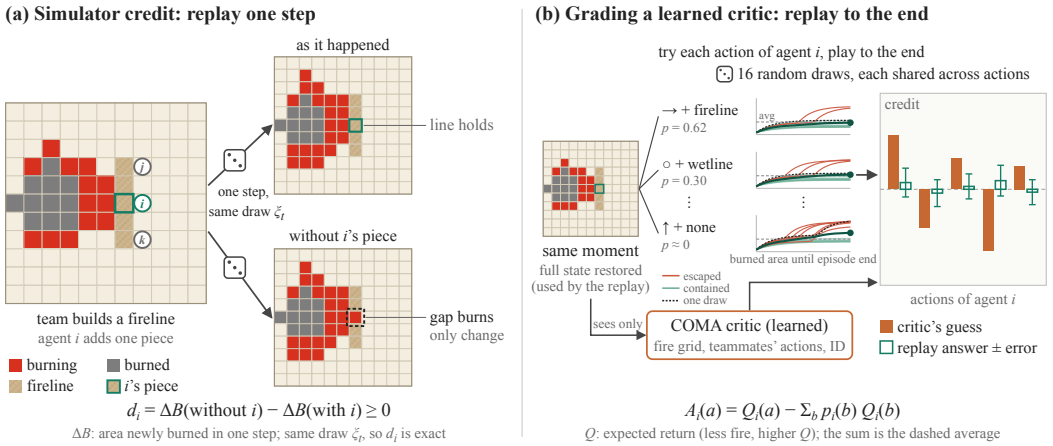


Figure 1: Two uses of replay. (a) Delete one accepted placement and replay the fire step under the same draw ξ_t; the normalized difference in newly burned area is training credit d_i (Equation 3). (b) Replay all twenty actions to episode end under sixteen independent draws, each shared across alternatives. The policy-weighted average centers the mean returns into a credit reference (Equation 5). Rows show three example actions; thin lines are sampled futures. The illustrated grid-only critic sees fire, teammate actions and identity; informed critics add tactical inputs. All grids, futures and bars are schematic.

Tracking team returns did not ensure calibrated credit. Four informed critics paired with responsive main-comparison actors reach factual-return $R^2 = 0.911$–0.954, yet raw credit RMSE is 8.81–15.59 times zero's. Matching the finite reference's RMS leaves MSE at 1.44–1.82 times zero's (1.80–2.24 across ten earlier critics). Held-out MSE fitting instead shrinks predictions by over 97%, approaching zero-credit error. Their highly similar cross-agent rankings complement these finite-reference calibration measurements. A finite-data argument explains why factual fit leaves alternative-action contrasts undetermined (Section 5).

The paper contributes:

- FLARE and distinct replay operations for immediate training feedback and long-horizon credit evaluation in a stochastic, rule-constrained team task.
- A matched eight-seed MAPPO comparison showing improved suppression with simulator credit over both shared reward and a non-replay local reward, plus an offline control measuring the noise reduction from shared random draws.
- Audits of fourteen saved critics, led by four main-comparison critics with responsive actors, separating raw calibration, RMS matching and MSE shrinkage, alongside action structure and replay precision.

Our contribution is the controlled measurement enabled by coupled-draw replay. Difference rewards, common random numbers and the estimator identities are established tools; here they separate immediate placement effects from fire randomness and make learned credit directly testable. The suppression gain is measured in a benchmark that permits cost-free ground wetline attack, whose use is illustrated in Appendix I.

2 Related Work

Difference rewards compare individual contributions with specified alternatives (Wolpert & Tumer, 1999; Castellini et al., 2025). Shapley Q-value averages across coalitions (Wang et al., 2020); FLARE deletes one accepted placement with teammates fixed. COMA subtracts a policy-weighted own-action baseline (Foerster et al., 2018); MACA combines individual, joint and correlated advantages (Zhao & Xie, 2025). VDN, QMIX and QPLEX factor joint values (Sunehag et al., 2018; Rashid et al., 2018; Wang et al., 2021a); DOP addresses centralized–decentralized mismatch (Wang

et al., 2021b). Our audit targets explicit COMA credit. Its weaker performance on many EPyMARL tasks (Papoudakis et al., 2021) motivates MAPPO against MAPPO as the headline comparison (Yu et al., 2022).

Ilyas et al. (2020) show that fitting supervised value targets need not imply accurate policy-return prediction. Multi-agent analyses identify gradient variance and state-critic bias under partial observability (Kuba et al., 2021; Lyu et al., 2023). Castellini et al. (2025) compare a trained COMA critic's action values with estimates from 100 policy-following rollouts. We grade the policy-centered differences that enter learning, which can be small relative to value-prediction errors, against replays sharing random draws across actions.

C3 restores language-agent text histories, audits counterfactual credit under a frozen policy, and reports gains over learned-critic baselines (Chen et al., 2026). Its observable text determines the restored history; expected returns are still estimated from finite rollouts. FLARE restores physical and tactical state and couples environmental and policy randomness across branches in a stochastic task with simultaneous, repeated decisions. COSAC measures advantage error in sequential bandits and cooperation in a reasoning task (Deshmukh et al., 2026); FLARE measures remaining-horizon contrasts in an evolving stochastic environment. CCPO uses agent-removal rewards (Li et al., 2026), distinct from replacing one current action while retaining the team.

Buesing et al. (2019) develop counterfactual policy search by conditioning a structural model's exogenous variables on logged experience. Oberst & Sontag (2019) study trajectory-level counterfactual policy evaluation with Gumbel-Max structural causal models. FLARE directly controls simulator randomness and pairs interventions with common random numbers, an established simulation variance-reduction technique (Glasserman & Yao, 1992). Hindsight-conditioned credit separates action effects from external variation (Mesnard et al., 2021). Unsupported-action extrapolation in batch RL connects our audit to data coverage (Fujimoto et al., 2019). Korkmaz (2026) studies distortions of counterfactual action values under adversarial training that enforces local ϵ-invariance, whereas FLARE compares each action's learned credit with a replay reference.

Wildfire RL already includes decentralized aerial firefighting (Haksar & Schwager, 2018) and GPU stochastic simulation (Çakır et al., 2025). The free-range-zoo suite includes stochastic multi-agent wildfire suppression with changing agents and tasks (Patino et al., 2025). CREW-Wildfire evaluates language-model teams with heterogeneous roles in stochastic wildfire response (Hyun et al., 2025). FLARE's focus is the controlled feedback comparison and direct grading of learned credit through explicitly defined interventions under shared random draws.

3 FLARE: Fireline Learning Agents in a Replayable Environment

A crew's move can prevent an aerial drop by entering its exclusion zone; an established fireline can make a ground placement admissible. Credit therefore depends on tactical variables as well as fire. Replay must restore every variable the rules read.

Task and state. The main task uses a 64×64 grid, four hand crews, two dozers and two aerial units. A randomly located 5×5 ignition starts each episode, which lasts at most 100 steps and ends earlier if no active fire or usable fuel remains. A stochastic cellular automaton tracks burning, cumulative burning, fuel and mitigation. Randomized wind speed raises spread; wind direction is unused.

For N agents, let $s_t = (x_t, \tau_t)$ contain the fire grid x_t and tactical state τ_t: positions, types, water, restriction timers, previous moves, activity and time. Agent i observes $o_{it} = O_i(s_t)$ and acts according to

$$h_{it} = (o_{i0}, a_{i0}, \ldots, a_{i,t-1}, o_{it}), \qquad \pi(\mathbf{a}_t \mid \mathbf{h}_t) = \prod_{i=1}^{N} \pi_i(a_{it} \mid h_{it}). \tag{1}$$

Here h_{it} is its local history; bold symbols collect agents. Implemented actors share parameters and use only current observations. These contain own-centered 21×21 fire crops, a pooled fire summary, own type and aerial resource/restriction channels, with no explicit teammate positions or resources.

Actions and rules. Each action $a_{it} = (m_{it}, u_{it}) \in \mathcal{A}$ combines five movements (stay, up, down, left, right) and four suppression choices: none, a type-specific primary intervention, one-cell wetline or one-cell scratchline, giving $|\mathcal{A}| = 20$. Agents move before placement acceptance, accepted effects are combined, then fire spreads. Ground primary interventions remove fuel in a one-cell hand-crew or up-to-three-cell dozer fireline; they require a boundary or an existing burned/mitigated neighbor. Aerial primary drops wet up to 3×3 cells for one transition and require water, inactive refill/no-fly timers, and no ground unit within Chebyshev distance one. No-fly periods are random intervals when aerial suppression is blocked.

Every aerial suppression action passes the water, refill, no-fly and ground-deconfliction checks and consumes one water unit when accepted. Ground one-cell wetline and scratchline remain unanchored direct attack: wetline wets one cell for one transition, and scratchline halves its fuel. Ground wetline has no modeled water cost and extinguishes burning at its cell. The earlier performance and transfer comparisons, ten consistency audits, and offline, tuning and no-clamp controls use earlier rules in which aerial one-cell wetline and scratchline bypassed these checks and used no water (Appendix A.2). Detailed timing, footprints and parameters appear in Appendix A.

Objective and replay. Let $B_t = \sum_c b_t(c)$ sum the fractional cumulative burned channel over $C = 64^2$ cells. All agents receive the same environment reward, and every method is evaluated by

$$r_t = -\frac{B_{t+1} - B_t}{C}, \qquad \text{BurnedFraction} = \frac{B_T}{C}. \tag{2}$$

Undiscounted return telescopes to $-(B_T - B_0)/C$; discounting changes this objective. A complete replay context is $z_t = (s_t, \mathbf{h}_t)$, together with restored implementation bookkeeping and random-generator state. Fixing all fire, no-fly and policy-sampling draws ω makes the continuation deterministic. Branches use *common random numbers*, but policies respond to their own branch observations. The replay reference rechecks the crew–aircraft interaction; placement-deletion training holds acceptance fixed. Sections 4–5 define these interventions precisely.

4 Credit Signals and Policy Learning

All four training arms share the actor architecture, Adam actor-optimizer settings, budget and logit clamp. Shared versus simulator credit isolates reward replacement within MAPPO; learned credit also changes the critic and advantage construction (Table 1).

Shared reward. Every agent receives $r^{\text{sh}}_{it} = r_t$, the same value estimate $V(x_t)$ and termination mask. Thus pre-normalization generalized advantage estimation (GAE) gives $\widehat{A}^{\text{sh}}_{it} = \widehat{A}^{\text{sh}}_{jt}$ (Schulman et al., 2016).

Signal	What "without my action" means	Exact or learned?	What the critic sees
Shared reward	No removal; common team reward	Observed reward; learned value baseline	Fire grid
Simulator credit	Delete accepted placement; hold movement and acceptance fixed	Exact realized effect; learned value baseline	Fire grid
Local reward	No removal; count accepted cells near fire	Observed local activity; learned value baseline	Fire grid
Learned credit	Policy-weighted own-action alternatives	Learned long-horizon contrast A^c	Grid, teammate actions, identity; informed adds tactical features

Table 1: The four feedback signals: the alternative each compares against, whether it is computed by replay or learned, and what the critic sees. Shared, simulator and local rewards enter PPO through returns and GAE; learned credit is a critic's long-horizon advantage.

Simulator credit. Let $P_t = (\ell_{1t}, \ldots, \ell_{Nt})$ collect accepted placements, with empty entries for rejected requests, and let P_t^{-i} delete agent i's placement. Let F reconstruct mitigation/fuel effects and advance fire once, using the fire draw ξ_t, the step-t fire component of ω. With Section 3's grid x_t, burned total B and cell count C, define

$$r(x, P, \xi) = -\{B(F(x, P, \xi)) - B(x)\}/C,$$
$$d_{it}^{\text{place}} = r(x_t, P_t, \xi_t) - r(x_t, P_t^{-i}, \xi_t). \tag{3}$$

Movement, acceptance decisions, other agents' placements and the fire draw stay fixed. Absent or redundant placements earn zero; identical placements are redundant under cellwise maximum mitigation and minimum fuel multipliers, while partial overlap need not be. For the burned-area-only reward and recorded suppression update, $d_{it}^{\text{place}} \geq 0$ up to roundoff: mitigation and fuel removal cannot increase same-draw burning. GAE can reward moves enabling later useful placements.

Local reward. Count unique accepted placement cells within Chebyshev distance one of pre-step burning, divide by $C = 64^2$, and feed MAPPO returns and GAE without replay. Overlapping agents each receive their own count.

Learned credit. The base critic receives $\chi_{it} = (x_t, \mathbf{a}_{-i,t}, i)$: fire grid, teammate action codes and focal identity. Let $q_\phi(\chi_{it}, b)$ be its value prediction after undoing target normalization, in reward-standardized training units. Its COMA-style credit is

$$A_{it}^c = q_\phi(\chi_{it}, a_{it}) - \sum_{b' \in \mathcal{A}} \pi_i(b' \mid h_{it}) q_\phi(\chi_{it}, b'). \tag{4}$$

Only the selected action receives a bootstrapped team-return label; alternatives can learn through selection elsewhere and shared parameters, but receive no counterfactual label from that transition. Informed critics add positions, resources, timers and other tactical features.

Training settings. All three MAPPO arms use GAE and a fire-grid value critic; reward replacement also changes value targets and the objective (Yu et al., 2022). Learned credit supplies A^c to PPO (Foerster et al., 2018; Schulman et al., 2017). Training uses reward standardization and PopArt-style target normalization (van Hasselt et al., 2016). Actors clamp raw action scores (logits) to $[-10, 10]$ before forming probabilities. We call a decision *fully saturated* when every raw logit lies strictly outside these bounds: its local policy and entropy gradients vanish. An actor is *saturated* when every decision on its checked final-policy trajectories is fully saturated. Section 6.4 tests this cap, separately from PPO ratio and gradient clipping.

5 Grading Learned Credit with a Replay Reference

The audit asks whether strong prediction of team returns also gives calibrated and selective action-level credit. It compares a saved Q-critic with a matched long-horizon replay target. This is complementary to Section 6.2's intervention on the immediate reward used by MAPPO.

Target and reference. Fix context z_t, joint action $\mathbf{a}_t$, frozen policy π and remaining horizon H. For training discount γ, write $\mathbf{a}_t^{i \leftarrow b} = (\mathbf{a}_{-i,t}, b)$, $p_i(b) = \pi_i(b \mid h_{it})$ and $G_H^\pi(z, \mathbf{a}; \omega) = \sum_{\ell=0}^{H-1} \gamma^\ell r_{t+\ell}$, with zero rewards after termination. Then

$$Q_i^\pi(b) = \mathbb{E}_\omega G_H^\pi(z_t, \mathbf{a}_t^{i \leftarrow b}; \omega), \qquad A_i^\pi(b) = Q_i^\pi(b) - \sum_{b' \in \mathcal{A}} p_i(b') Q_i^\pi(b'). \tag{5}$$

This is Equation (4)'s contrast using complete-context expected returns. We replay all twenty alternatives with teammates' current actions fixed, re-apply the rules, and follow each branch's observations. For $M = 16$ independent replicates, each sharing environmental and policy draws across alternatives, define

$$Z_{i,m}(b) = G_H^\pi(z_t, \mathbf{a}_t^{i \leftarrow b}; \omega_m) - \sum_{b' \in \mathcal{A}} p_i(b') G_H^\pi(z_t, \mathbf{a}_t^{i \leftarrow b'}; \omega_m), \qquad \widehat{A}_i(b) = \frac{1}{M} \sum_m Z_{i,m}(b). \tag{6}$$

Centering each replicate cancels return variation common to all actions. The standard Monte Carlo identities give unbiased estimates with variance proportional to $1/M$ under the specified replay distributions (Appendix G); their Monte Carlo standard errors use the sample variance of $Z_{i,m}$.

Why factual fit is insufficient. For fixed regression targets and an unrestricted function class, the predicted value of an unobserved alternative with positive policy probability can change freely without altering factual predictions, and each change shifts the credit (Appendix G). The audit measures the contrasts that parameter-sharing, bootstrapped neural critics learn in practice.

Grading. Stored normalizers convert critic predictions to raw reward units. Here $A_i^c(b)$ denotes the resulting raw-scale contrast for alternative b. *Factual* R^2 and root mean squared error (RMSE) compare predictions for actions actually taken with their mean replayed returns. Credit error and its post hoc difference from a zero-credit reference are

$$\mathrm{MSE}_{\mathrm{credit},i} = \sum_b p_i(b)[A_i^c(b) - \widehat{A}_i(b)]^2, \tag{7}$$

$$\Delta_i = \sum_b p_i(b)\{[A_i^c(b) - \widehat{A}_i(b)]^2 - \widehat{A}_i(b)^2\}. \tag{8}$$

Positive Δ_i establishes greater raw numerical credit error than zero. We average agents within each context, then contexts, separately by checkpoint, and repeat with uniform action weights. Appendix E reports whole-episode bootstrap intervals conditional on the stored replay means and separate Monte Carlo errors.

Calibration and direction. PPO batch standardization preserves the standardized signal under a common positive scaling, up to numerical stabilization; it does not shrink that signal to zero. We distinguish two diagnostics. Let W average contexts and agents with saved-policy action weights, $R_c^2 = W[(A^c)^2]$, $R_A^2 = W[\widehat{A}^2]$ and $\rho_{c,A} = W[A^c\widehat{A}]/(R_c R_A)$. Reference-RMS matching, $\widetilde{A}^c = (R_A/R_c)A^c$, gives

$$\frac{W[(\widetilde{A}^c - \widehat{A})^2]}{W[\widehat{A}^2]} = 2(1 - \rho_{c,A}). \tag{9}$$

This removes global scale inflation while matching the finite reference's signal energy. Separately, a nonnegative MSE multiplier is fitted on seven contexts and evaluated on the eighth (Appendix H). Ranks and independent-half checks describe action structure. Frozen actor-score gradients compare credit-weighted log-policy derivatives, holding probabilities and credits fixed; their cosine, Monte Carlo error and split-half agreement assess local alignment, separately from PPO updates (Equation 16).

6 Experiments and Results

6.1 Setup

Shared-reward, simulator-credit and local-reward MAPPO, plus informed learned credit, each use eight seeds (1–8), completing 732 updates and 2,998,272 agent-steps. Frozen policies take argmax actions on the same thirty fresh fires, none used in training. We average each policy's fire outcomes and report the interquartile mean (IQM) of the eight policy scores (Agarwal et al., 2021). The 95% intervals use 10,000 training-seed bootstrap resamples, paired for contrasts. Appendix A gives training settings and Appendix B gives evaluation and sensitivity checks.

6.2 Does simulator credit improve learning?

Yes, in matched MAPPO training. One-step simulator credit reduces IQM burned area from 26.6% to 20.9%, winning all eight seed pairs (Figure 2). The shared-minus-simulator IQM difference is 5.76 percentage points (95% interval $[3.65, 7.73]$), a 21.7% relative improvement.

A simple local reward helps, but simulator credit improves further. Both arms can use cost-free, unanchored ground wetline. The local control directly rewards accepted suppression cells within Chebyshev distance one of pre-step burning, including wetlines, without replay. It reaches 24.7% IQM and beats shared reward in all eight pairs; simulator credit (20.9%) beats local reward in all eight, by a paired mean 3.898 points (95% paired-t interval $[2.117, 5.678]$; Appendix B). Thus the gain exceeds this wetline-accessible proximity reward. Local reward counts overlapping placements

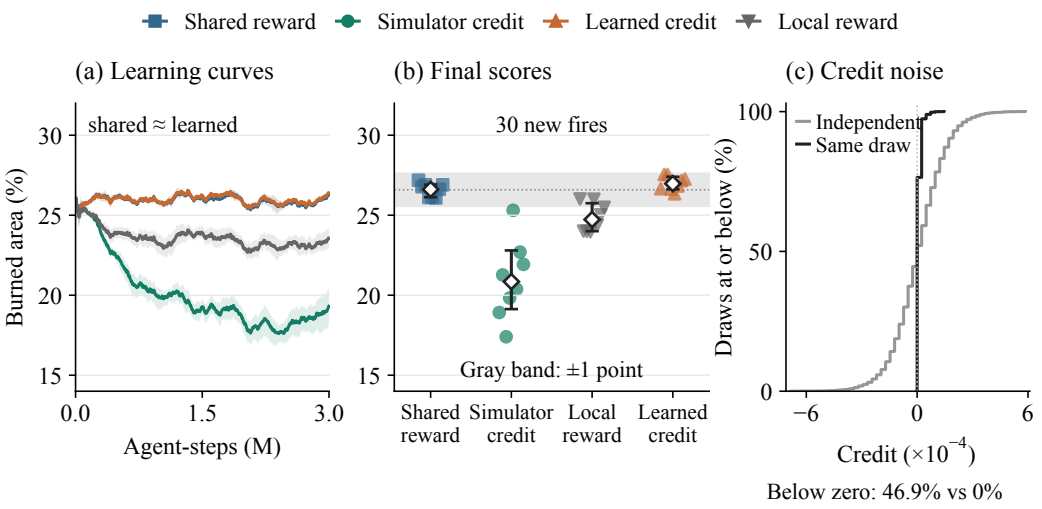


Figure 2: Simulator credit improves suppression. (a) Four-arm training means over eight seeds, with 95% intervals. (b) Thirty-fire seed means (dots), IQMs with 95% intervals (diamonds), and the ±1-point band around shared reward. (c) Earlier-rule one-step credit CDFs: independent draws give negative credit in 46.9% of samples; shared draws never do (Appendix D).

separately and rewards proximity even when a placement saves no area in that transition; simulator credit measures incremental burned-area reduction.

Learned credit, using informed critics, scores 27.0% and loses all eight pairs to simulator credit. Its mean difference from shared reward is 0.405 percentage points (90% paired-t interval $[0.119, 0.690]$), inside the ±1-point equivalence margin fixed before inspecting these outcomes. Both one-sided tests pass at $\alpha = .05$ (Schuirmann, 1987). This establishes practical equivalence, with a small estimated advantage to shared reward.

Four of eight learned-credit actors (seeds 2, 3, 5 and 6) are saturated on their checked final-policy trajectories. In a post hoc descriptive comparison, the responsive four (seeds 1, 4, 7 and 8) average 26.9% burned area, versus 26.7% for their paired shared-reward seeds (Appendix F). The six earlier-rule no-clamp runs also fail to improve either critic group's mean evaluation score (Table 7).

An earlier-rule comparison also favors simulator credit in all eight seed pairs and over the two other trained arms across five simulated scenarios (Appendix B). On the preselected fire in Figure 3, simulator-credit ground crews favor immediate wetline attack.

6.3 Does learned credit match the replay?

Current-rule critics track returns but miscalibrate credit. We audit four informed critics paired with locally responsive actors from Section 6.2 (seeds 1, 4, 7, 8) under current rules; none of their 256 audited agent/context decisions is fully saturated. Factual-return $R^2 = 0.911$–0.954, while raw credit RMSE is 8.81–15.59 times zero's (77.6–243.2 in MSE), with four strictly positive conditional 95% episode intervals for Δ (Figure 4; Table 9). Ten earlier-rule critics show the same raw-error ordering under policy and uniform weights, including three informed critics with $R^2 = 0.957$–0.970 and two base critics paired with responsive actors, with RMSE 9.09 and 7.01 times zero's.

RMS matching and MSE shrinkage answer different questions. Matching the finite reference's RMS leaves credit MSE at 1.440–1.818 times zero's (earlier: 1.797–2.235). Current critic/reference cosines are $+0.091$–$+0.280$. Held-out MSE fitting instead uses multipliers 0–0.0238, attenuating raw predictions by 97.62–100%, and reaches 0.994–1.044 times zero's MSE (earlier: 1.0000–1.0073). All fourteen fitted-error intervals include or touch zero. This attenuation approaches the zero predictor; it is distinct from PPO standardization (Appendix H).

Critic rankings are highly similar across agents. Median cross-agent Spearman correlations are 0.980–0.988 (earlier informed: 0.997), and 219/256 rowwise largest-magnitude credits concern

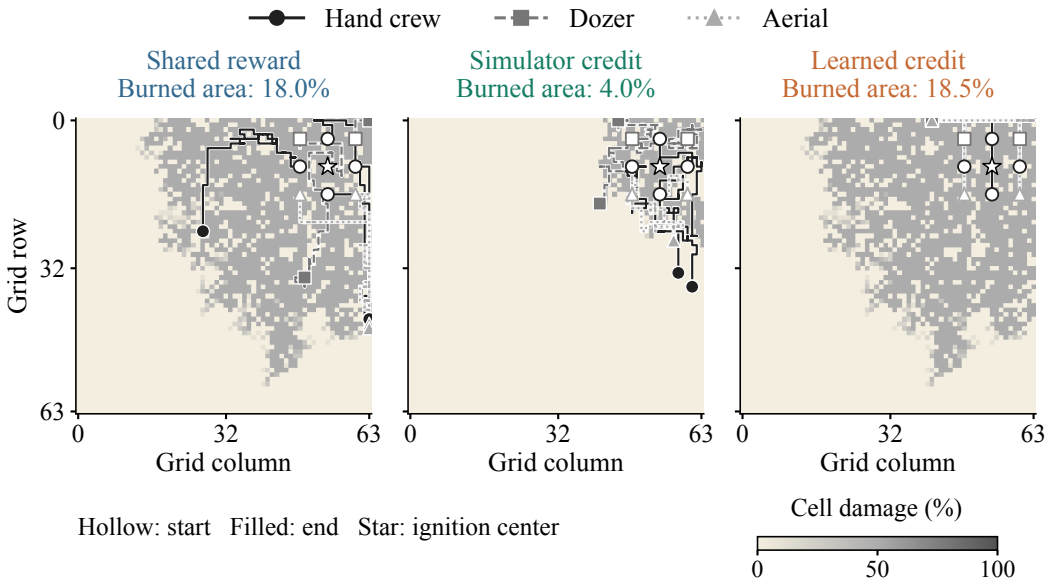


Figure 3: One preselected fire (training seed 1, first fresh evaluation fire): final cell damage (gray) and agent paths under each method. Simulator-credit crews stay at the fire's edge and the fire burns out at step 69 (4.0% burned); under shared reward and learned credit it spreads until the 100-step limit (18.0% and 18.5% burned).

actions with saved-policy probability below 1%. These critic-only summaries require no rollout estimates. Separately, 19–47 of each critic's 64 agent/context rows have equal action returns within every sampled draw; all agents are flat in ten of 32 contexts. Figure 4b shows one such context with absolute critic credit reaching 6.11×10^{-3}.

Long-horizon reference precision. With uniform post-burn-in sampling ($t \geq 10$), seven of eight registered query steps fall at $t \geq 55$. Finite-reference credit RMS is 0.072–0.171×10^{-3}, comparable to sixteen-draw Monte Carlo error RMS 0.102–0.159×10^{-3} (Table 9). These late-context measurements motivate testing horizon-dependent signal dilution. Appendix E reports reference agreement; the raw and RMS-matched errors above measure calibration against the finite reference, while cross-agent ranking similarity is computed solely from the critic.

6.4 Ablations and Controls

Shared random draws reduce credit noise. In an offline earlier-rule control on 32 saved contexts, independent fire draws increase one-step credit variance 934-fold, or standard deviation 30.6-fold (Figure 2c). Independent-draw credit is below zero in 46.9% of the saved samples, versus none with shared draws (Appendix D).

Replacing accepted-placement deletion with either of two no-suppression action baselines leaves immediate credit unchanged in all 256 sampled agent/context rows while changing some tactical states (Appendix D). None of twelve tested critic settings improves on shared reward in the late-training mean. Removing the logit clamp improves neither critic group's mean evaluation score. The tuning and paired no-clamp controls are detailed in Appendix C.

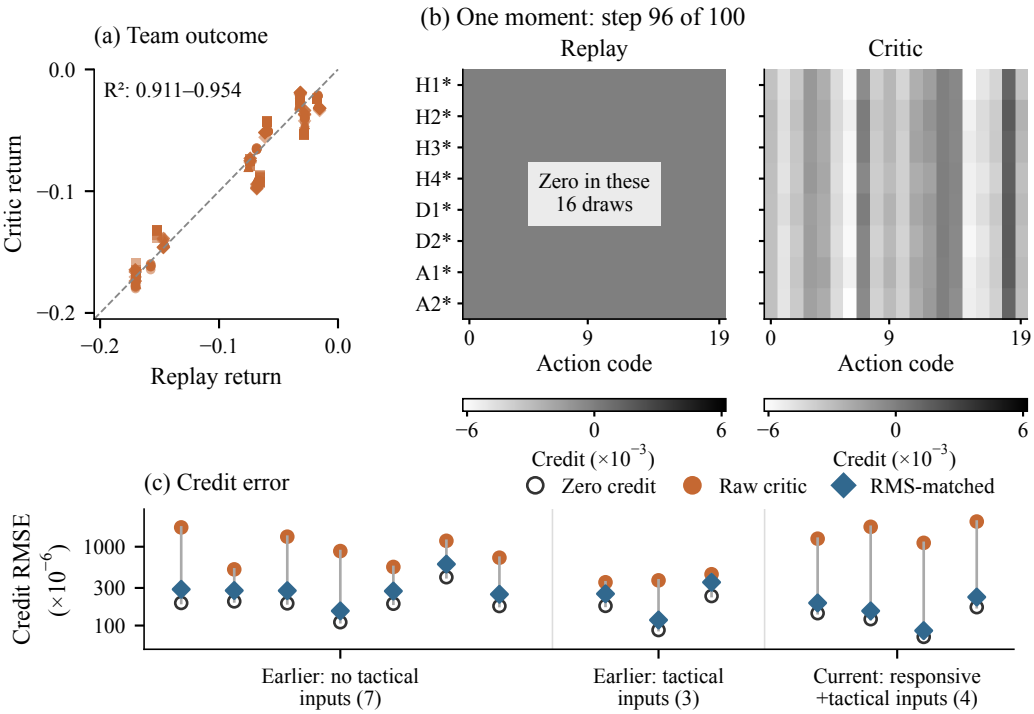


Figure 4: Return fit and credit calibration. (a) Taken-action critic/replay returns; shapes identify the four current seeds. (b) A fixed context: replay credit is zero in all sixteen draws (stars), while critic credit reaches 6.1×10^{-3}. (c) Raw critic RMSE (orange circles), reference-RMS-matched RMSE (blue diamonds) and zero-credit RMSE (open circles) for fourteen checkpoints. Both critic errors exceed zero's; held-out shrinkage and reference precision are reported in Appendices E and H.

7 Discussion and Conclusion

Simulator credit reduces MAPPO's IQM burned area from 26.6% with shared reward and 24.7% with a fixed non-replay local reward to 20.9%, winning every paired seed against both, while the implemented learned-credit arm meets the ±1-point practical-equivalence criterion against shared reward. Audits of four informed critics paired with responsive main-comparison actors show that strong factual return tracking ($R^2 = 0.911$–0.954) coexists with severe raw scale inflation (8.81–15.59 times zero's RMSE) and nearly identical action rankings across heterogeneous agent roles (median Spearman 0.980–0.988). Across all fourteen critics, including ten earlier-rule audits with R^2 up to 0.970, matching the replay reference's RMS still leaves squared credit error above zero's (1.44–2.24 times), whereas cross-validated nonnegative rescaling removes excess error only by shrinking predictions toward zero. FLARE thus supports two complementary replay measurements: immediate placement effects for policy learning and graded learned action contrasts.

Scope and future directions. These findings concern MAPPO feedback replacement and the evaluated COMA-style critic in FLARE. Because ground wetline has no modeled water cost, resource constraints and further local-reward variants would separate counterfactual deletion from local feedback density. Sampled long-horizon action contrasts have substantial sixteen-draw uncertainty, motivating larger replay budgets and replay-supervised critic training.

Reproducibility Statement

Section 3 defines the environment and Table 4 maps each reported study to its rules and sample counts. The main comparison archives source/configuration records, complete training-reset ledgers, checkpoints and fixed fresh-fire evaluations. Earlier comparisons retain actor checkpoints, evaluation arrays and training summaries; logs and per-run summaries document the auxiliary controls. The online audits reuse four main-comparison actor/Q pairs selected for actor responsiveness and ten earlier-rule pairs, with eight contexts and sixteen replays per action. They archive checkpoints, saved normalizers and replay arrays; scale and frozen-gradient diagnostics additionally bind cross-validation, resampling and gradient outputs. Figure sources bind saved arrays and logs by input hashes. Sections 4 and 6.1 define learning and aggregation; Appendices A and E provide settings, sampling units and uncertainty calculations.

Ethics Statement

The simulator is an abstract research environment. Its policies have not been validated for deployment or incident command, and the results should not guide real suppression operations.

AI Use Statement

Generative AI tools assisted with literature retrieval, code inspection and implementation, experiment and diagnostic design, interpretation of recorded results, formulation and checking of mathematical arguments, figure-generation code, and manuscript drafting and editing. Numerical figures in this paper are computed from recorded experiments. The use of automated checks and source comparisons does not replace the authors' responsibility for the final claims and submission.

References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2108.13264.

Lars Buesing, Theophane Weber, Yori Zwols, Nicolas Heess, Sebastien Racaniere, Arthur Guez, and Jean-Baptiste Lespiau. Woulda, coulda, shoulda: Counterfactually-guided policy search. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJG0voC9YQ.

Ufuk Çakır, Victor-Alexandru Darvariu, Bruno Lacerda, and Nick Hawes. JaxWildfire: A GPU-accelerated wildfire simulator for reinforcement learning. *arXiv preprint arXiv:2512.06102*, 2025. URL https://arxiv.org/abs/2512.06102.

Jacopo Castellini, Sam Devlin, Frans A. Oliehoek, and Rahul Savani. Difference rewards policy gradients. *Neural Computing and Applications*, 37:13163–13186, 2025. doi: 10.1007/s00521-022-07960-5. URL https://link.springer.com/article/10.1007/s00521-022-07960-5. First published online in 2022.

Yanjun Chen, Yirong Sun, Hanlin Wang, Jinghan Wang, Xinming Zhang, Xiaoyu Shen, Wenjie Li, and Wei Zhang. Exact is easier: Credit assignment for cooperative LLM agents, 2026. URL https://arxiv.org/abs/2603.06859v2.

Shripad Deshmukh, Jayakumar Subramanian, Raghavendra Addanki, and Nikos Vlassis. COSAC: Counterfactual credit assignment in sequential cooperative teams, 2026. URL https://arxiv.org/abs/2604.17693v2.

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. doi: 10.1609/aaai.v32i1.11794. URL https://ojs.aaai.org/index.php/AAAI/article/view/11794.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2052–2062, 2019. URL https://proceedings.mlr.press/v97/fujimoto19a.html.

Paul Glasserman and David D. Yao. Some guidelines and guarantees for common random numbers. *Management Science*, 38(6):884–908, 1992. doi: 10.1287/mnsc.38.6.884. URL https://pubsonline.informs.org/doi/10.1287/mnsc.38.6.884.

Ravi N. Haksar and Mac Schwager. Distributed deep reinforcement learning for fighting forest fires with a network of aerial robots. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1067–1074. IEEE, 2018. doi: 10.1109/IROS.2018.8593539. URL https://msl.stanford.edu/bibliography/haksar_distributed_2018.

Jonathan Hyun, Nicholas R. Waytowich, and Boyuan Chen. CREW-Wildfire: Benchmarking agentic multi-agent collaborations at scale. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=8mr27qFzKR.

Andrew Ilyas, Logan Engstrom, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. A closer look at deep policy gradients. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxdEkHtPS.

Ezgi Korkmaz. How to lose inherent counterfactuality in reinforcement learning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=2kutK2Y8Sv.

Jakub Grudzien Kuba, Muning Wen, Linghui Meng, Shangding Gu, Haifeng Zhang, David Henry Mguni, Jun Wang, and Yaodong Yang. Settling the variance of multi-agent policy gradients. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13458–13470, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/6fe6a8a6e6cb710584efc4af0c34ce50-Abstract.html.

Zhongyi Li, Wan Tian, Jinju Chen, Huiming Zhang, Yang Liu, Yikun Ban, and Fuzhen Zhuang. Counterfactual credit policy optimization for multi-agent collaboration, 2026. URL https://arxiv.org/abs/2603.21563v5.

Xueguang Lyu, Andrea Baisero, Yuchen Xiao, Brett Daley, and Christopher Amato. On centralized critics in multi-agent reinforcement learning. *Journal of Artificial Intelligence Research*, 77: 295–354, 2023. doi: 10.1613/jair.1.14386. URL https://www.jair.org/index.php/jair/article/view/14386.

Thomas Mesnard, Theophane Weber, Fabio Viola, Shantanu Thakoor, Alaa Saade, Anna Harutyunyan, Will Dabney, Thomas S. Stepleton, Nicolas Heess, Arthur Guez, Eric Moulines, Marcus Hutter, Lars Buesing, and Remi Munos. Counterfactual credit assignment in model-free reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7654–7664, 2021. URL https://proceedings.mlr.press/v139/mesnard21a.html.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with Gumbel-Max structural causal models. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4881–4890. PMLR, 2019. URL https://proceedings.mlr.press/v97/oberst19a.html.

Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/a8baa56554f96369ab93e4f3bb068c22-Abstract-round1.html.

Ceferino Patino, Tyler J. Billings, Alireza Saleh Abadi, Daniel Redder, Adam Eck, Prashant Doshi, and Leen-Kiat Soh. Inaugural MOASEI competition at AAMAS'2025: A technical report. *arXiv preprint arXiv:2507.05469*, 2025. URL https://arxiv.org/abs/2507.05469.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4295–4304, 2018. URL https://proceedings.mlr.press/v80/rashid18a.html.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15:657–680, 1987. doi: 10.1007/BF01068419. URL https://link.springer.com/article/10.1007/BF01068419.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1506.02438.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2085–2087. International Foundation for Autonomous Agents and Multiagent Systems, 2018. URL https://www.ifaamas.org/Proceedings/aamas2018/pdfs/p2085.pdf. Extended abstract.

Hado van Hasselt, Arthur Guez, Matteo Hessel, Volodymyr Mnih, and David Silver. Learning values across many orders of magnitude. In *Advances in Neural Information Processing Systems*, 2016. URL https://arxiv.org/abs/1602.07714.

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. URL https://openreview.net/forum?id=Rcmk0xxIQV.

Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, and Yunjie Gu. Shapley Q-Value: A local reward approach to solve global reward games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7285–7292, 2020. doi: 10.1609/aaai.v34i05.6220. URL https://ojs.aaai.org/index.php/AAAI/article/view/6220.

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://openreview.net/forum?id=6FqKiVAdI3Y.

David H. Wolpert and Kagan Tumer. An introduction to collective intelligence. *arXiv preprint cs/9908014*, 1999. URL https://arxiv.org/abs/cs/9908014.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9c1535a02f0ce079433344e14d910597-Abstract.html.

Xutong Zhao and Yaqi Xie. Multi-level advantage credit assignment for cooperative multi-agent reinforcement learning. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2971–2979, 2025. URL https://proceedings.mlr.press/v258/zhao25c.html.

A Environment, training and evaluation configuration

Setting	Value
Training scenario / seeds per main arm	S1 / 8
Grid / roster / maximum episode steps	64×64 / 4 hand crew, 2 dozer, 2 aerial / 100
Parallel environments / rollout length	4 / 128
Agent-steps per update / number of updates	4096 / 732
Actor and base critic hidden width	128
Actor / critic learning rate	5×10^{-4} / 5×10^{-4}
Discount / GAE or TD mixing	0.99 / 0.95
PPO clip / entropy coefficient	0.2 / 0.01
Actor logits in the main runs	Hard clamp to $[-10, 10]$
Minibatch / maximum PPO epochs	64 / 4
KL early-stop target / gradient clipping	0.01 / 10
Reward normalization	Running standard deviation
Evaluation / fires per policy	Argmax actions / 30
Reported metric	Episode-final burned fraction

Table 2: Mixed-team training and evaluation settings; the no-clamp control changes the stated actor-logit bound. Actor settings are shared across arms; V and Q critics and their targets differ as described in the main text.

Environment setting	Main mixed-team configuration (S1)
Grid / episode limit	64×64 cells / 100 steps
Roster	Four hand crews, two dozers, two aerial units
Initial ignition	One randomly located 5×5 square; radius parameter 2
Initial fuel / slope	Fuel 1 per cell / slope 0
Agent placement	Near the ignition, spawn-offset parameter 6
Local actor input	Own-centered 21×21 burning, mitigation, and fuel crops
Other actor inputs	Coarse burning mean and uncertainty; own type, water fraction, refill flag, global no-fly flag
Ensemble Kalman filter	Disabled; belief mean equals simulated state and uncertainty is zero
Action labels	Five movements $\times$ four suppression primitives $= 20$
Ground / aerial constraints	Enabled / enabled
Aerial water capacity / debit	8 units / 1 per accepted gated placement
Refill cooldown	5 blocked steps including the empty-tank request that triggers refill
No-fly onset probability	0.1, sampled only if the timer is inactive at step entry
No-fly duration	Uniform integer in $\{3, 4, 5, 6\}$; onset step counts toward duration
Ground–aerial separation	Chebyshev radius 1 around the aerial target center
Aerial primary footprint	Patch radius 1; up to 3×3 distinct cells
Dozer primary footprint	Line length 3 in movement direction; staying affects one cell
Wet mitigation	Strength 0.6 for one fire transition; cleared afterwards
Scratchline effect	Multiply remaining fuel by 0.5
Ground fireline effect	Mitigation strength 1; multiply remaining fuel by 0
Mitigation mode / reward shaping	Post-threshold mode / additional shaping disabled
Wind in spread calculation	Magnitude only; no directional transport
Aerial suppression checks	Apply to every suppression primitive; one water unit per accepted placement

Table 3: Main S1 settings; scenario variants are stated separately. Action semantics are defined in Section 3. An empty-tank request reaching the water check starts refill; water returns after five blocked steps. Footprints are clipped at boundaries. Wet suppression clears pre-existing fireline markings on its footprint after the transition while preserving zero fuel, so neighboring ground targets can lose that source of anchoring.

A.1 Transition and replay

The simulator updates time/no-fly state, moves agents, accepts placements, combines their effects and advances fire. Combination uses maximum mitigation, minimum fuel multipliers and the union of wet footprints; after fire spread, wet mitigation clears and refill timers tick. A ground primary fireline needs a boundary or an eight-neighbor whose pre-transition burned/mitigation channel exceeds 0.5, so a same-step placement cannot anchor another crew. Ground wetline and scratchline are unanchored; wetline has no modeled water cost and extinguishes its cell during the transition. Aerial acceptance requires inactive no-fly/refill timers, positive water and no post-movement ground position within Chebyshev distance one. Full-action replay rechecks acceptance; placement deletion holds it fixed. Complete restoration includes tactical bookkeeping and RNG states. Given a frozen policy and underlying environmental/policy draws, each replay is deterministic, with branch-specific observations and zero rewards after termination.

A.2 Experiment inventory and rule sets

Current rules apply all aerial checks and water debits to every suppression primitive. Earlier rules applied them to the primary patch drop; one-cell wetline and scratchline were exempt. Ground rules are unchanged. Every evaluation and audit matches its policy's training rules. Training seeds identify policies; episodes, actions and replicates are repeated measurements within them.

Study	Rules	Purpose and sample units
Main learning + local control	Current	8 seeds per 4 methods, 30 fresh fires each; local CPU, other arms CUDA. Profiles reuse evaluations; paths/action mix reuse seed 1 on the first fire.
Scenario transfer	Earlier	24 policies (8 seeds per 3 methods), 5 scenarios, 30 fires each; S1 reused. Heuristic/no-suppression references: 30 fires per scenario.
Responsive main audit	Current	Reuse learned-credit seeds 1, 4, 7, 8; 8 contexts, 16 replays per action per critic.
Consistency / information	Earlier	10 trained critic/actor pairs: 3 base, 1 tuned c08, 3 informed, 3 capacity-matched zero-input; 8 contexts, 16 replays per action each.
Audit sensitivity	Both	Reuse all 14 audits for scale, weighting and reference checks; frozen gradients use the current 4. Checkpoints analyzed separately.
Critic tuning	Earlier	12 configurations $\times$ 3 seeds; final-ten-update training means.
No-clamp training	Earlier	6 new runs paired with the 3 informed and 3 zero-input runs; frozen-policy evaluation.
Randomness / removal	Earlier	Reuse 8 simulator actors, 4 contexts each; 256 paired draw sets per context; 2 action-replacement baselines on 256 agent/context rows.
Actor saturation	Both	Reuse 16 earlier MAPPO, 8 current learned-credit and 8 local actors; 4 trajectories and selected gradient contexts per actor.

Table 4: Reported experiments and policy reuse. Full mixed-team training runs complete 732 updates and 2,998,272 agent-steps.

B Performance and scenario transfer

All current policies use argmax actions on fires 2026092400–2026092429, excluded by the saved training-reset ledgers. Each seed's score averages thirty fires; IQM averages the middle four of eight seed scores. Unless stated otherwise, 95% intervals use 10,000 whole-seed bootstrap draws, paired for contrasts and conditional on these fixed fires.

Practical equivalence. The ± 0.01 margin for learned minus shared was locked before inspecting current outcomes. The paired-t analysis uses eight seed differences, mean 0.004046, SD 0.004265, SE 0.001508 and seven degrees of freedom. Its 90% interval is $[0.001190, 0.006903]$; one-sided

Method	Burned fraction (IQM)	Descriptive 95% interval
Shared reward (MAPPO)	0.2661	[0.2613, 0.2691]
Simulator credit (MAPPO)	0.2085	[0.1913, 0.2280]
Informed learned credit	0.2698	[0.2662, 0.2740]
Local reward	0.2473	[0.2400, 0.2575]

Table 5: S1 performance: eight training seeds per arm, each evaluated on the same thirty fresh fires. Lower is better. Intervals resample training seeds and condition on the fixed fires. MAPPO feedback arms use shared, simulator or fixed local rewards; learned credit uses an informed Q critic. All arms retain the actor-logit clamp.

Seed	Shared	Simulator	Informed learned	Learned − shared	Local
1	0.271906	0.189196	0.266535	−0.005371	0.260199
2	0.267826	0.212686	0.275475	+0.007650	0.240091
3	0.269023	0.174030	0.275475	+0.006452	0.239967
4	0.261043	0.198280	0.266326	+0.005282	0.239664
5	0.261488	0.253081	0.263376	+0.001888	0.260044
6	0.260782	0.203773	0.268258	+0.007476	0.243898
7	0.266229	0.226977	0.271595	+0.005366	0.250273
8	0.268991	0.219263	0.272618	+0.003627	0.254977

Table 6: Current-rule policy means over the same thirty fires. Eight seed pairs are the replication units for method contrasts; together they aggregate the 240 episode outcomes per arm.

p-values are 1.70×10^{-5} and 0.00277. A preselected 100,000-draw paired-bootstrap check gives $[0.001492, 0.006122]$.

Fixed local-reward control. The registered replacement reward counts unique accepted placement cells within Chebyshev distance one of cells burning before the real transition, divided by 64^2. It uses no counterfactual transition and adds no shared reward. The action set includes the same cost-free ground wetline, and accepted wetline cells near burning earn this local reward. Rejected placements receive zero; overlapping agents each receive their own footprint count. Simulator credit measures realized incremental benefit, canceling redundant or immediately ineffective placements. The comparison tests a gain beyond this fixed proximity reward while keeping wetline access unchanged.

The eight local-reward runs use the same MAPPO hyperparameters, current rules, actor architecture and full training budget as the shared-reward and simulator-credit arms. All use the same fresh evaluation fires.

The registered primary local-minus-simulator mean difference is 0.0390 (95% paired-t, seven degrees of freedom: $[0.0212, 0.0568]$); paired seed-bootstrap sensitivity gives $[0.0256, 0.0531]$. The secondary local-minus-shared mean is -0.0173, with paired-t interval $[-0.0247, -0.0098]$ and bootstrap $[-0.0229, -0.0113]$. Simulator beats local and local beats shared in every seed pair.

Earlier-rule transfer. Eight frozen S1-trained policies per method use argmax actions on seeds 10000–10029. S2 increases ignition radius from two to four; S3 has three ignitions; S4 reduces aerial water to four and raises no-fly probability to 0.20; S5 has four hand crews and four aerial units. Reference policies are no suppression and a type-aware anchor-and-progress heuristic with access to fire truth and positions, using FIRELINE/NONE actions. Simulator credit has the lowest learned-policy IQM in all five scenarios and beats the heuristic mean in S1, S2 and S4; the heuristic is stronger in S3 and S5. In S1 its IQM is 0.1728 $[0.1584, 0.1825]$, versus shared 0.2643 $[0.2593, 0.2677]$ and learned 0.2680 $[0.2614, 0.2716]$. All eight shared/simulator pairs favor simulator credit. S2–S5 simulator scores are 0.241, 0.337, 0.191, 0.152, versus learned scores 0.305, 0.361, 0.269, 0.267; all five paired simulator/learned intervals exclude zero.

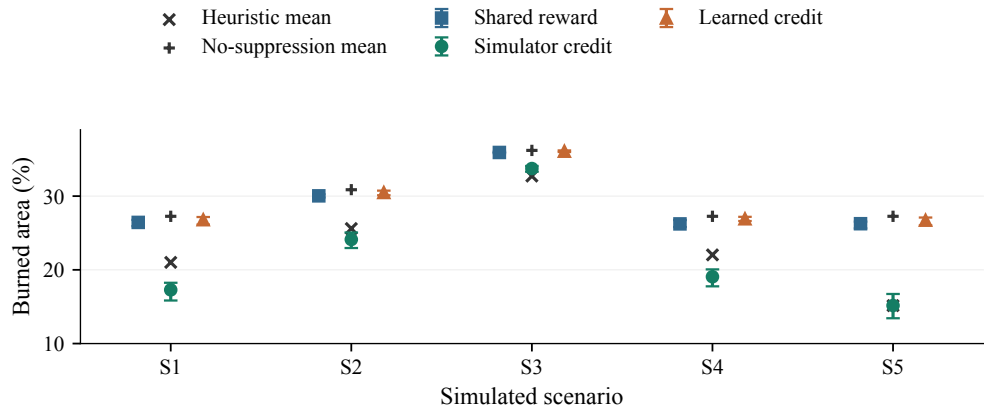


Figure 5: Scenario transfer. Colored symbols show IQMs and 95% training-seed bootstrap intervals; black markers show reference means. All learned policies are frozen S1 checkpoints. S1 is the earlier-rule training scenario; all scenarios use the same frozen policy cohort. S2: larger initial fire; S3: multiple ignitions; S4: tighter aerial constraints; S5: aerial-heavy roster.

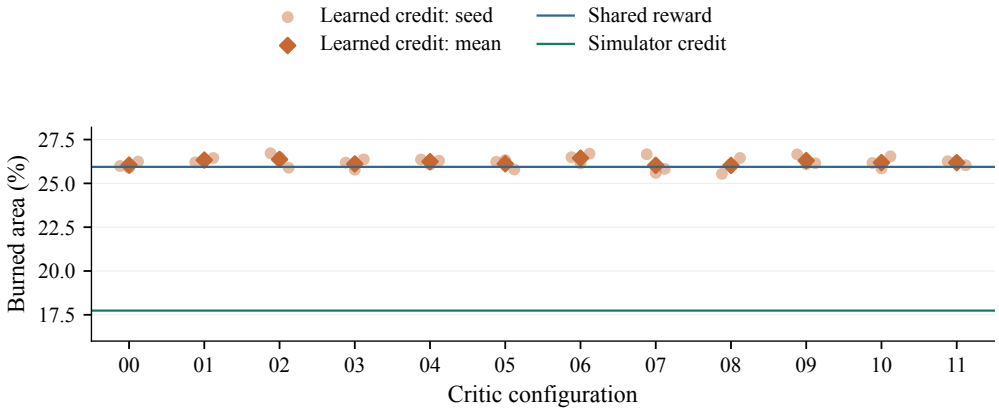


Figure 6: Learned-credit critic sweep: twelve configurations, three training seeds each. Each point is a seed's mean end-of-episode burned fraction over its last ten training iterations; diamonds average the three seeds. Dashed references use the same metric and seeds 1–3 from the main sweep. These are training summaries, not held-out tuned-policy scores or uncertainty intervals.

C CRITIC TUNING AND ACTOR-LOGIT CONTROL

The earlier-rule critic sweep uses seeds 1–3 and twelve configurations: base; learning rates 10^{-3} or 2×10^{-4}; width 256; TD mixing 0.8 or 1; a target critic; and five combinations. Target copies occur every 200 updates. All 36 runs complete the full budget. Three-seed final-ten-update means range from 0.2602 to 0.2644, versus shared 0.2594 and simulator 0.1774. Tuned c08 (target critic, mixing 1) has the lowest sweep mean. The six earlier-rule no-clamp runs change only the actor-logit bound across sampling, optimization and evaluation; PPO ratio and gradient clipping remain. All finish the full budget. Group means change from 0.2641 to 0.2683 for informed critics and 0.2667 to 0.2683 for zero-input critics.

Critic inputs	Seed	Clamped	No clamp	Difference
Informed	1	0.259793	0.268792	+0.008999
Zero-input	1	0.259793	0.270819	+0.011025
Informed	2	0.269488	0.267045	−0.002443
Zero-input	2	0.267701	0.267047	−0.000654
Informed	3	0.263088	0.268929	+0.005841
Zero-input	3	0.272616	0.266965	−0.005651

Table 7: Six matched actor-logit controls. Entries are mean burned fractions on fires 10000–10029; positive differences favor the clamped actor. PPO ratio clipping and gradient clipping remain enabled. The three seeds per condition are shown separately.

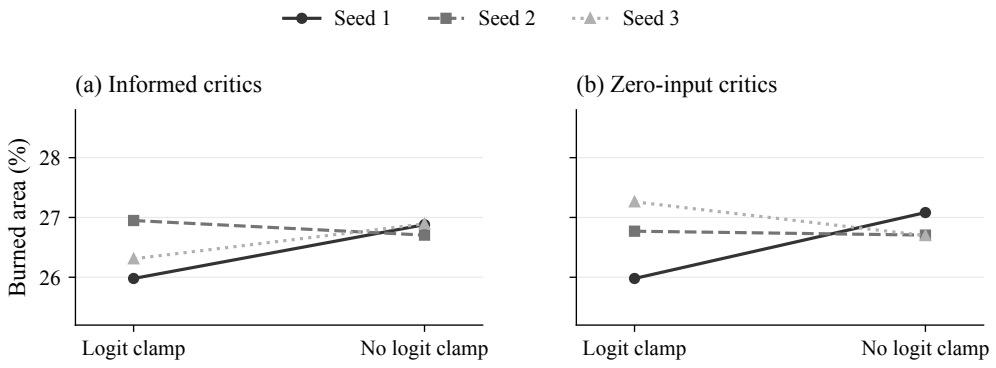


Figure 7: Removing the actor-logit clamp. Each line joins a matched training seed's mean burned area on thirty fixed fires with and without the clamp. Across these six runs, the three informed and three zero-input changes have mixed signs, with higher group mean burned area without the clamp in each condition.

D Randomness and intervention controls

Eight earlier-rule simulator actors each generate four stochastic S1 episodes (seeds 40000–40003). Uniform reservoir sampling selects one post-step-10 context with an accepted placement per episode; all 2,880 eligible steps qualify. The resulting 256 agent/context rows contain 195 nonempty placements and 180 deletions changing combined fields. Snapshots fix tactical state, movement, acceptance and all RNG states.

For raw reward R_e, placements P, focal deletion P_{-i} and independent fire keys k, l, compute

$$D_{ei}^{\text{coupled}} = R_e(P;k) - R_e(P_{-i};k), \quad D_{ei}^{\text{independent}} = R_e(P;k) - R_e(P_{-i};l). \tag{10}$$

Each context uses 256 independent key pairs, shared across agents. Marginal expectations agree; variance differs through branch covariance. Dynamics stay float32 and burned fields are summed in float64. At factual keys, all branches match serial fire updates exactly and all realized credits match training within 8.95×10^{-9}. The independent-draw credit is negative in $30{,}749/65{,}536$ samples (46.92%); shared-draw credit is never negative. Figure 2c uses the full empirical distribution, with equal context/draw/agent weights. The variance ratio averages within-context/agent variances, including 196 rows with zero sampled coupled variance.

Choice of removal. At each factual draw, full-state replay replaces the focal action by (m_i, none) or $(\text{stay}, \text{none})$, retaining teammates' actions and reapplying rules. Against placement deletion, both show exactly zero immediate-credit difference in all 256 rows (predeclared tolerance 10^{-7}) and no changed teammate placement. Each changes water in 15 rows and refill in three; staying changes positions in 243. These measurements concern the immediate reward, with tactical differences recorded separately.

Included focal rows	Count	Coupled variance	Independent variance	Variance ratio
All	256	2.08×10^{-11}	1.94×10^{-8}	934 [675, 1416]
Accepted placement	195	2.73×10^{-11}	1.94×10^{-8}	709 [537, 1055]
Deletion changes fields	180	2.96×10^{-11}	2.00×10^{-8}	677 [503, 1013]

Table 8: Mean within-context/agent variance, in raw reward squared, estimated across 256 replicate pairs. Ratios divide aggregate variances, avoiding undefined rowwise ratios. Brackets are descriptive 95% intervals from 2,000 nested bootstrap draws: resample whole actors with their four fixed episode seeds, then paired replicate rows jointly across focal agents within each context. Intervals condition on these fire seeds and selected contexts.

E AUDIT PROTOCOL AND CURRENT-RULE RESULTS

Four informed main-comparison critics (seeds 1, 4, 7, 8) were selected for responsive final actors. Ten earlier-rule critics provide consistency/information controls. Each is paired with its own actor, inputs and normalizers after full-budget training. Informed critics receive 109 bounded features including positions, types, resources, timers, previous movement, activity, time and dynamics indicators; zero-input critics retain this architecture but zero the tactical vector. Each audit uses its training rule set.

Sampling and uncertainty. Each checkpoint supplies one uniform reservoir context at or after step 10 from each of eight stochastic episodes (seeds 40000–40007), then sixteen paired remaining-horizon draws per action at $\gamma = .99$. Current query steps are 96, 63, 89, 55, 82, 92, 65, 18. Contexts are independent within a checkpoint; checkpoints are analyzed separately. Metrics average agents and contexts equally, with saved-policy action weights. Whole-episode intervals use 10,000 resamples conditional on stored replay means. Stored PopArt parameters and final running reward SD convert predictions to raw units.

With centered draws $Z_{ei,m}(b)$ from Equation (6), the separate standard error is

$$\widehat{\mathrm{SE}}_{ei}(b) = \sqrt{\frac{\sum_m [Z_{ei,m}(b) - \widehat{A}_{ei}(b)]^2}{M(M-1)}}. \tag{11}$$

For learned-minus-zero error define $D_{e,m} = N^{-1} \sum_{i,b} p_{ei}(b)[(A^c_{ei}(b))^2 - 2A^c_{ei}(b)Z_{ei,m}(b)]$. Its MCSE is $E^{-1}\sqrt{\sum_e \widehat{\mathrm{Var}}_m(D_{e,m})/M}$. This retains action/agent covariance and cancels the squared-reference term. RMSE ratios below are point estimates; inferential intervals use paired error differences.

Seed	Factual R^2	Factual RMSE	Credit RMSE	Zero RMSE	Replay MCSE
1	0.954	11.716	1.264	0.143	0.149
4	0.911	15.910	1.789	0.120	0.102
7	0.950	11.617	1.120	0.072	0.115
8	0.935	13.395	2.087	0.171	0.159

Table 9: Current-rule responsive main-comparison critics. All RMSE and replay MCSE columns are multiplied by 10^3. Credit and zero errors use equal context/agent weights and saved-policy action weights. Replay MCSE is the square root of the correspondingly weighted variance of the sixteen-draw credit mean.

Ranking and selectivity. For each action pair, set $D_m = G_m(a) - G_m(b)$. One eight-draw half selects pairs with $|\bar{D}| > t_{0.95,7} s_D/\sqrt{8} + 10^{-14}$ and matching nonzero signs in its two four-draw submeans; the other half checks direction, then roles swap. Held-out ties use the same 10^{-14} tolerance, and a tied critic gets half credit. Coverage counts unique pairs; direction agreement excludes held-out ties and gives tied critic predictions half credit. Fold/agent summaries average within episodes, then across represented episodes; reversals retain all selected pairs. Flat vectors have undefined

Seed	$10^6\Delta$ [95% interval]	Separate MCSE	RMSE / zero
1	1.577 [0.629, 2.843]	0.027	8.81
4	3.185 [1.308, 5.720]	0.045	14.85
7	1.248 [0.234, 2.552]	0.043	15.59
8	4.324 [1.381, 7.858]	0.071	12.20

Table 10: Current-rule raw learned-minus-zero error and error ratios. The separate MCSE column is also multiplied by 10^6. Intervals reuse the registered 10,000 complete-episode bootstrap samples, conditional on each checkpoint and its stored replay means. RMSE ratios are uncorrected finite-reference point estimates.

Seed	Unique pairs	Coverage (%)	Contexts	Reversals (%)	Critic agreement (%)
1	823	6.77	6/8	40.29	51.60
4	732	6.02	7/8	24.41	51.36
7	313	2.57	3/8	60.52	56.55
8	545	4.48	5/8	24.49	60.90

Table 11: Current-rule independent-half pair screening. Each checkpoint has 12,160 possible unique context/agent/action pairs. One eight-draw half selects contrasts and the other evaluates them, then the halves swap roles. Coverage counts each pair once; reversal and critic-agreement summaries retain the registered episode-macro weighting across both folds.

Seed	Spearman	95% episode interval	Defined contexts	Defined agent rows
1	+0.001	[-0.049, 0.050]	6/8	33/64
4	-0.036	[-0.138, 0.062]	7/8	45/64
7	+0.067	withheld	3/8	17/64
8	-0.025	[-0.090, 0.038]	6/8	35/64

Table 12: Within-agent action rankings for the current-rule cohort, averaged first within each episode and then over defined episodes. Flat reference rows have undefined rank correlation and remain in the reported denominators. The interval is withheld when fewer than five episodes define the metric.

Spearman correlation; intervals require five defined contexts. The sampled-flat counts (all twenty returns equal within each draw) are $31/64, 19/64, 47/64, 29/64$ for current seeds 1, 4, 7, 8. Median cross-agent Spearman correlations over each critic's 8×28 pairs are $0.986, 0.988, 0.985, 0.980$. This is a critic-only structural diagnostic; finite-reference directional precision is reported alongside it.

Independent-half reference precision. For current seeds 1, 4, 7 and 8, policy-weighted credit cosines between independent eight-draw halves are 0.002, 0.147, -0.449 and 0.104. Frozen actor-score gradient cosines range from -0.109 to $+0.662$, with all cross-product Monte Carlo intervals spanning zero (Table 16). Action-pair screening is reported in Table 11. These checks accompany the finite-reference calibration results; they do not identify a causal dilution mechanism or noise-free action ordering.

F Earlier audits and actor responsiveness

The earlier cohort comprises three base critics, tuned c08, three informed and three capacity-matched zero-input critics. All ten have positive learned-minus-zero error under policy and uniform weights. Base 1 and 3 have responsive checked actors and raw RMSE 9.09 and 7.01 times zero's. Base 2, tuned c08 and all three informed actors are fully saturated on their checked trajectories. The earlier informed cohort's pooled cross-agent rank correlation is 0.997.

Final-actor diagnostics. Four trajectories per actor (two stochastic, two argmax) and one post-step-10 gradient context per trajectory test final-policy sensitivity. Current learned-credit checks use stochastic seeds 40000–40001 and evaluation seeds 2026092400–2026092401; all deterministic outcomes reproduce saved evaluations. Table 14 identifies the responsive subset, whose descriptive

Critic	R^2	Credit	Zero	$10^6\Delta$ [95%]	MCSE
Base seed 1	0.938	1.758	0.193	3.0545 [1.7178, 4.4575]	0.0561
Base seed 2	0.808	0.516	0.203	0.2252 [0.1416, 0.3106]	0.0094
Base seed 3	0.890	1.340	0.191	1.7604 [0.5492, 3.5325]	0.0666
Tuned c08 seed 1	0.902	0.880	0.110	0.7631 [0.4884, 1.1470]	0.0110
Informed seed 1	0.957	0.354	0.178	0.0938 [0.0689, 0.1148]	0.0080
Informed seed 2	0.970	0.374	0.088	0.1326 [0.1164, 0.1487]	0.0063
Informed seed 3	0.962	0.447	0.237	0.1442 [0.0769, 0.2240]	0.0116
Zero-input seed 1	0.879	0.554	0.189	0.2709 [0.1370, 0.4254]	0.0040
Zero-input seed 2	0.907	1.189	0.410	1.2457 [0.3814, 2.4101]	0.0691
Zero-input seed 3	0.755	0.726	0.177	0.4955 [0.1159, 0.9879]	0.0029

Table 13: Earlier-rule consistency and information controls, eight contexts and sixteen draws each. Credit/zero columns are RMSE in 10^{-3} return; paired differences and separate MCSE are in 10^{-6} return squared. Intervals resample episodes conditional on stored means. All ten checkpoints remain represented.

mean burned fraction is 0.2693, versus paired shared reward 0.2670. Earlier MAPPO checks are in Table 14; all 16 actors have nonzero sampled NLL and entropy gradients. All eight local-reward

	Current learned credit			Earlier MAPPO			
Seed	Saturated	NLL	Entropy	Shared sat.%	Sim. sat.%	Shared NLL	Sim. NLL
1	0/3200	1.107	2.762	0.00	0.28	2.89	2.04
2	3200/3200	0.000	0.000	0.00	0.46	3.31	3.05
3	3200/3200	0.000	0.000	0.00	0.44	4.63	2.07
4	0/3200	5.945	3.697	0.00	0.00	3.66	2.18
5	3200/3200	0.000	0.000	0.00	1.71	4.29	3.98
6	3200/3200	0.000	0.000	0.00	3.66	3.74	2.49
7	0/3200	0.934	3.168	1.03	0.09	3.11	2.76
8	0/3200	3.775	4.177	0.00	0.81	3.66	4.28

Table 14: Final-actor checks: four trajectories and four gradient contexts per actor. Current saturation counts are out of 3,200 decisions; earlier percentages cover 2,568–3,200 decisions. NLL/entropy columns are minimum actor-parameter gradient norms over checked contexts. Decisions within trajectories are dependent.

actors have nonzero gradients on at least one checked context. Fully saturated decisions out of 3,200 are $250, 122, 112, 186, 2837, 2053, 106, 330$ for seeds 1–8. On the six local-reward seeds with fewer than 10.4% saturated decisions (1–4, 7–8), mean burned fraction is 0.2475 versus 0.2034 for their paired simulator-credit policies, which win all six pairs (Table 6). This post hoc subset retains the performance gap after excluding seeds 5 and 6. Every training seed remains in the primary comparison.

G Matched counterfactual targets and finite-data coverage

These standard identities specify the measurement framework. Use the complete context, frozen policy and remaining horizon from Equations (5)–(6); teammates' actions are fixed at the intervention, then all actors follow branch-specific observations.

Matched replay estimator. For fixed $z_t, \mathbf{a}, \pi$, let $Z_{i,m}$ be the centered factual-action return under draw m and $\widehat{A} = M^{-1}\sum_m Z_{i,m}$. Correct branch marginals, finite variance and independence across the M noise arrays give, by linearity and independence,

$$\mathbb{E}[\widehat{A}] = A_i^{\pi}(a_i), \qquad \mathrm{Var}(\widehat{A}) = \mathrm{Var}(Z_{i,1})/M. \tag{12}$$

Within an array, branches share draws. For two returns with fixed marginals, $\mathrm{Var}(G - G') = \mathrm{Var}(G) + \mathrm{Var}(G') - 2\,\mathrm{Cov}(G, G')$; positive covariance reduces difference variance relative to independent draws.

Finite-data counterfactual non-identification. Let $L(q) = \sum_{k=1}^{n}(q(z_k, \mathbf{a}_k) - y_k)^2$ have fixed regression labels and evaluated inputs $\mathcal{D} = \{(z_k, \mathbf{a}_k)\}_{k=1}^{n}$. For factual $(z, \mathbf{a}) \in \mathcal{D}$, consider an unobserved alternative $u = (z, (\mathbf{a}_{-i}, b)) \notin \mathcal{D}$, where $b \neq a_i$ and $p_b = \pi_i(b \mid h_i) > 0$. If the function class permits $q_\lambda = q + \lambda \mathbf{1}_{\{u\}}$, then

$$L(q_\lambda) = L(q), \qquad A_i[q_\lambda](z, \mathbf{a}) = A_i[q](z, \mathbf{a}) - p_b\lambda. \tag{13}$$

Here $A_i[q]$ is Equation (5)'s factual-action contrast with action values replaced by q. The perturbation vanishes on observed inputs and changes only the baseline's term for b. This holds for an unrestricted tabular class; regularization and structure can constrain unseen values. Full action support and zero population error against the correct conditional-mean target identify the contrasts.

Tabular action coverage. Fix a context z and anchor joint action $\mathbf{a}$ independently of n i.i.d. draws from a stationary behavior distribution $\mu(\cdot \mid z)$. Let U_i count unvisited focal alternatives with positive policy probability. An alternative of probability p is absent with probability $(1-p)^n$, giving

$$\mathbb{E}[U_i] = \sum_{\substack{b \neq a_i \\ \pi_i(b|h_i) > 0}} (1 - \mu((\mathbf{a}_{-i}, b) \mid z))^n . \tag{14}$$

Under independent uniform K-action sampling by all N agents in both μ and π, this is $(K-1)(1-K^{-N})^n$ per agent and $N(K-1)(1-K^{-N})^n$ across the distinct single-agent alternatives. Here n counts visits to the same complete context. The unseen fraction grows with N at fixed n, K; $20^8 = 25.6 \times 10^9$ counts joint actions. Neural generalization additionally depends on structure and factorization; this occupancy identity motivates a coverage intervention.

H Scale sensitivity and frozen actor-score gradients

RMS matching and held-out fitting. Equation (9) removes global magnitude inflation by matching the finite replay reference's RMS. It differs from PPO's centering and standardization of sampled batch advantages. Separately, define $W_e[f] = N^{-1}\sum_{i,b} p_{ei}(b) f_{ei}(b)$ and fit one nonnegative scalar on seven contexts, testing on the eighth:

$$\alpha_{-e} = \max\left\{0, \frac{\sum_{d \neq e} W_d[A^c \widehat{A}]}{\sum_{d \neq e} W_d[(A^c)^2]}\right\},$$

$$\Delta^{\text{scale}} = \frac{1}{8}\sum_{e=1}^{8} W_e\left[\alpha_{-e}^2 (A^c)^2 - 2\alpha_{-e} A^c \widehat{A}\right]. \tag{15}$$

Whole-episode bootstrap draws refit each fold, excluding all copies of its held-out original episode. Intervals condition on saved replay draws. Table 15 retains current raw, RMS-matched and fitted errors separately. Earlier-rule RMS-matched error is 1.797–2.235 times zero's; held-out fitting gives 1.0000–1.0073, with every fitted-error interval including or touching zero. Earlier critic/reference cosines range from -0.118 to $+0.101$, half cosines from -0.314 to $+0.573$, and estimated noise fractions from 0.535 to 1.505. The last quantity is MC variance divided by finite-reference squared magnitude, not a fraction of known true signal.

Frozen actor-score gradients. Using the saved current actors and exact replay observations, define

$$g(A) = \frac{1}{8N}\sum_{e=1}^{8}\sum_{i=1}^{N}\sum_{b=1}^{20} p_{ei}(b) A_{ei}(b) \nabla_\theta \log \pi_\theta(b \mid o_{ei}). \tag{16}$$

Credits and probabilities are held fixed; derivatives include all actor parameters. This is a frozen score gradient, distinct from a full PPO/GAE/entropy/Adam update. For per-context replay gradient $g^r_{e,m}$ and $g^c = g(A^c)$, the cross-product MCSE is

$$\widehat{\text{SE}}_{\text{MC}} = \sqrt{\frac{1}{8^2 M}\sum_{e=1}^{8} \widehat{\text{Var}}_m\left[g^c \cdot g^r_{e,m}\right]}. \tag{17}$$

Seed	Raw / zero	RMS / zero	Fitted / zero	$10^{10}\Delta^{\text{scale}}$ [95%]	Cosine
1	77.64	1.818	1.005	$1.1\ [-0.00338, 3.12]$	0.091
4	220.48	1.619	0.994	$-0.834\ [-13.2, 6.13]$	0.190
7	243.19	1.440	1.019	$0.973\ [-0.654, 3.99]$	0.280
8	148.79	1.803	1.044	$12.8\ [0, 52.4]$	0.098

Seed	RMS multiplier	Minimum α_{-e}	Maximum α_{-e}
1	0.1130	0.0004	0.0129
4	0.0666	0.0042	0.0151
7	0.0631	0.0029	0.0238
8	0.0816	0	0.0187

Table 15: Current-rule policy-weighted calibration and finite-reference alignment. "Raw / zero", "RMS / zero" and "Fitted / zero" divide raw, reference-RMS-matched and held-out MSE by zero's MSE. Cosine compares critic/replay credit. Independent-half agreement is reported in Appendix E. Episode intervals refit the common scale in duplicate-safe held-out folds. The lower panel gives the descriptive RMS-matching multiplier and the minimum/maximum MSE-fit multipliers across all eight held-out folds.

Table 16 uses normal-approximation Monte Carlo intervals; whole-context bootstrap uncertainty is separate. Zero norms give undefined cosines. In all ten zero-gradient contexts, action returns are identical within every sampled draw. The measured half agreement and MC intervals quantify the precision available for directional interpretation.

Seed	Gradient cosine	Half cosine	Zero contexts	10^7 cross product [95% MC]
1	-0.012	-0.151	2/8	$-0.033\ [-3.317, 3.251]$
4	0.662	0.410	1/8	$9.135\ [-8.197, 26.466]$
7	0.652	-0.895	5/8	$1.369\ [-0.874, 3.611]$
8	-0.109	0.287	2/8	$-1.463\ [-5.996, 3.070]$

Table 16: Frozen actor-score diagnostics for the four current-rule critics. Cosines use the context-averaged gradients. All four critic/reference cross-product Monte Carlo intervals span zero. "Zero contexts" counts exactly zero sampled mean reference gradients, whose per-context cosines are undefined.

I SUPPLEMENTARY VISUAL DIAGNOSTICS

Plotting protocol. Figure 2 uses trailing 50-update means within seeds and whole-seed bootstrap bands; the equivalence band centers on the shared arithmetic mean. Figure 4 fixes its example to seed 1's first stored context (episode 40000, step 96), uses stable centering for sampled-flat rows, and retains separate errors for fourteen checkpoints. Illustration captions state the fixed fire, policy seed and episode lengths.

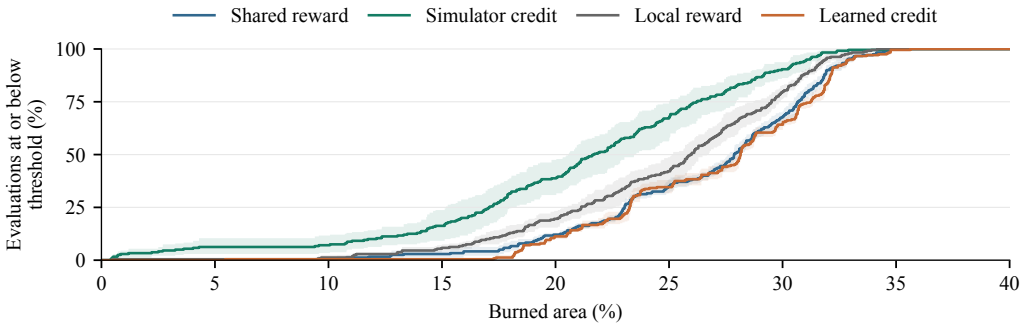


Figure 8: Empirical performance profiles for all four training arms. At each burned-area threshold, the curve gives the fraction of the 8×30 recorded evaluations at or below that threshold. Shading gives pointwise descriptive 95% intervals from 10,000 whole-training-seed resamples, retaining all thirty fixed fires within each seed. The 240 outcomes per arm retain their training-seed clusters; the curve describes all 240 fixed-fire observations.

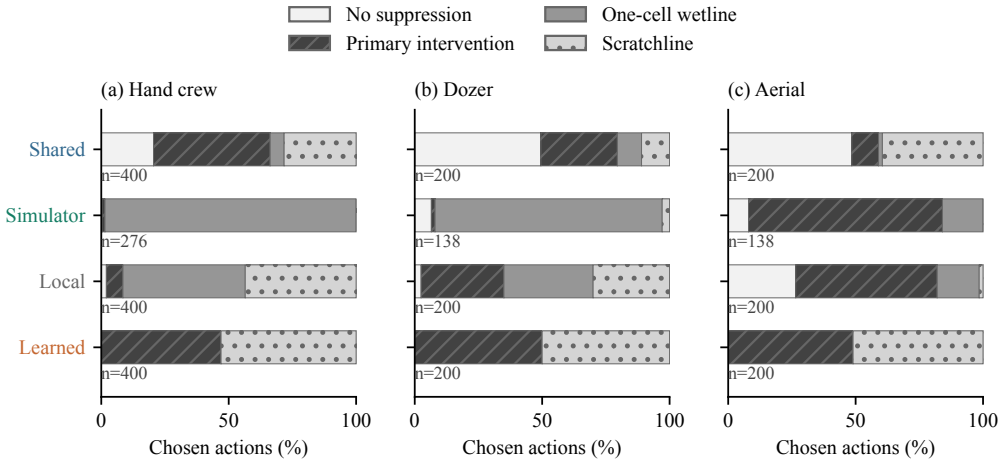


Figure 9: Chosen suppression primitives by agent type for all four methods, using training seed 1 and fire 2026092400. The three original episodes match Figure 3; the saved local-reward episode reproduces its recorded evaluation. Each row sums to 100% of n decisions; movement is omitted. Primary intervention is ground fireline or aerial patch drop. Simulator credit terminates after 69 steps, the other arms after 100. Ground wetline is unanchored and has no modeled water cost.

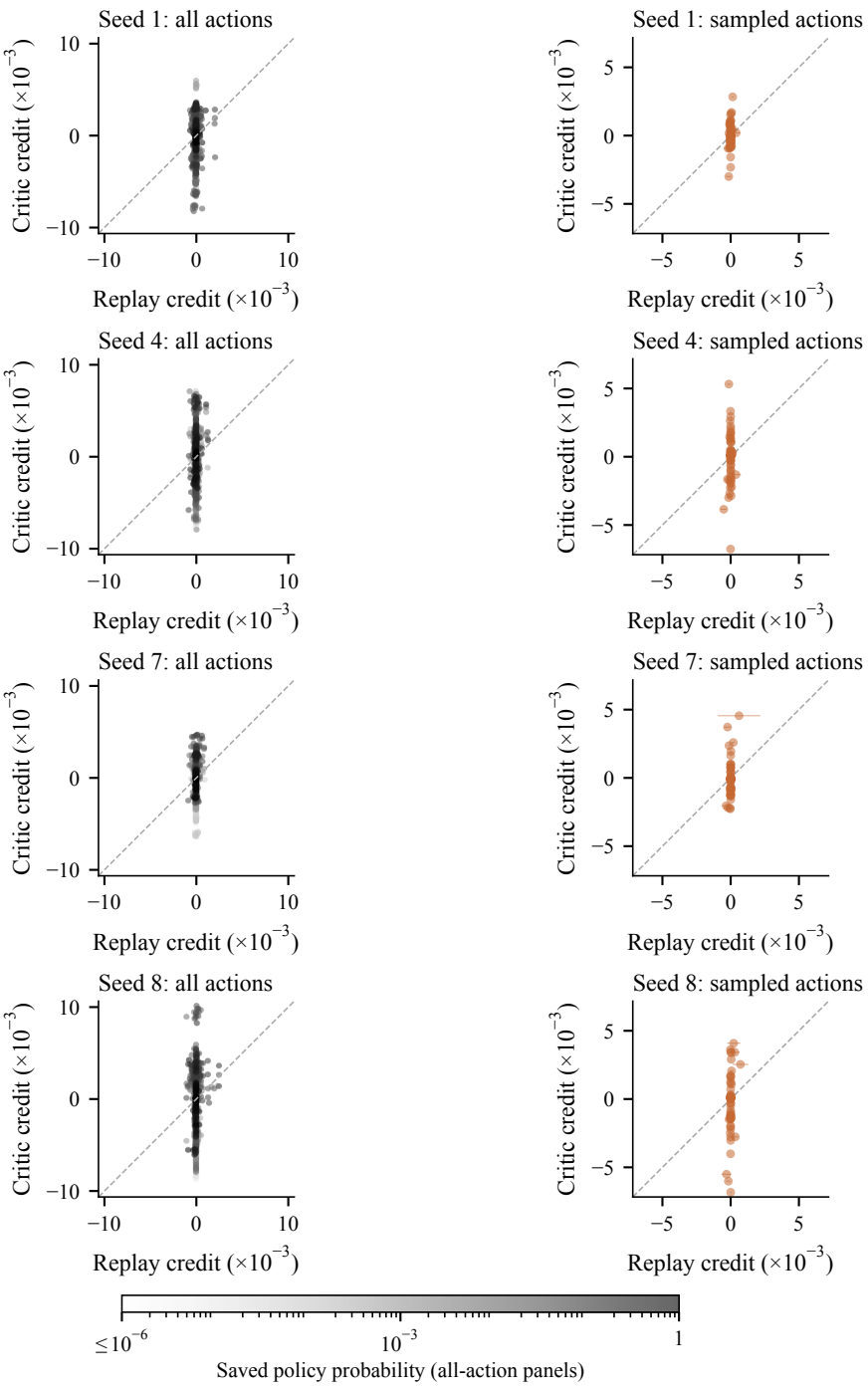


Figure 10: Detailed current-rule audits, one row per responsive actor and its informed critic (seeds 1, 4, 7 and 8). Left: all twenty actions, 1,280 dependent points per checkpoint, shaded by saved action probability. Right: the 64 sampled actions, with horizontal ± 1 Monte Carlo standard errors. Each column uses common equal numerical axes; credits are policy-centered and the primary error metric uses saved-policy weights. Replay estimates use sixteen paired full-horizon continuations.