PRIME: Plasticity Recovery in Multi-Agent Environments for UAV-Assisted Emergency Communication Networks

Wen Qiu, *Member, IEEE,* Zhiqiang He, *Member, IEEE,* Wei Zhao, *Member, IEEE,* and Hiroshi Masui

arXiv:2607.17922v1 [cs.MA] 20 Jul 2026

***Abstract*—When disasters destroy terrestrial infrastructure, fleets of unmanned aerial vehicles (UAVs) can restore connectivity as temporary aerial base stations, coordinating trajectories, spectrum, and power while the mission itself keeps changing. Most reinforcement learning controllers for these networks assume stationary conditions, and the few that handle change react to the external environment while leaving the network's internal state unexamined. We show that sustained non-stationarity damages this internal state directly: as objectives shift, neurons progressively fall dormant and the shared policy loses the capacity to learn. The obvious remedy, resetting dormant neurons, is unsafe under shared-parameter multi-agent training: many neurons that appear inactive are still receiving strong training gradients, and whether a neuron appears dormant depends on which agent's observations it processes. PRIME (Plasticity Recovery In Multi-agent Environments) therefore verifies both directions before intervening. Extending the bidirectional Silent Neuron framework to cooperative multi-agent reinforcement learning, it aggregates activation and gradient statistics over the full team batch, reads the backward signal from the gradient the training loss has already deposited , not from a hand-crafted proxy, and reinitializes only neurons that are simultaneously activation-dormant and gradient-silent. Useful representations are preserved while learning capacity is restored. On a phase-switching UAV emergency communication simulator, PRIME improves interquartile mean return by 24.9% over MAPPO and holds dormant neuron fractions at 10–20% versus 40–45%; ablations attribute the gains to the gradient signal and team-level aggregation rather than to the specific reset operator. A dynamic regret bound shows that the perturbation cost scales with the small silent-subspace dimension rather than the full parameter count.**

***Index Terms*—Multi-Agent Reinforcement Learning, Neural Plasticity, Dormant Neuron, UAV-Assisted Emergency Communication, Non-Stationary Environments**

I. Introduction

WHEN large-scale disasters strike, terrestrial communication infrastructure is often among the first casualties, and restoring connectivity in the affected area is a prerequisite for coordinating subsequent rescue and relief operations [1], [2]. Deploying multiple unmanned aerial vehicles (UAVs) as temporary aerial base stations has emerged as a practical solution, forming what is commonly referred to as a UAV-assisted emergency communication network (UAV-ECN) [3], [4]. What makes this problem distinct from other UAV deployments is that ground-level demand is driven by the evolving disaster situation rather than by a pre-designed traffic model: as rescue operations progress through distinct phases, the number, location, and priority of users shift in ways no prior distribution adequately captures. Non-stationarity is therefore not an occasional disturbance but a structural feature of the task.

Reinforcement learning (RL) and its multi-agent extension (MARL) have been widely adopted for UAV-ECN control, since they learn policies directly from interaction with the environment and jointly optimize trajectories, spectrum, and transmit power without requiring closed-form channel or mobility models [5]–[10]. Most of these studies, however, assume the environment is stationary or that its variation can be perceived externally. A smaller line of work addresses non-stationarity directly, through prioritized experience replay [11], meta-RL for rapid adaptation [12], or communication-assisted phase handling such as MACPH [13]. These methods help the agent respond to external change, but they leave open a more basic question: what happens *inside* the policy network as it trains under sustained non-stationarity?

The answer matters in practice: internal degradation does not surface in standard reward curves or training loss plots, so an agent may continue to produce actions while its representational capacity quietly erodes, leaving it unable to adapt when the disaster situation transitions to a new phase. Recent work in continual reinforcement learning has documented exactly such a degradation, termed plasticity loss: as training progresses under sustained distributional shift, a growing fraction of neurons drift toward near-zero activation and become dormant, and the network's capacity to fit new data distributions steadily declines [14]–[16]. The operating condition of UAV emergency communication matches this trigger exactly: demand undergoes abrupt transitions between disaster phases, and features learned for one phase become ineffective in the next.

The cooperative multi-agent setting compounds this problem in ways that single-agent treatments do not anticipate. Most modern MARL systems, including multi-agent proximal policy optimization (MAPPO) [17], train a single policy network whose parameters are shared across all agents and updated from a batch that pools every agent's experience. Under this regime, a neuron that becomes dormant degrades the

W. Qiu and H. Masui are with the Department of Information and Communication Engineering, Kitami Institute of Technology, Japan (e-mail: w-qiu@ieee.org; hgmasui@mail.kitami-it.ac.jp).

Z. He is with the Graduate School of Informatics and Engineering, University of Electro-Communications, Japan (e-mail: hezhiqiang@ieee.org).

W. Zhao is with the School of Computer Science and Technology, Anhui University of Technology, China (e-mail: zhaowei@ahut.edu.cn).

Corresponding authors: Wei Zhao and Hiroshi Masui.

decision quality of every agent on the team rather than a single learner, so the cost of leaving plasticity loss untreated scales with team size. Diagnosis is also less straightforward: the same neuron may produce near-zero activation on one agent's input while remaining active on another's, and a single per-neuron statistic must summarize this heterogeneous behavior before any reset decision can be made. These properties of the shared-parameter regime: amplified damage, ambiguous diagnosis, and the composite nature of the training gradient discussed below, make plasticity recovery in cooperative MARL a problem that needs to be solved at the team level, not merely transplanted from single-agent recipes.

A natural intervention is to detect dormant neurons and reset their weights so that they can re-engage in learning [14], [18]. Carrying this out reliably in shared-parameter multi-agent training, however, raises two coupled challenges. The first is that existing methods declare a neuron dormant when its forward activation falls below a threshold [18]. Yet under distributional shift, a neuron with near-zero activation may be in the process of being steered by gradient updates toward a new functional configuration, and resetting it on forward evidence alone destroys learning already underway [16]. The second is that adding a backward dimension, the natural refinement, is non-trivial: prior single-agent methods rely on hand-crafted proxies such as a utility score combining activation magnitude with outgoing weight norm [14], or a gradient saliency value from an auxiliary forward-backward pass on the network output [19]. In MAPPO with shared parameters, the effective training gradient additionally reflects experience aggregation across agents, importance-ratio clipping, and advantage estimation, and whether hand-crafted surrogates remain reliable under this composite signal has not previously been tested; we test it directly in Section V.

Existing plasticity maintenance methods have advanced unevenly along these two dimensions, and only a small subset has been examined under cooperative multi-agent training. Early countermeasures apply global perturbation indiscriminately, either by injecting noise into all parameters [20] or by periodically resetting weights toward their initial values [21], at the cost of disrupting features that remain useful. ReDo refines the granularity to individual neurons, resetting only those whose forward activation falls below a threshold [18], but inherits the limitation of single-dimensional detection. Notably, ReSiN [19] follows a similar neuron-level philosophy and adds a backward criterion, classifying a neuron as silent only when it is simultaneously forward-dormant and gradient-silent; it has been developed and evaluated only in single-agent settings, with the backward signal obtained through a separate forward–backward pass on the network output. On a parallel track, PE-MAMoE [22] addresses plasticity loss in cooperative multi-agent UAV emergency communication through an architectural route: a mixture-of-experts actor with a phase controller that perturbs experts after switches, preserving expressivity through expert diversity rather than neuron-level diagnosis and departing from the shared-MLP backbones standard in MARL practice.

Building on this analysis, we propose PRIME (**P**lasticity **R**ecovery **I**n **M**ulti-agent **E**nvironments), which extends bidirectional silent neuron detection to the shared-parameter multi-agent regime while remaining compatible with the standard MAPPO backbone [17]. The centralized-training-with-decentralized-execution (CTDE) architecture pools every agent's experience into a single set of network weights, providing a unified substrate for team-level diagnosis: PRIME aggregates forward activations and training gradients over the full team batch and classifies a neuron as *silent* only when it is simultaneously forward-dormant and gradient-silent under layer-normalized thresholds. A forward-dormant neuron still receiving gradient updates is preserved.

The key distinction from prior bidirectional designs lies in the source of the backward signal. Rather than constructing a proxy gradient, PRIME reads the gradient that PPO training has already deposited on the network parameters, a signal that already encodes multi-agent data aggregation, importance-ratio clipping, and advantage computation. Detection is performed at fixed periodic intervals rather than triggered by external phase signals, decoupling plasticity recovery from any assumed knowledge of when the environment changes. When a silent neuron is reset, PRIME re-initializes its incoming weights while zeroing its outgoing connections, preserving policy output continuity across the reset.

The main contributions of this paper are as follows.

1) **Problem identification.** We provide what is, to our knowledge, the first systematic characterization of how dormant neuron accumulation limits adaptation across disaster phases in shared-parameter cooperative MARL. Diagnostic measurements show that forward dormancy and the actual training gradient can diverge substantially, particularly in the policy network, making a forward-only reset criterion prone to false positives. This internal-state perspective is distinct from prior multi-agent work based on architectural specialization.
2) **Method.** We propose PRIME, a plasticity maintenance module integrated into shared-parameter MAPPO without altering its backbone architecture. Its core contribution is a bidirectional silent neuron criterion built on three deliberate choices: the backward signal is the gradient the PPO training loss has already deposited, so diagnosis reflects each neuron's actual role in the composite multi-agent objective (experience aggregation, importance-ratio clipping, advantage estimation); detection statistics aggregate over the full team batch rather than any single agent's slice; and the trigger is a fixed internal period rather than an external phase signal. To our knowledge, this is the first plasticity maintenance method designed for the shared-parameter cooperative MARL regime that performs neuron-level diagnosis on the live training gradient.
3) **Theoretical and empirical validation.** Extending the single-agent tracking analysis of ReSiN [19] to the team-averaged MAPPO objective, we establish a dynamic regret bound showing that selectively resetting only the identified silent neurons incurs lower tracking cost than perturbing all parameters. Experiments in a non-stationary UAV-ECN environment with eight phase transitions show a 24.9% IQM improvement over MAPPO

and suppression of dormant neuron fractions from 40–45% to 10–20%, with consistent gains in coverage and collision safety; design-axis ablations that substitute the backward signal, the aggregation slice, and the reset operator in isolation confirm that the first two choices carry the effect while the operator is exchangeable.

The remainder of this paper is organized as follows. Section II presents the measurements that motivate the design. Section III describes the system model and problem formulation. Section IV details the PRIME framework, including the bidirectional detection mechanism, the reset procedure, and the theoretical analysis. Section V reports experimental results and ablation studies. Section VI reviews related work, and Section VII concludes the paper.

II. Motivation and Analysis

This section establishes three findings on a vanilla MAPPO controller in our UAV-ECN simulator: dormant neurons accumulate and persist under standard gradient updates; forward dormancy and backward gradient silence diverge in three of four hidden layers, so a forward-only reset criterion would discard neurons that are still being actively updated; and dormancy decisions vary by 5–18% across agents, making per-agent reset criteria structurally inconsistent. Together these findings define the design target for Section IV.

A. Plasticity Loss Accumulates and Persists

In a UAV-ECN mission the optimization objective changes at every phase boundary (Section III details the phase model), violating the stationary-MDP assumption underlying standard PPO convergence guarantees [23] and creating conditions under which plasticity loss is both rapid and compounding.

To quantify the effect, we instrument a vanilla MAPPO controller (shared two-layer MLP, width $H{=}32$, ReLU activations) under two regimes: a fixed-phase *Normal* mode ($3.07{\times}10^6$ environment steps) and a cyclic-phase-switching *Change* mode with eight phase transitions ($18.43{\times}10^6$ steps). At every training iteration we record (i) the *dormant neuron fraction* (DNF), defined as the proportion of hidden neurons whose batch-averaged absolute activation falls below a normalized threshold τ_d; and (ii) the *dormant overlap* (LDO), the fraction of all hidden neurons that are dormant at both the current and previous checkpoint.

Figure 1 reveals a non-uniform but broadly severe dormancy landscape. Normal-mode counterparts of this and the two following figures are collected in Appendix D of the supplementary material. Value Layer 1 saturates near 60% in both modes, leaving a majority of that layer effectively unused, and the remaining three layers fluctuate in 14–58% bands without spontaneously decaying. Across all four layers, DNF and LDO traces are nearly indistinguishable, meaning that the current dormant set is almost entirely a subset of the previous one: dormancy persists across checkpoints rather than rotating. In Change mode, the eight phase boundaries produce no spontaneous recovery; DNF in the affected layers oscillates around its phase-conditional level but never relaxes. Standard gradient updates alone are therefore insufficient to reactivate dormant units, and the cooperative CTDE architecture compounds the cost: because all n agents share the same network, every dormant neuron degrades every agent's output simultaneously.

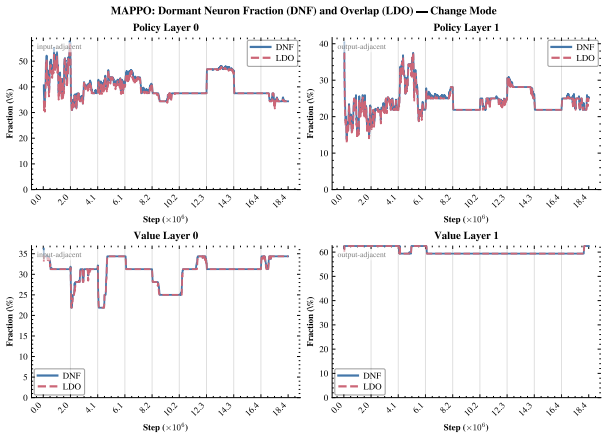


Fig. 1: Dormant Neuron Fraction (DNF, solid blue) and Dormant Overlap (LDO, dashed red) for the four hidden layers of a vanilla MAPPO controller in Change mode ($\tau_d{=}0.5$; vertical gray lines mark the eight phase boundaries, none of which produces spontaneous recovery). The near-coincidence of LDO with DNF indicates that the current dormant set is almost entirely a subset of the previous one rather than rotating across iterations. Value Layer 1 saturates near 60%; Policy Layer 0 ranges 31–58%, Policy Layer 1 14–40%, and Value Layer 0 22–36%.

B. Forward Dormancy Diverges from Backward Gradient Silence

Forward DNF measures whether a neuron contributes to the network's output at inference time, but it does not measure whether the optimizer is still investing learning capacity in that neuron. To make the distinction concrete, we additionally measure the *backward silence fraction* $|\mathcal{G}_l|/H_l$, where $\mathcal{G}_l$ is the set of neurons whose per-parameter gradient magnitude, normalized within the layer, falls below a threshold τ_g. Both quantities come from the same run under the same per-layer normalization ($\tau_d{=}0.5$, $\tau_g{=}0.08$), so the two fractions are directly comparable.

Figure 2 shows the resulting time series. The forward-dormant set is strictly larger than the backward-silent set in three of four hidden layers throughout training. In Policy Layer 0 the gap reaches 20–30 percentage points ($|\mathcal{D}_l|{\approx}45\%$ vs $|\mathcal{G}_l|{\approx}17\%$); in Policy Layer 1, $|\mathcal{D}_l|{\approx}30\%$ versus $|\mathcal{G}_l|{\approx}10\%$; in Value Layer 0, $|\mathcal{D}_l|{\approx}33\%$ versus $|\mathcal{G}_l|{\approx}18\%$. Value Layer 1 is the only exception: $|\mathcal{D}_l|$ and $|\mathcal{G}_l|$ both saturate at $\approx$62% and remain numerically indistinguishable for the entire run, reflecting a fully degenerate state in which the two criteria converge on the same neurons.

The persistent gap in the three non-saturated layers carries a concrete consequence for reset design: the set difference $\mathcal{D}_l \setminus \mathcal{G}_l$ consists of neurons whose forward activation is near zero but whose gradients remain large, neurons the optimizer is still actively steering toward the incoming distribution. A

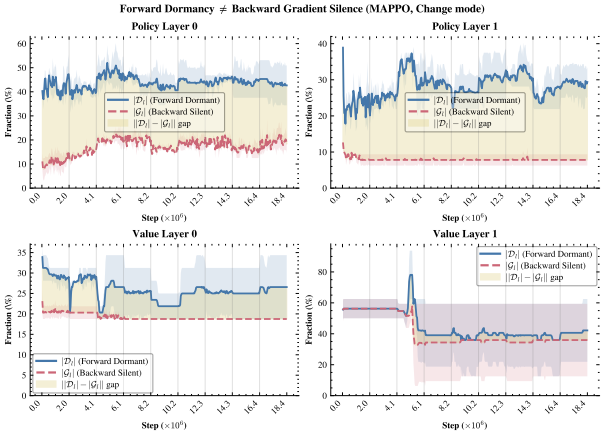


Fig. 2: Forward-dormant fraction $|\mathcal{D}_l|/H_l$ (solid blue, τ_d=0.5) and backward-silent fraction $|\mathcal{G}_l|/H_l$ (dashed red, τ_g=0.08) on a vanilla MAPPO controller in Change mode; both quantities use the same per-layer normalization. The shaded region $||\mathcal{D}_l| - |\mathcal{G}_l||$ marks the magnitude of decoupling. Policy Layers 0–1 and Value Layer 0 show a persistent gap with $|\mathcal{D}_l| > |\mathcal{G}_l|$, while Value Layer 1 saturates with $|\mathcal{D}_l| \approx |\mathcal{G}_l|$.

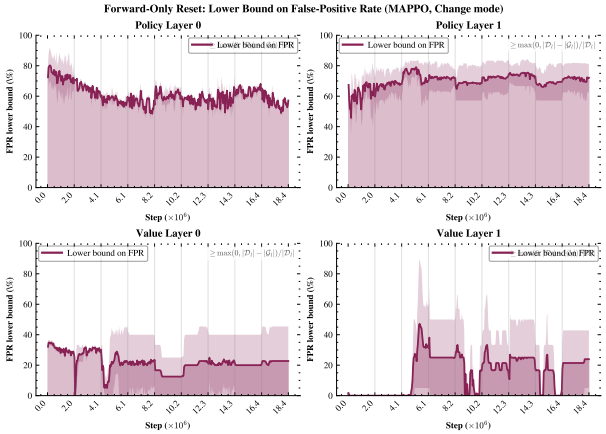


Fig. 3: Lower bound on the false-positive rate of a forward-only reset, $\max(0, |\mathcal{D}_l| - |\mathcal{G}_l|)/|\mathcal{D}_l|$, on a vanilla MAPPO controller in Change mode. The bound under-estimates the rate at which a forward-only criterion would flag neurons that still receive non-trivial gradient signal; Value Layer 1 remains near zero because dormancy and gradient silence have converged at saturation.

reset criterion that flags neurons solely on the basis of forward dormancy, as in ReDo [18], would discard the optimizer's accumulated progress on this population. Figure 3 bounds this cost from below by $\max(0, |\mathcal{D}_l| - |\mathcal{G}_l|)/|\mathcal{D}_l|$, an under-estimate of the false-positive rate that a forward-only reset would incur. In Policy Layer 0 this bound reaches 50–90%; in Policy Layer 1 it stays in 55–75%; in Value Layer 0 it ranges 25–45%; the true false-positive rate is likely higher.

A criterion that intersects $\mathcal{D}_l$ with $\mathcal{G}_l$ resolves this by definition, preserving forward-dormant neurons that still receive gradient signal. Value Layer 1 illustrates the complementary case: at saturation the intersection gracefully reduces to the forward-only criterion ($\mathcal{D}_l \cap \mathcal{G}_l = \mathcal{D}_l$), so the combined criterion is strictly safer than either component alone.

C. Cross-Agent Dormancy Disagreement

Shared-parameter CTDE routes n agents' observations through a single network, but each agent sees the joint state from its own perspective, so whether a neuron is dormant can depend on whose observation is being processed.

Figure 4(a) shows per-agent forward dormancy at Policy Layer 0 across training. The three agents trace visibly distinct trajectories within a 35–55% band. Figure 4(b) quantifies this dependence through the disagreement rate, defined as $(|\bigcup_i \mathcal{D}_l^{(i)}| - |\bigcap_i \mathcal{D}_l^{(i)}|)/H_l$, where $\mathcal{D}_l^{(i)}$ is the dormant set under agent i's observations alone. Disagreement is modest but persistent at 5–18% in policy layers; in value layers it remains near zero because the centralized critic operates on a joint-state encoding common to all agents.

A per-agent reset criterion therefore yields inconsistent decisions at every checkpoint; resolving the conflicts requires either a voting scheme, which discards information, or a unified statistic over the full $B \times n$ team batch. PRIME adopts the latter (Section IV-B).

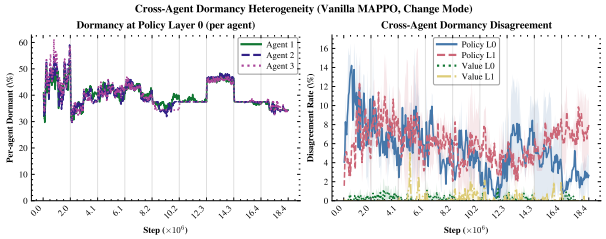


Fig. 4: Cross-agent dormancy heterogeneity on a vanilla MAPPO controller in Change mode. (a) Per-agent forward dormancy at Policy Layer 0: the three agents trace visibly distinct trajectories within a 35–55% band, indicating that the network's effective dormancy varies with input source. (b) Cross-agent disagreement rate, $(|\bigcup_i \mathcal{D}_l^{(i)}| - |\bigcap_i \mathcal{D}_l^{(i)}|)/H_l$, across all four layers. Policy layers exhibit 5–18% persistent disagreement; value layers remain near zero because the centralized critic operates on a joint-state encoding.

III. System Model

Following the UAV-ECN setup of [22], this section introduces the system model and its Dec-POMDP abstraction; Fig. 5 illustrates the scenario.

A. Network Topology and Kinematics

Consider a post-disaster square area $\Omega = [0, S]^2$ in which terrestrial infrastructure has been destroyed (Fig. 5). A fleet of U rotary-wing UAVs, indexed by $\mathcal{U} = \{1, \ldots, U\}$, is deployed as aerial base stations to provide temporary communication coverage for N ground users. At discrete time slot t, the three-dimensional position and velocity of UAV u are

$$\mathbf{q}_u(t) = \big[x_u(t),\, y_u(t),\, h_u(t)\big]^\top, \mathbf{v}_u(t) = \big[v_{h,u}(t),\, v_{z,u}(t)\big]^\top, \tag{1}$$

with the kinematic update $\mathbf{q}_u(t+1) = \mathbf{q}_u(t) + \mathbf{v}_u(t)\,\Delta t$, subject to maximum horizontal speed $V_{\max}^h$, maximum ver-

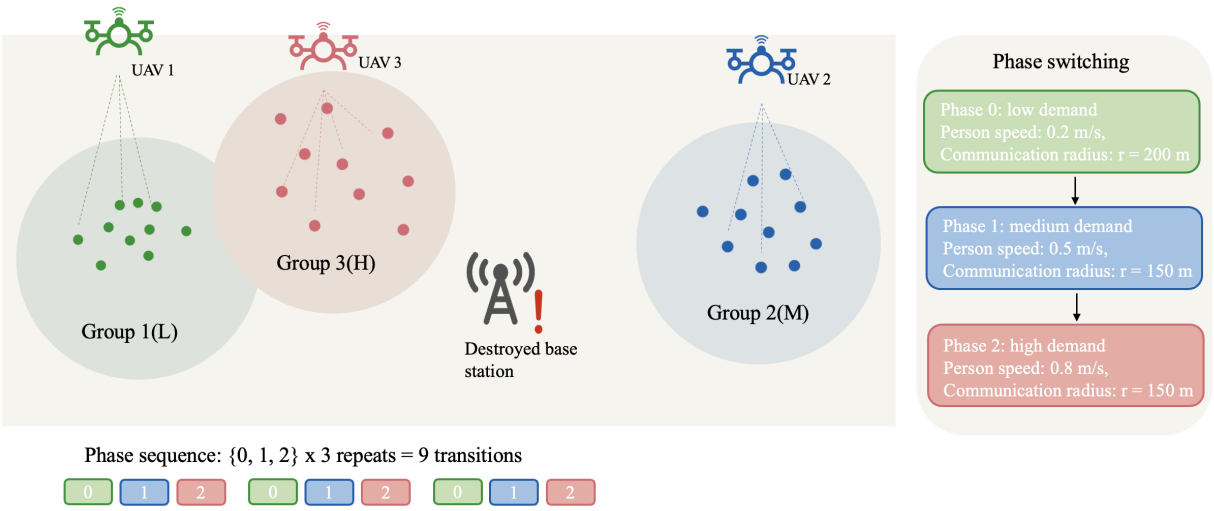


Fig. 5: System model of the UAV-assisted emergency communication network. Three UAVs serve as aerial base stations for N=20 ground users partitioned into G=3 RPGM groups under a cyclic three-phase demand regime. Each phase alters user mobility speed, communication coverage radius, and demand class (L/M/H). The phase sequence $\{0, 1, 2\}$ repeats three times, yielding nine phase segments (eight phase transitions) that drive the non-stationarity analyzed in this work.

tical speed $V^z_{\max}$, altitude bounds $[h_{\min}, h_{\max}]$, and the area boundary $\mathbf{q}_u(t) \in \Omega$.

B. Channel and Interference Model

The line-of-sight (LoS) probability for UAV–user pair (u, n) follows the 3GPP Urban Micro model with altitude correction [24]:

$$P_{\mathrm{LoS}}(r, h) = \frac{18}{r} + e^{-r/36}\left(1 - \frac{18}{r}\right)\big(1 + C_0(h)\big), \qquad (2)$$

where $C_0(h) = \max\{h-13,\, 0\}/101.5$ and r is the horizontal distance. This closed-form approximation of the 3GPP TR 36.777 LoS table reduces simulator overhead and is adopted from [22]. Path loss follows the 3GPP TR 38.901 large-scale models with Rician fading under LoS and Nakagami-m fading under NLoS. A frequency reuse factor F_{reuse} partitions UAVs into co-channel groups. The aggregate interference experienced by user n when served by UAV u is

$$I_{n,u}(t) = \sum_{\substack{u' \neq u \\ c(u')=c(u)}} P_{u'} g_{n,u'} + \rho_{\mathrm{adj}} \sum_{\substack{u' \\ c(u')\neq c(u)}} P_{u'} g_{n,u'}, \qquad (3)$$

where $\rho_{\mathrm{adj}} \in [0, 1)$ is the adjacent-channel leakage ratio, $P_{u'}$ is the transmit power, and $g_{n,u'}$ is the composite channel gain including path loss and small-scale fading.

C. Energy Consumption Model

Per-slot propulsion energy follows the three-dimensional rotary-wing model of [25], [26]:

$$\begin{aligned} P_{\mathrm{rot}}(V_h, V_z) =& P_0\left(1 + \frac{3V_h^2}{U_{\mathrm{tip}}^2}\right) \\ &+ P_i\left(\sqrt{1 + \frac{(V_h^2+V_z^2)^2}{4v_0^4}} - \frac{V_h^2+V_z^2}{2v_0^2}\right)^{1/2} \\ &+ \tfrac{1}{2} d_0 \rho s A V_h^3, \end{aligned} \qquad (4)$$

with standard rotary-wing aerodynamic parameters (blade profile and induced power P_0, P_i; tip speed U_{tip}; hover induced velocity v_0; drag ratio d_0; air density ρ; solidity s; disc area A). This energy cost enters the reward function as a penalty term that the policy must balance against coverage objectives.

D. User Mobility and Phase Model

N ground users are distributed among G hotspots according to the Reference Point Group Mobility (RPGM) model [27], [28]. Individual velocities combine a Gauss–Markov process with a group-following attraction:

$$\mathbf{v}_n^+(t) = (1-\kappa)\,\mathbf{v}_n^{\mathrm{GM}}(t) + \kappa v_0 \frac{\mathbf{c}_g(t) - \mathbf{x}_n(t)}{\|\mathbf{c}_g(t) - \mathbf{x}_n(t)\| + \varepsilon} + \boldsymbol{\eta}_n(t), \qquad (5)$$

where $\kappa \in [0, 1]$ controls the group-cohesion strength.

Time is partitioned into phases $\varphi_t \in \{0, 1, 2\}$, each of duration L iterations, cycling as $\varphi_t = \lfloor t/L \rfloor \bmod 3$. As shown in the right panel of Fig. 5, each phase defines a distinct mobility–demand regime: Phase 0 (low demand, user speed 0.2, coverage radius r=200 m), Phase 1 (medium demand, speed 0.5, r=150 m), and Phase 2 (high demand, speed 0.8, r=150 m). The demand level of each user group is drawn from classes $\ell \in \{\mathrm{L}, \mathrm{M}, \mathrm{H}\}$ with corresponding rate thresholds and reward weights (Table I). The sequence $\{0, 1, 2\}$ repeats three times, yielding nine phase segments and eight phase transitions that constitute the structured non-stationarity of the UAV-ECN task.

E. Multi-Objective Optimization

At each slot t, the controller jointly selects UAV poses, binary user associations $a_{n,u} \in \{0, 1\}$, and per-link band-

width/power allocations to maximize

$$\max_{\mathbf{q},\mathbf{v},\mathbf{a},\mathbf{b},\mathbf{p}} \mu_t^{\mathrm{qoe}} \sum_{n,u} a_{n,u}\, w_{\ell_n} \big[R_{n,u} \geq r_{\ell_n}\big] - \mu_t^{\mathrm{ene}} \sum_{u} E_u^{\mathrm{prop}} - \mu_t^{\mathrm{col}} \sum_{u<u'} \big[\|\mathbf{q}_u - \mathbf{q}_{u'}\| < d_{\min}\big], \tag{6}$$

subject to battery dynamics, association and power-budget constraints, and minimum inter-UAV separation $d_{\min}$ [22]. The phase-varying weights $(\mu_t^{\mathrm{qoe}}, \mu_t^{\mathrm{ene}}, \mu_t^{\mathrm{col}})$ shift at every phase boundary, jointly defining a piecewise-stationary objective whose transitions drive the plasticity dynamics this paper addresses.

F. Dec-POMDP Formulation

We cast Problem (6) as a cooperative decentralized partially observable Markov decision process (Dec-POMDP) [29], defined by the tuple $\langle n, \mathcal{S}, \{\mathcal{A}_u\}_{u=1}^n, \{\mathcal{O}_u\}_{u=1}^n, P, R, \gamma\rangle$, where n is the team size, $\mathcal{S}$ is the joint state space encompassing UAV poses, user positions, channel realizations, and the current phase index φ_t, and P is the joint transition kernel induced by UAV kinematics, user mobility, and channel dynamics. At each slot t, UAV u receives a local observation $o_u(t) \in \mathcal{O}_u$ comprising its own 3D position, relative positions of teammates, residual battery level, positions and achieved rates of users within its service radius, and the current-phase demand weights. Its action $a_u(t) \in \mathcal{A}_u$ selects the next pose displacement and the associations and resources for users it currently serves. The team shares a scalar reward R equal to the objective in (6), normalized per step, and discounts with $\gamma \in [0, 1)$.

We adopt the centralized training with decentralized execution (CTDE) paradigm [30]: during training, a centralized critic conditioned on the joint state evaluates a shared actor that conditions on local observations only, while at execution each UAV acts independently using its local policy. All n UAVs share the same actor parameters, which both reduces sample complexity and ensures that the dormancy and gradient statistics aggregated at the team level (Section IV-B) are well-posed.

IV. Methodology: PRIME

PRIME augments the standard MAPPO training loop with a periodic neuron activity assessment and selective reinitialization step. The network architecture remains unchanged: a shared MLP actor with two hidden layers of width H and ReLU activations, paired with a centralized MLP critic of identical topology. No routing modules, expert sub-networks, or auxiliary regularization terms are added. The only modification is a diagnostic-and-reset procedure executed every F mini-batch steps on both the actor and the critic.

The design is guided by a tradeoff between two failure modes observed in Section II. A reset criterion that is too permissive (such as the forward-only rule of ReDo [18]) reinitializes neurons that still receive informative gradient signal, incurring substantial false positives (Section II-B). A reset criterion that is too conservative leaves dormant neurons in place, allowing plasticity to erode across phases. PRIME navigates between these regimes by intersecting two complementary signals: forward dormancy identifies neurons whose information contribution has collapsed, while backward silence verifies that the optimizer has indeed abandoned them.

A second design consideration, the 5–18% per-agent dormancy disagreement measured in Section II-C, is addressed by aggregating both detection statistics over the full $B \times n$ team batch (Section IV-B).

A. Preliminaries: Dormant and Silent Neurons

Under the phase-driven non-stationarity of Section III, the Dec-POMDP becomes a family $\{M_0, M_1, M_2\}$ whose kernels P_φ and rewards R_φ are indexed by the demand regime; the mission cycles $\varphi_t \in \{0, 1, 2\}$ three times, producing eight transitions. The shared policy π_θ must therefore track a piecewise-stationary optimum w_φ^* that jumps at each transition, motivating the path-length-based tracking analysis in Section IV-E.

Within each M_φ, let $f_\theta \colon \mathbb{R}^{d_{\mathrm{in}}} \to \mathbb{R}^{d_{\mathrm{out}}}$ denote the shared policy network. Layer l contains H_l neurons with post-activation outputs $h_{l,i}(\mathbf{x}) = \sigma(\mathbf{w}_{l,i}^\top \mathbf{x} + b_{l,i})$.

Definition 1 (Dormant Neuron [18])**.** *The* dormancy index *of neuron (l, i) over data distribution $\mathcal{D}$ is the ratio of its mean absolute activation to the layer average:*

$$s_{l,i} \;=\; \frac{\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\big|h_{l,i}(\mathbf{x})\big|}{\frac{1}{H_l}\sum_{j=1}^{H_l} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\big|h_{l,j}(\mathbf{x})\big|}. \tag{7}$$

Neuron (l, i) is dormant *if $s_{l,i} \leq \tau_d$ for a user-specified threshold $\tau_d > 0$.*

Definition 2 (Silent Neuron [19])**.** *The* activity index *of neuron (l, i) combines its forward activation with its backward gradient:*

$$\xi_{l,i} \;=\; \frac{\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\big|h_{l,i}(\mathbf{x})\big| \;+\; \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\big|g_{l,i}(\mathbf{x})\big|}{\frac{1}{H_l}\sum_{j=1}^{H_l} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\big|h_{l,j}(\mathbf{x})\big|}, \tag{8}$$

where $g_{l,i}(\mathbf{x}) = \partial \sum_n f_\theta(\mathbf{x}_n)/\partial h_{l,i}$ is the gradient of the aggregated network output with respect to the neuron's activation. A neuron is silent *when $\xi_{l,i} < \varepsilon$, which, under the boundedness and non-degeneracy conditions of [19], holds if and only if both $\mathbb{E}|h_{l,i}|$ and $\mathbb{E}|g_{l,i}|$ are vanishingly small.*

The key distinction between dormancy and silence is that a dormant neuron may still receive substantial gradient signal, indicating that the optimizer is actively steering it toward a useful configuration.

B. Multi-Agent Silent Neuron Detection

PRIME adapts the single-agent definitions above to the batch structure of shared-parameter CTDE training. Each rollout of length T_{roll} collects observations from all n UAVs, producing a flattened batch $\mathbf{O}$ of shape $(T_{\mathrm{roll}} \cdot n,\ d_{\mathrm{obs}})$. The shared actor processes this batch in a single forward pass, yielding an activation tensor of shape $(T_{\mathrm{roll}} \cdot n,\ H_l)$ at each hidden layer l.

Forward activity index. Per-neuron activations $\{h_{l,i}^{(k)}\}_{k=1}^{B\cdot n}$ recorded during the PPO update are aggregated into a normalized dormancy score:

$$\hat{s}_{l,i} \;=\; \frac{\frac{1}{Bn}\sum_k |h_{l,i}^{(k)}|}{\frac{1}{H_l}\sum_j \frac{1}{Bn}\sum_k |h_{l,j}^{(k)}|}. \tag{9}$$

Neuron (l,i) is flagged as *forward-dormant* when $\hat{s}_{l,i} \le \tau_d$. Because the batch spans all n agents, this score measures team-level dormancy rather than any single agent's observation distribution.

Backward activity index. During the K_{epoch} PPO update epochs, per-neuron weight gradients $\partial\mathcal{L}_{\text{PPO}}/\partial\mathbf{w}_{l,i}$ are accumulated across all mini-batches. The normalized gradient magnitude is

$$\hat{g}_{l,i} \;=\; \left\| \frac{\partial \mathcal{L}_{\text{PPO}}}{\partial \mathbf{w}_{l,i}} \right\|_1 \Big/ |\mathbf{w}_{l,i}|, \tag{10}$$

where $|\mathbf{w}_{l,i}|$ denotes the number of elements in the weight vector of neuron i. Neuron (l,i) is flagged as *backward-silent* when $\hat{g}_{l,i} \le \tau_g$.

Intersection criterion. Let $\mathcal{D}_l = \{i\colon \hat{s}_{l,i} \le \tau_d\}$ denote the forward-dormant set and $\mathcal{G}_l = \{i\colon \hat{g}_{l,i} \le \tau_g\}$ the backward-silent set. The *silent neuron set* is their intersection:

$$\mathcal{S}_l \;=\; \mathcal{D}_l \;\cap\; \mathcal{G}_l. \tag{11}$$

Only neurons in $\mathcal{S}_l$ are candidates for reinitialization. No neuron carrying an informative signal in *either* the forward or the backward direction is disturbed.

Two properties of this construction deserve note. The backward index costs nothing beyond training itself: the gradient in (10) is the one the optimizer has already computed, at the scale the optimizer actually operates on, so no auxiliary pass is required. A known concern with loss-based gradients is masking: a neuron may receive a vanishing loss gradient while remaining active in the forward pass [19]. The intersection in (11) neutralizes this failure mode, since such a neuron fails the forward-dormancy test and is never reset. Section V tests the converse design directly, replacing the training gradient with an output-sensitivity proxy.

C. Selective Reinitialization

For each silent neuron $(l,i) \in \mathcal{S}_l$, three coordinated operations restore its capacity while preventing immediate disruption of downstream layers:

$$\mathbf{w}_{l,i} \leftarrow \text{Uniform}\big(-b_l,\, b_l\big)^{d_{l,\text{in}}}, \quad b_l = \sqrt{3/d_{l,\text{in}}}, \tag{12}$$

$$b_{l,i} \leftarrow 0, \tag{13}$$

$$[\mathbf{W}_{l+1}]_{:,i} \leftarrow \mathbf{0}. \tag{14}$$

Step (12) draws the incoming weights from the Kaiming uniform distribution, matching the variance of the original initialization. Step (13) zeros the bias. Step (14) sets the corresponding column of the next layer's weight matrix to zero, so the freshly reinitialized neuron produces no immediate output perturbation and must be activated through subsequent gradient updates [18]. The Adam optimizer state (first and second moment estimates) for all affected parameter slices is also cleared to prevent stale momentum from biasing the relearning trajectory. The operator ablation in Section V indicates that the hard reset itself is not performance-critical: a matched-cadence stochastic perturbation of the same silent set performs on par, and the reset is retained for its output continuity at the moment of intervention.

The same reset procedure is applied independently to the shared actor and the centralized critic.

Algorithm 1 PRIME-MAPPO Training

Require: Shared actor π_θ, centralized critic V_ψ (both two-layer MLP, width H, ReLU); thresholds τ_d, τ_g; reset period F (in mini-batch steps); n agents; PPO hyperparameters (K_{epoch}, clip ϵ, γ, λ_{GAE})
Output: Trained policy π_θ

Initialize θ, ψ via Kaiming uniform; set step counter $c \leftarrow 0$
for iteration $t = 1, 2, \ldots, T$ **do**
  // Step 1: Multi-agent rollout
  for step $k = 1, \ldots, T_{\text{roll}}$ **do**
    for each UAV $u = 1, \ldots, n$ **in parallel do**
      Observe $o_u^{(k)}$; sample $a_u^{(k)} \sim \pi_\theta(\cdot \mid o_u^{(k)})$
    end for
    Execute joint action; receive shared reward $r^{(k)}$
  end for
  Flatten: $\mathbf{O} \leftarrow \{o_u^{(k)}\}_{u,k}$, shape $(T_{\text{roll}}\cdot n,\ d_{\text{obs}})$
  // Step 2: MAPPO update with activity recording
  Compute GAE advantages $\hat{A}$ using V_ψ
  for epoch $e = 1, \ldots, K_{\text{epoch}}$ **do**
    for each mini-batch $\mathcal{B} \subset \mathbf{O}$ **do**
      Forward pass: record activations $\{h_{l,i}^{(k)}\}$
      Compute clipped PPO loss $\mathcal{L}_{\text{PPO}}$ and value loss $\mathcal{L}_V$
      Backpropagate; update θ, ψ via Adam; $c \leftarrow c + 1$
      // Step 3: Periodic silent neuron detection & reset
      if $c \bmod F = 0$ **then**
        for each hidden layer l in π_θ and V_ψ **do**
          Compute $\hat{s}_{l,i}$ via (9), $\hat{g}_{l,i}$ via (10)
          $\mathcal{D}_l \leftarrow \{i : \hat{s}_{l,i} \le \tau_d\}$; $\mathcal{G}_l \leftarrow \{i : \hat{g}_{l,i} \le \tau_g\}$; $\mathcal{S}_l \leftarrow \mathcal{D}_l \cap \mathcal{G}_l$
          for each $i \in \mathcal{S}_l$ **do**
            Apply (12)–(14); clear Adam states for affected slices
          end for
        end for
      end if
    end for
  end for
end for

D. PRIME-MAPPO Training Procedure

Algorithm 1 summarizes the procedure. Each iteration collects a multi-agent rollout, flattens it into the $(T_{\text{roll}}\cdot n,\ d_{\text{obs}})$ batch of Section IV-B, and runs standard clipped-PPO updates with GAE [31]; the only additions are the per-neuron activation and gradient statistics recorded during each forward and backward pass, at a cost of two scalars per neuron per layer. Every F mini-batch steps, the detection of Section IV-B and the reset of Section IV-C are applied to both the actor and the critic; with $F{=}200$ and $K_{\text{epoch}} \times 32 = 256$ mini-batch steps per iteration, detection fires approximately once per training iteration.

E. Theoretical Analysis

We analyze PRIME's tracking performance over the Dec-POMDP family $\{M_0, M_1, M_2\}$ formalized in Section IV-A. Because all n agents share the same policy parameter $w_t \in \mathbb{R}^d$, the MAPPO surrogate objective at iteration t can be written as

$$\mathcal{L}_t(w) \;=\; \frac{1}{n}\sum_{u=1}^{n} \mathcal{L}_t^{(u)}(w), \tag{15}$$

where $\mathcal{L}_t^{(u)}$ is the clipped PPO loss evaluated on agent u's rollout data. The gradient $\nabla\mathcal{L}_t(w) = \frac{1}{n}\sum_u \nabla\mathcal{L}_t^{(u)}(w)$ averages over the full multi-agent batch.

We adopt the following standard assumptions, which extend those used in the single-agent ReSiN analysis [19] to the shared-parameter CTDE setting.

Assumption 1 (L-Smoothness)**.** *There exists $L > 0$ such that $\|\nabla\mathcal{L}_t(w) - \nabla\mathcal{L}_t(w')\| \le L\|w - w'\|$ for all $w, w' \in \mathbb{R}^d$ and all t.*

Assumption 2 (μ-Strong Convexity (Local))**.** *There exists $\mu > 0$ such that $\langle \nabla\mathcal{L}_t(w) - \nabla\mathcal{L}_t(w'),\, w - w'\rangle \;\ge\; \mu\,\|w - w'\|^2$ for all w, w' within the PPO trust region at iteration t. This local condition is enforced by the clipped surrogate objective, which restricts each policy update to a neighborhood where the loss surface is approximately quadratic [23], [32].*

Assumption 3 (Bounded Non-Stationarity)**.** *The cumulative drift of the phase-conditional optimum $w_t^* = \arg\min_w \mathcal{L}_t(w)$ across the eight transitions of the family $\{M_0, M_1, M_2\}$ is bounded: $P_T = \sum_{t=1}^{T-1} \|w_{t+1}^* - w_t^*\| < \infty$.*

Assumption 4 (Selective Reset Energy)**.** *At each reset event, parameters in the silent set $\mathcal{S}_t = \bigcup_l \mathcal{S}_l$ of cardinality $d_t^* = |\mathcal{S}_t|$ are reinitialized. The perturbation is modeled as a projection-restricted noise injection: $E_t = \eta\gamma_{\mathrm{r}}\Pi_t\zeta_t$ with $\zeta_t \sim \mathcal{N}(0, I_d)$, where Π_t is the orthogonal projection onto the silent subspace of dimension d_t^*, and $\gamma_{\mathrm{r}} > 0$ controls the perturbation strength. The expected perturbation energy satisfies $\mathbb{E}\|E_t\|^2 = \eta^2\gamma_{\mathrm{r}}^2 d_t^*$, with $d_t^* \le d$ [19].*

The PRIME update rule augments the standard PPO gradient step with the selective perturbation:

$$w_{t+1} \;=\; w_t \;-\; \eta\,\nabla\mathcal{L}_t(w_t) \;+\; \eta\gamma_{\mathrm{r}}\,\Pi_t\zeta_t. \tag{16}$$

Between detection events the silent set is empty, so $\Pi_t = 0$ and (16) reduces to the plain PPO step; the sums over d_t^* below therefore accumulate only at the periodic reset events of Algorithm 1.

Theorem 1 (PRIME Dynamic Regret Bound)**.** *Under Assumptions 1–4, with learning rate $\eta \le 1/L$, the time-averaged squared tracking error of the shared policy parameter satisfies*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|e_t\|^2 \;\le\; \frac{2}{\mu\eta T}\left(\mathbb{E}\|e_0\|^2 + \frac{2P_T^2}{\mu\eta}\right) + \frac{2\eta\gamma_{\mathrm{r}}^2}{\mu}\cdot\frac{1}{T}\sum_{t=1}^{T} d_t^*, \tag{17}$$

where $e_t = w_t - w_t^*$ is the tracking error and $P_T = \sum_{t=1}^{T-1}\|w_{t+1}^* - w_t^*\|$ is the path length of the optimal parameters.

The bound takes the same form as ReSiN's single-agent tracking analysis [19]: the multi-agent structure enters through the team-averaged objective $\mathcal{L}_t$ and the path length P_T, and the practical content lies in how small the intersection criterion keeps d_t^*. The proof, together with the supporting noise-energy lemma, is given in Appendix B of the supplementary material.

Comparison with blanket perturbation. Undiscriminated noise injection (e.g., shrink-and-perturb [21]) corresponds to $\Pi_t = I_d$ and $d_t^* = d$ for all t, yielding a perturbation term $\frac{2\eta\gamma_{\mathrm{r}}^2 d}{\mu}$. Because the bidirectional intersection criterion (11) ensures $d_t^* \ll d$ in practice, PRIME's perturbation term is smaller by a factor of $\bar{d}^*/d$, where $\bar{d}^* = T^{-1}\sum_t d_t^*$ is the time-averaged silent-subspace dimension.

Role of phase non-stationarity. PRIME reduces the perturbation term in (17) without affecting the drift term P_T^2, improving tracking by lowering the cost of plasticity maintenance. The experiments in Section V confirm this: PRIME's dormant fraction stays at 10–20% while MAPPO reaches 40–45%, so $\bar{d}^*$ is a small fraction of d. Appendix C of the supplementary material discusses the assumptions underlying Theorem 1 in more detail, including how the additive-noise reset model relates to PRIME's actual reset operations.

V. Experiments

A. Setup

Environment. We evaluate on a 1 km × 1 km post-disaster urban grid with U=3 UAVs serving N=20 ground users organized in G=3 RPGM groups, with slot duration Δt=60 s and an episode horizon of 32 steps. Fixed simulation parameters and the rotary-wing energy model are listed in the upper part of Table I. Two training regimes follow Section II: *Normal* mode holds a single fixed phase for 1,500 iterations ($\sim 3\times10^6$ environment steps), and *Change* mode cycles $\{0, 1, 2\}$ three times (9,000 iterations, $\sim 18.4\times10^6$ steps, 1,000 iterations per phase, eight transitions). Each phase simultaneously alters user mobility speed, service radius, and the demand pattern with its overlap penalty λ_{ov} (lower part of Table I); these concurrent shifts create the non-stationarity that motivates PRIME's plasticity maintenance.

Baselines. We compare three methods sharing the same shared two-layer MLP architecture (H=32, ReLU) and MAPPO training pipeline, with the plasticity strategy as the only difference: (i) **MAPPO**, no plasticity mechanism; (ii) **PRIME-Forward**, which resets all neurons with $\hat{s}_{l,i} \le \tau_d$ regardless of gradient magnitude, subsuming the forward-only family (ReDo [18], shrink-and-perturb [21], Plasticity Injection [33]); and (iii) **PRIME**, the bidirectional intersection criterion with τ_d=0.5, τ_g=0.08, and reset period F=200 mini-batch steps. PRIME constitutes the multi-agent CTDE adaptation of ReSiN's bidirectional criterion [19] with the shared-batch aggregation in Section IV-B. The full hyperparameter list is provided in Appendix A of the supplementary material.

Evaluation protocol. We report episode return, coverage rate (fraction of users receiving above-threshold data rate), served-user count, collision rate, and energy usage rate, using interquartile mean (IQM) [34] as the primary scalar metric for robustness to outlier episodes. The evaluation focuses

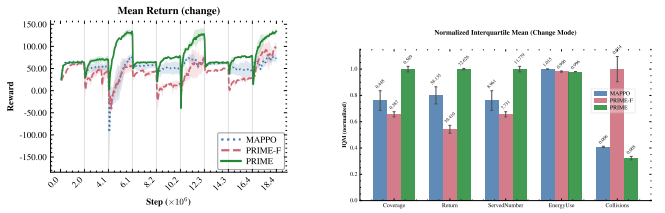
TABLE I: UAV-ECN Simulation Parameters

Parameter	Value
Fixed across all phases	
Area size $S \times S$	$1000\,\text{m} \times 1000\,\text{m}$
Number of UAVs U	3
Number of ground users N	20
RPGM groups G	3
Slot duration Δt	60 s
Episode horizon	32 steps
Max horizontal speed	12.0 m/s
Action step distance	100.0 m
Min inter-UAV distance $d_{\min}$	10.0 m
Max users per UAV	5
Frequency reuse factor	1
Adjacent-channel leakage ρ_{adj}	10^{-3}
Gauss–Markov α / σ	0.9 / 0.08
RPGM follow gain κ	0.3
Rotary-wing P_0 / P_i	90 W / 110 W
Tip speed U_{tip}	120 m/s
Phase-specific (Change mode)	
	Phase 0 / Phase 1 / Phase 2
User speed (normalized)	0.2 / 0.5 / 0.8
Service radius (m)	200 / 150 / 150
Demand pattern	all L / half→M / half→H
Demand thresholds (Mbps)	L: 0.5, M: 1.0, H: 2.0
Reward weights	L: 5, M: 10, H: 20
Overlap penalty λ_{ov}	20 / 40 / 80
Training schedule	
Phase sequence (Change mode)	$\{0, 1, 2\} \times 3$ repeats
Iterations per phase (Change)	1,000
Total iterations (Normal / Change)	1,500 / 9,000

on a single ECN scenario (U=3, H=32) so that the effect of bidirectional detection is isolated under a controlled non-stationarity profile with clearly identifiable transition points. The small network (1,184 trainable actor parameters) places a high premium on detection precision, since every unnecessary reset removes a significant share of capacity. PRIME's mechanism is architecture- and task-agnostic, requiring no environment-specific tuning beyond τ_d and τ_g (robustness confirmed by ablation, Fig. 9); extensions to larger teams and networks, standard MARL benchmarks (SMAC [35], MPE [30]), and bootstrap confidence intervals over larger seed populations [34] are left to future work. All experiments are repeated with two random seeds (42 and 43) under deterministic PyTorch settings. Reported IQM values are the mean of per-seed IQMs; curves with shaded regions show the cross-seed mean and the min–max band across seeds, ablation points carry min–max error bars, and the remaining per-neuron diagnostic traces are from a single representative seed (42).

B. Task and ECN Performance

Stationary baseline. In the stationary Normal setting, vanilla MAPPO attains the highest IQM return (83.392), followed by PRIME (78.352) and PRIME-Forward (50.622). PRIME's periodic resets trade a small amount of steady-state performance for sustained representational health (DNF



(a) Training reward curves

(b) Normalized IQM across five ECN metrics

Fig. 6: Change-mode task performance. (a) Across eight phase transitions (gray lines), PRIME recovers to the highest within-phase reward; PRIME-Forward suffers progressively deeper post-switch dips. (b) PRIME leads on return, coverage, and served users, with the lowest collision rate among the compared methods.

at 10–18% versus MAPPO's 40–45%), an investment whose value becomes clear under non-stationarity. Full Normal-mode curves and bar charts are reported in Appendix E of the supplementary material.

Change mode. Under cyclic phase switching, the ordering reverses (Fig. 6a). All three methods dip at each phase boundary, but they differ in dip depth and recovery speed. PRIME recovers to within-phase peaks of ≈ 140 in each phase, whereas MAPPO peaks near 85 and PRIME-Forward near 80 in later phases. However, PRIME-Forward suffers much deeper post-switch dips than MAPPO, and the dips deepen progressively, reflecting compounding disruption from indiscriminate resets.

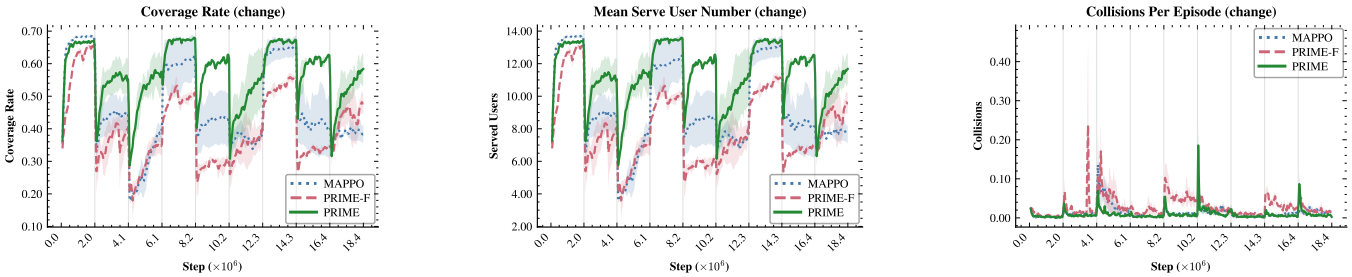
In terms of IQM return (Fig. 6b), PRIME reaches 72.626, compared with 58.135 for MAPPO (+24.9%) and 39.410 for PRIME-Forward. The PRIME-Forward gap is the more informative one: forward-only resets destroy gradient-active neurons precisely when the optimizer needs stable learning signals to adapt to a new phase.

ECN metrics. Figure 7 reports three application-level metrics. PRIME recovers coverage fastest after each transition (~ 0.68 at within-phase peaks, versus ~ 0.60 for MAPPO and ~ 0.45 for PRIME-Forward in later phases) and serves the most users (~ 14 per slot at peak, versus $\sim$9–11); PRIME-Forward's deepest post-switch dips ($\sim$4–5 served users) reflect the transient disruption of over-resetting. All methods produce brief collision spikes at phase boundaries as trajectories are re-planned; among the compared methods, PRIME maintains the lowest collision rate (IQM $= 0.005$), suggesting that the selective reset criterion preserves neurons responsible for inter-UAV separation while reclaiming genuinely inactive capacity.

C. Plasticity Diagnostics

We examine four neural-network-level indicators that explain the performance differences above.

Dormant neuron fraction. Figure 8(a) shows that PRIME maintains the lowest DNF throughout Change-mode training ($\sim$10–20%), compared with MAPPO ($\sim$40–45%) and PRIME-Forward ($\sim$15–30%). PRIME's DNF exhibits phase-synchronous oscillations: a brief rise at each phase boundary followed by a drop upon the next reset sweep. This pattern



(a) Coverage rate (b) Served users (c) Collisions

Fig. 7: ECN performance in Change mode: (a) coverage rate, (b) served users, and (c) collision rate; PRIME is lowest among the compared methods on (c).

confirms that the bidirectional criterion targets genuinely silent neurons at each phase transition.

Feature rank. MAPPO retains the highest effective rank ($\approx$8–10; Fig. 8(b)). This may appear favorable, but is consistent with noisy activations from its $\sim 40\%$ dormant neurons inflating statistical diversity in the feature matrix without a corresponding gain in policy quality. PRIME's rank ($\approx$7–8) is lower but corresponds to a more coherent representation where every active neuron contributes to the output. PRIME-Forward's rank collapses to $\approx$5–6 over time, confirming that indiscriminate resets destroy learned structure faster than the optimizer can rebuild it.

Policy entropy. PRIME sustains higher entropy across phase transitions (Fig. 8(c)), retaining the exploratory capacity needed to adapt when the objective changes. MAPPO's entropy declines monotonically, consistent with a shrinking effective action space as dormancy accumulates.

Silent neuron decomposition. Figure S10 in the supplementary material decomposes PRIME's neuron population into three sets per layer: forward-dormant ($\mathcal{D}_l$), backward-silent ($\mathcal{G}_l$), and their intersection ($\mathcal{S}_l = \mathcal{D}_l \cap \mathcal{G}_l$). Policy Layer 0 is the highest-dormancy layer under PRIME, consistent with Section II-A's finding that input-adjacent layers accumulate dormancy most aggressively; the team-aggregate DNF reported as 10–20% in Fig. 8(a) averages across all four hidden layers and is pulled down by the lower-dormancy value layers. In Policy Layer 0 specifically, $|\mathcal{G}_l|/H_l \approx$ 60–80%, $|\mathcal{D}_l|/H_l \approx$ 30–40%, and $|\mathcal{S}_l|/H_l \approx$ 20–30%. Two facts follow. First, $|\mathcal{D}_l \setminus \mathcal{G}_l|/|\mathcal{D}_l| \approx$ 20–30% of forward-dormant neurons are preserved by the bidirectional criterion because they still receive non-trivial gradient signal. Second, this preserved fraction is the empirical counterpart to the forward-only false-positive lower bound of 50–90% measured on vanilla MAPPO in Section II-B: PRIME's bidirectional gate brings the equivalent rate of would-be over-resets down by a factor of two to three, directly confirming the diagnostic from Section II.

D. Ablation Studies

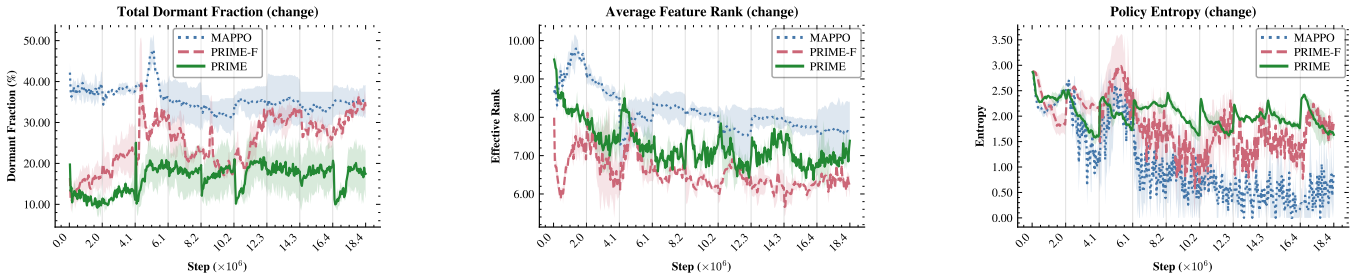
We ablate seven design choices in Change mode: the two detection thresholds and the reset period (with PRIME-Forward as the forward-only reference), phase-signal dependence, and the three design-axis substitutions of Table II (backward-signal source, aggregation granularity, and intervention operator).

Dormancy threshold τ_d. We vary $\tau_d \in \{0.05, 0.5, 1.0\}$ with τ_g=0.08 and F=200 fixed (Fig. 9(a)). PRIME's IQM return stays above 64 across the full range, with only a mild decrease at τ_d=1.0 where the dormancy criterion becomes overly inclusive. PRIME-Forward drops from $\sim$69 at τ_d=0.05 (few neurons qualify) to $\sim$28 at τ_d=1.0 (nearly all neurons are reset). This divergence illustrates the protective effect of the gradient-silence criterion: even when the forward threshold admits many neurons, the backward filter prevents PRIME from resetting those that are still receiving useful gradient signal.

Gradient threshold τ_g. Varying $\tau_g \in \{0.03, 0.08, 0.15\}$ with τ_d=0.5 fixed (Fig. 9(b)), PRIME's IQM return rises from $\sim$70 at τ_g=0.03 to 72.626 at the default τ_g=0.08, and dips slightly to $\sim$71 at 0.15. A moderately permissive gradient threshold captures the most beneficial resets without over-inclusion. PRIME-Forward is invariant to τ_g (IQM = 39.410 at all three values), since it ignores gradient information by construction.

Reset period F. With τ_d=0.5 and τ_g=0.08, we test $F \in \{50, 200, 300\}$ (Fig. 10(a)). PRIME's IQM return stays within 71–73 across all three values, confirming that the intersection criterion is selective enough to prevent harmful resets even at high frequency. PRIME-Forward is sensitive to F: its IQM rises from $\sim$39 at F=50 (frequent over-resetting) to $\sim$62 at F=300 (less frequent resets allow partial recovery between interventions).

Phase-signal independence. Algorithm 1 specifies a purely periodic mechanism ($c \bmod F = 0$, no knowledge of phase boundaries), but the main experimental runs (variant A in Fig. 10(b)) additionally trigger a full reset sweep at each boundary; whether the periodic-only mechanism suffices is therefore an empirical question, answered by comparing three configurations. Variant B removes the phase-triggered sweep, leaving exactly the procedure of Algorithm 1; variant C further shifts the period to a co-prime F=211, so the offset between resets and regime changes drifts across phases and is effectively randomized. PRIME's IQM return is 72.626, 75.301, and 55.790 under A, B, and C respectively; coverage rates follow the same ordering (0.589, 0.617, 0.520). Variant B matches and slightly exceeds the default, demonstrating



(a) Total dormant fraction (b) Average feature rank (c) Policy entropy

Fig. 8: Plasticity diagnostics (Change mode). (a) PRIME maintains the lowest dormant fraction ($\sim$10–20%) through periodic selective resets. (b) MAPPO's high rank is inflated by noisy dormant-neuron activations; PRIME's lower rank reflects a more coherent representation. (c) PRIME retains higher entropy, preserving exploration across phase transitions.

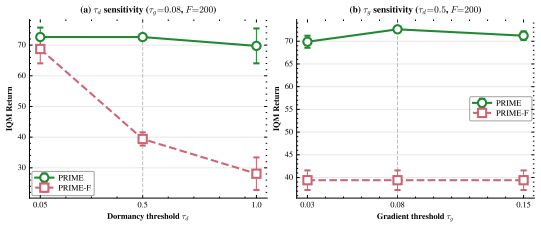
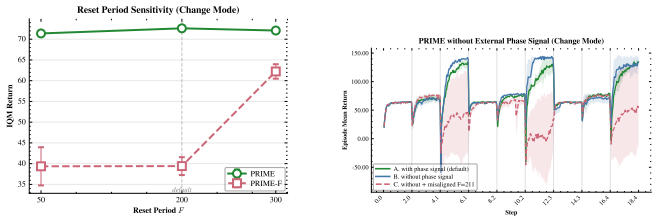


Fig. 9: Threshold sensitivity (Change mode). (a) IQM return vs. $\tau_d \in \{0.05, 0.5, 1.0\}$ (τ_g=0.08, F=200). PRIME is robust across the range; PRIME-Forward degrades sharply at larger τ_d. (b) IQM return vs. $\tau_g \in \{0.03, 0.08, 0.15\}$ (τ_d=0.5, F=200). PRIME improves with moderately permissive τ_g; PRIME-Forward is invariant.



(a) Reset period F (b) Phase-signal independence

Fig. 10: Reset-schedule ablations (Change mode). (a) IQM return for $F \in \{50, 200, 300\}$: PRIME is stable across reset periods; PRIME-Forward improves with less frequent resets. (b) Three reset-schedule configurations: A: default schedule from the main experiments, which augments periodic detection (F=200) with a full reset sweep at each phase boundary; B: pure periodic detection at F=200, as specified by Algorithm 1; C: pure periodic detection at F=211. PRIME's IQM return is comparable in A and B and degrades gracefully in C, confirming that the periodic detection mechanism alone is sufficient.

that the phase-triggered augmentation is not required—the periodic detection in Algorithm 1 suffices on its own. Variant C degrades gracefully rather than failing: with the schedule intentionally adversarial to the phase structure, performance stays comparable to MAPPO (58.135) and well above the forward-only baseline, though with the largest cross-seed spread. Together these results verify the design claim in Section IV: PRIME's plasticity recovery is decoupled from any external signal about when the environment changes.

The remaining three ablations vary choices that Algorithm 1 fixes implicitly: where the backward signal originates, over whose experience the detection statistics aggregate, and which operator acts on the silent set. Each variant changes exactly one choice while keeping thresholds, layer-wise normalization, the intersection gate, and the period F=200 identical to the main configuration. Table II reports Change-mode results; the Seeds column distinguishes two-seed protocol values from single-seed sweep points.

Backward-signal source. PRIME's gradient gate reads the gradient that the PPO training loss has already deposited on the weights. The *AuxGrad* variant replaces this signal with the loss-independent output-sensitivity gradient used by ReSiN [19], $g_W = \nabla_W \sum_n f_W(x_n)$, obtained from one additional forward–backward pass per network at each detection event; the per-neuron reduction, layer-wise normalization, intersection gate, and reset operator are unchanged. Because this proxy is not scaled by the optimizer that produced it, its threshold requires calibration, and we sweep $\tau_g^{\text{aux}} \in \{0.03, 0.08, 0.15, 0.30\}$ on seed 42 before committing to a two-seed comparison at the best value. The sweep spans 47 IQM points: 34.095, 45.146, 81.208, and 71.079 respectively, with two of the four settings falling below vanilla MAPPO

TABLE II: Design-Axis Ablations of the Detection Pipeline (Change Mode)

Variant	Modified choice	τ_g	Seeds	Return IQM
MAPPO	no reset	–	42, 43	58.135
PRIME-Forward	gradient gate removed	–	42, 43	39.410
PRIME	none (reference)	0.08	42, 43	72.626 [72.2, 73.0]
AuxGrad	signal → output gradient	0.03	42	34.095
		0.08	42	45.146
		0.15	42, 43	73.439 [65.7, 81.2]
		0.30	42	71.079
SingleSlice	statistics → agent-0 slice	0.08	42, 43	64.480 [60.0, 69.0]
Noise	reset → perturbation	0.08	42, 43	74.414 [73.0, 75.8]

(58.135). Over the same grid, the training-gradient signal varies by roughly 2.6 points (Fig. 9(b)). Cumulative reset counts track the swing. At 0.03 the gate essentially never opens: 5,012 resets over the full run, 3.9% of PRIME's volume on the same seed (127,492), so the variant degenerates to vanilla MAPPO with detection overhead. At 0.08 the reset schedule repeatedly stalls, flattening after roughly 10M steps at 49.8% of PRIME's volume, while the value-network zero-activation fraction spikes above 60% inside the stalled windows. At 0.15 clearing is continuous ($1.27\times$ PRIME's volume) and performance reaches parity. The tuned threshold does not transfer across seeds, however. On seed 43 the same setting reproduces the stalling mode, with no interventions after roughly 11M steps and 48% of the seed-42 reset volume, and lands at 65.669, which is 6.6 points below PRIME's weaker seed. The resulting two-seed mean, 73.439, carries a 15.5-point seed range, against 0.8 points for PRIME (72.246 and 73.006). Cumulative volume does not separate the two signals: PRIME's own count varies from 127,492 to 231,085 across its seeds while performance moves by less than a point. What separates them is whether the gate stays responsive. PRIME clears continuously on both seeds; the proxy's gate closes outright on one of them, and performance falls with it. In the stationary Normal mode the Change-tuned threshold transfers: AuxGrad reaches 77.488 against PRIME's 78.352, with a 0.04-point seed band; we note that the Change-mode stalls emerged only after roughly 10M steps, beyond the 3.07M-step horizon of the stationary runs. The training gradient thus costs no additional computation, arrives pre-scaled by the optimizer, and holds performance within 2.6 points over the threshold grid and 0.8 points across seeds; the output-sensitivity proxy reaches parity only after a per-environment threshold search and remains seed-fragile at its best setting under non-stationarity.

Aggregation granularity. Section II-C measured 5–18% cross-agent disagreement in per-agent dormancy statistics; the *SingleSlice* variant quantifies the behavioral cost of acting on one agent's view. Both criteria are computed from agent 0's $(B\times1)$ slice: forward statistics from that agent's rows, and the gradient from the PPO loss restricted to the same rows, while resets still apply to the shared network. This emulates a single-agent detector operating inside a shared-parameter system. The variant keeps part of the benefit, 64.480 against MAPPO's 58.135, but gives up 8.1 IQM points relative to PRIME, an 11.2% drop, with disjoint per-seed ranges: its stronger seed (68.957) remains below PRIME's weaker one (72.246). Coverage falls from 0.589 to 0.543 while collisions hold at PRIME's 0.005, so the cost concentrates in service quality rather than safety. The disagreement measured in Section II-C is the phenomenon; the 8.1-point gap is its consequence: team-level aggregation is not only the well-posed choice but a measurably better one.

Intervention operator. The *Noise* variant keeps detection unchanged and replaces the output-preserving reset with the perturbation family used by ReSiN [19]: additive Gaussian noise with standard deviation $0.1\times$ the layer's initialization bound, applied to the incoming weights and bias of the silent set with no outgoing zeroing, at the matched cadence $F{=}200$. Performance matches PRIME: 74.414 against 72.626, with overlapping per-seed ranges and near-identical intervention volumes across seeds (148,850 and 148,041, within 0.6%, versus 127,492 for PRIME on the same seed). The roughly $1.17\times$ higher volume is consistent with perturbation lifting a neuron out of dormancy more gradually than reinitialization, so that some neurons are treated repeatedly. Two conclusions follow. The performance load rests on detection: once the intersection criterion selects the neurons, hard reset and stochastic perturbation are interchangeable at matched cadence. And the main-experiment gains do not depend on the output-preserving convention; we retain the hard reset as the default for its output continuity at the moment of intervention.

Activation function. All experiments use ReLU, whose exact zeros are the natural setting for dormancy detection; for smooth rectifiers (Leaky-ReLU, GELU, Swish) the normalized threshold τ_d already implements near-zero-magnitude detection, and the gradient criterion is activation-agnostic by construction. We therefore expect PRIME to transfer, with the optimal τ_d possibly shifting; empirical validation is left to future work.

Summary. Across all seven ablation axes (dormancy threshold, gradient threshold, reset period, phase-signal independence, backward-signal source, aggregation granularity, and intervention operator), PRIME is robust to every implementation choice, while the sensitivities that remain are exactly the claimed ones. Substituting the training-gradient signal exposes a 47-point threshold sensitivity and a 15.5-point seed spread; restricting aggregation to a single agent's slice costs 8.1 IQM points; exchanging the reset operator for a matched-cadence perturbation costs nothing; and wherever the forward-only criterion admits many resets, PRIME-Forward degrades sharply while PRIME does not. This confirms the central claim of Section IV-E: the intersection criterion $\mathcal{S}_l = \mathcal{D}_l \cap \mathcal{G}_l$ restricts resets to the genuinely expendable subspace, keeping the perturbation term $\bar{d}^*$ in the regret bound (17) small. Neurons in $\mathcal{D}_l \setminus \mathcal{G}_l$ carry non-zero gradients, are actively participating in learning, and must not be reset.

E. Computational Overhead

PRIME adds three costs to vanilla MAPPO: per-neuron activation logging during the forward pass, an ℓ_1 reduction over each neuron's weight gradients at detection events (the gradients themselves are reused from the PPO backward pass), and the reset writes, which touch $|\mathcal{S}_l| \approx$ 2–5 of $H{=}32$ neurons per layer in practice (Fig. S13 in the supplementary material). In our setting ($H{=}32$, $n{=}3$, $F{=}200$), wall-clock training time increases by approximately 3.8% over MAPPO (23h12m vs. 22h21m for $18.4{\times}10^6$ environment steps on the same hardware). Per-neuron scoring is $O(H \cdot d_{l,\text{in}})$ per layer, dominated by the PPO forward and backward passes that scale identically, and the detection logic operates on per-neuron aggregates, adding no cost in n beyond the batch aggregation itself. Memory overhead is two floats per neuron (512 bytes for the current architecture).

VI. Related Work

A. UAV-Assisted ECN and Non-Stationarity

UAV swarms for post-disaster connectivity restoration have been studied from trajectory optimization [5], [26] to energy-aware scheduling and user association [36], [37]; Zhao et al. [38] provide a recent survey of deep reinforcement learning (DRL)-based UAV communications. Single-agent DRL has been applied to adaptive placement and spectrum management [39], and cooperative MARL formulations extend this to multi-UAV coordination using CTDE architectures [40], [41]. PE-MAMoE [22] introduces sparse MoE routing with phase-aware noise injection for objective switching. These works, however, assume that the policy network retains its representational capacity across phase boundaries.

When the environment does change, MARL methods typically handle non-stationarity by modeling or anticipating the external shift. Experience-replay stabilization [42] and opponent modeling [43], [44] address the policy-drift form of non-stationarity that arises from co-learning agents. Meta-RL [45], [46] targets task-distribution shift but requires representative meta-training tasks. Wang et al. [13] recently proposed MACPH for non-stationary environments with unknown change points, combining composite replay buffers with adaptive parameter-space noise; the noise, however, is applied globally to actor parameters without neuron-level selectivity. Galashov et al. [47] introduced Soft Reset, which models parameter drift explicitly and pulls weights toward their initialization, but the reset is uniform across all parameters rather than targeted at silent neurons. A common limitation runs through these approaches: they assume, implicitly or explicitly, that environmental change can be tracked through external signals such as reward shifts, transition changes, or task labels. None addresses the network-level capacity degradation that compounds across phase switches, the problem that PRIME targets.

B. Neural Plasticity Loss and Reset Methods

Plasticity loss has been studied along two axes: diagnosis and mitigation. On the diagnostic side, Dohare et al. [14] established that continual training degrades deep networks to the point where they offer no advantage over single-layer models. Sokar et al. [18] formalized the dormant neuron fraction as a proxy for capacity loss and proposed the ReDo algorithm, which reinitializes neurons whose forward activations fall below a threshold. Lyle et al. [15], [48] connected this neuron-level phenomenon to a geometric signature at the representation level: the effective dimensionality of learned features shrinks as dormancy accumulates. Juliani and Ash [49] showed that on-policy PPO is equally susceptible, with several off-policy fixes such as CReLU and plasticity injection [33] failing to transfer.

On the mitigation side, existing methods span a spectrum from coarse to fine-grained. Global perturbation approaches, including continual backpropagation [14], the primacy-bias reset [21], weight clipping [50], and regenerative regularization [51], perturb or constrain the full parameter space without distinguishing active neurons from inactive ones. MoE architectures [52] can reduce dormancy through structural diversity [53], but add routing overhead and load-balancing losses that complicate on-policy training [54]. ReDo [18] and plasticity injection [33] are more targeted, resetting or augmenting individual neurons, yet both rely solely on forward-activation scores and cannot distinguish genuinely expendable neurons from those receiving strong gradient signal during phase adaptation. ReSiN [19] introduced the Silent Neuron activity index, combining forward dormancy with backward gradient silence into a single criterion: only neurons inactive in both directions qualify for reset. The method was validated on single-agent adaptive video streaming with tracking-error guarantees. ReSiN itself builds on PA-MoE [55], which addressed plasticity under QoE shifts in the same setting through a plasticity-aware mixture-of-experts controller.

In the multi-agent setting, the dormancy problem takes a distinct form. Qin et al. [56] found that dormant neurons in QMIX concentrate in the mixing network and that their prevalence grows with team size; their ReBorn method redistributes weights from over-active to dormant neurons through a monotonicity-aware transfer. Yuan et al. [41] showed that coordination skills degrade across regime shifts even when individual performance is maintained, and proposed progressive task contextualization to mitigate inter-agent forgetting. Neither ReBorn nor progressive contextualization uses backward gradient information to guide resets. PRIME extends the bidirectional criterion of ReSiN to shared-parameter CTDE. The extension is not cosmetic: Section II-C shows that per-agent dormancy disagreement reaches 5–18% in policy layers under vanilla MAPPO, so the single-rollout activation and gradient statistics used by ReSiN do not directly transfer to multi-agent training without team-level aggregation. PRIME instead computes activation and gradient scores over the full $B\times n$ team-aggregated batch, which both eliminates per-agent inconsistency and exploits the richer statistics afforded by cooperative training. Section V makes the comparison empirical: ReSiN's output-sensitivity signal reaches PRIME's Change-mode performance only after a per-environment threshold sweep and loses it to gate closure on one seed, while ReSiN's perturbation operator, paired with PRIME's detector at matched cadence, performs on par with the hard reset. To our knowledge, PRIME is the first method in multi-agent communication systems to detect non-stationarity through internal neuron-level plasticity monitoring rather than external environment classification or phase signals.

VII. Conclusion

This paper argued that plasticity loss is a fundamental, under-addressed bottleneck in non-stationary UAV-ECN control, and proposed PRIME to address it through internal monitoring rather than external environment modeling. PRIME uses the shared-network batch structure of CTDE to compute team-level activation and gradient statistics, identifies neurons that are simultaneously forward-dormant and backward-silent, and reinitializes only this intersection set with an output-preserving reset. The procedure requires no architectural changes, auxiliary losses, or phase-aware scheduling.

Direct measurement on a vanilla MAPPO controller established the three findings behind the design: dormancy accumulates and persists, forward dormancy decouples from the actual training gradient in three of four hidden layers, and per-agent dormancy decisions disagree in policy layers. PRIME answers each in turn with periodic resets, the bidirectional intersection, and team-level aggregation.

A dynamic regret bound, obtained by extending ReSiN's single-agent tracking analysis to the team-averaged objective, shows that the perturbation energy of PRIME's resets scales with the empirical silent-subspace dimension d_t^* rather than the full parameter count d, providing a tighter tracking guarantee than undiscriminated noise injection. On a phase-switching UAV-ECN simulator, PRIME improves IQM return by 24.9% over vanilla MAPPO and outperforms the forward-only baseline (PRIME-Forward) on return, coverage, served users, and collision rate. Ablation studies confirm robustness to the dormancy threshold, gradient threshold, reset period, and reset–phase alignment, and locate the performance-critical choices in the training-gradient signal and team-level aggregation rather than in the intervention operator; per-neuron diagnostics validate that dormant fractions remain at 10–20% throughout training.

Several directions remain open. The fixed reset period F could be replaced by an adaptive trigger driven by online monitoring of DNF or feature rank; heterogeneous agent architectures would require a principled scheme for aggregating activation statistics across distinct sub-networks; hardware-in-the-loop validation would test robustness under real actuator latency and sensor noise; and combining silent neuron resets with complementary plasticity techniques such as layer normalization [57] or spectral regularization may yield further gains.

REFERENCES

[1] N. Zhao, W. Lu, M. Sheng, Y. Chen, J. Tang, F. R. Yu, and K.-K. Wong, "Uav-assisted emergency networks in disasters," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 45–51, 2019.

[2] M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, and M. Debbah, "A tutorial on uavs for wireless networks: Applications, challenges, and open problems," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2334–2360, 2019.

[3] B. Li, Z. Fei, and Y. Zhang, "Uav communications for 5g and beyond: Recent advances and future trends," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2241–2263, 2019.

[4] Y. Zeng, J. Lyu, and R. Zhang, "Cellular-connected uav: Potential, challenges, and promising technologies," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 120–127, 2019.

[5] Q. Wu, Y. Zeng, and R. Zhang, "Joint trajectory and communication design for multi-uav enabled wireless networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 2109–2121, 2018.

[6] C. H. Liu, Z. Chen, J. Tang, J. Xu, and C. Piao, "Energy-efficient uav control for effective and fair communication coverage: A deep reinforcement learning approach," *IEEE J.Sel. A. Commun.*, vol. 36, no. 9, p. 2059–2070, Sep. 2018. [Online]. Available: https://doi.org/10.1109/JSAC.2018.2864373

[7] L. Wang, K. Wang, C. Pan, W. Xu, N. Aslam, and A. Nallanathan, "Deep reinforcement learning based dynamic trajectory control for uav-assisted mobile edge computing," *IEEE Trans. Mobile Comput.*, vol. 21, no. 10, pp. 3536–3550, 2022.

[8] Y. Jiang, D. Zhai, M. Yang, Z. Lin, and Y. Li, "Non-position-based uav trajectory optimization for coverage maximization," in *Proc. ACM MobiCom Workshop Drone Assisted Wireless Commun. 5G Beyond*, ser. DroneCom '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 67–72. [Online]. Available: https://doi.org/10.1145/3555661.3560866

[9] Y. Xing, C. N. Mathur, M. A. Haleem, R. Chandramouli, and K. Subbalakshmi, "Dynamic spectrum access with qos and interference temperature constraints," *IEEE Trans. Mobile Comput.*, vol. 6, no. 4, pp. 423–433, 2007.

[10] W. Xu, T. Zhang, X. Mu, Y. Liu, and Y. Wang, "Trajectory planning and resource allocation for multi-uav cooperative computation," *IEEE Trans. Commun.*, vol. 72, no. 7, pp. 4305–4318, 2024.

[11] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," 2016. [Online]. Available: https://arxiv.org/abs/1511.05952

[12] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, ser. ICML'17. JMLR.org, 2017, p. 1126–1135.

[13] S. Wang, Q. Yue, Z. Xu, P. Qiao, Z. Lyu, and F. Gao, "A collaborative multi-agent reinforcement learning approach for non-stationary environments with unknown change points," *Mathematics*, vol. 13, no. 11, 2025. [Online]. Available: https://www.mdpi.com/2227-7390/13/11/1738

[14] S. Dohare, J. F. Hernandez-Garcia, Q. Lan, P. Rahman, A. R. Mahmood, and R. S. Sutton, "Loss of plasticity in deep continual learning," *Nature*, vol. 632, no. 8026, pp. 768–774, 2024.

[15] C. Lyle, Z. Zheng, E. Nikishin, B. A. Pires, R. Pascanu, and W. Dabney, "Understanding plasticity in neural networks," in *Proc. Int. Conf. Mach. Learn.*, ser. ICML'23. JMLR.org, 2023.

[16] Z. Abbas, R. Zhao, J. Modayil, A. White, and M. C. Machado, "Loss of plasticity in continual deep reinforcement learning," in *Proc. Conf. Lifelong Learn. Agents*. PMLR, 2023, pp. 620–636.

[17] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of ppo in cooperative multi-agent games," in *Proc. Adv. Neural Inf. Process. Syst.*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.

[18] G. Sokar, R. Agarwal, P. S. Castro, and U. Evci, "The dormant neuron phenomenon in deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, ser. ICML'23. JMLR.org, 2023.

[19] Z. He and Z. Liu, "Silent neuron theory and plasticity preservation for deep reinforcement learning in adaptive video streaming," *IEEE Trans. Mobile Comput.*, 2025, early access.

[20] J. T. Ash and R. P. Adams, "On warm-starting neural network training," in *Proc. Adv. Neural Inf. Process. Syst.*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[21] E. Nikishin, M. Schwarzer, P. D'Oro, P.-L. Bacon, and A. Courville, "The primacy bias in deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2022, pp. 16 828–16 847.

[22] W. Qiu, Z. He, W. Zhao, and H. Masui, "Plasticity-enhanced multi-agent mixture of experts for dynamic objective adaptation in uavs-assisted emergency communication networks," 2026. [Online]. Available: https://arxiv.org/abs/2604.09028

[23] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017. [Online]. Available: https://arxiv.org/abs/1707.06347

[24] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," 3rd Generation Partnership Project, Technical Report TR 38.901, Nov. 2020.

[25] H. Yan, Y. Chen, and S.-H. Yang, "New energy consumption model for rotary-wing uav propulsion," *IEEE Wireless Commun. Lett.*, vol. 10, no. 9, pp. 2009–2012, 2021.

[26] Y. Zeng, J. Xu, and R. Zhang, "Energy minimization for wireless communication with rotary-wing uav," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2329–2345, 2019.

[27] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad hoc network research," *Wireless Commun. Mobile Comput.*, vol. 2, no. 5, pp. 483–502, 2002. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/wcm.72

[28] X. Hong, M. Gerla, G. Pei, and C.-C. Chiang, "A group mobility model for ad hoc wireless networks," in *Proc. ACM Int. Workshop Model. Anal. Simul. Wireless Mobile Syst.*, ser. MSWiM '99. New York, NY, USA: Association for Computing Machinery, 1999, p. 53–60. [Online]. Available: https://doi.org/10.1145/313237.313248

[29] C. Amato, "An introduction to centralized training for decentralized execution in cooperative multi-agent reinforcement learning," 2024. [Online]. Available: https://arxiv.org/abs/2409.03052

[30] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. Adv. Neural Inf. Process. Syst.*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6382–6393.

[31] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," 2018. [Online]. Available: https://arxiv.org/abs/1506.02438
[32] R.-A. Lascu, D. Šiška, and Łukasz Szpruch, "Ppo in the fisher-rao geometry," 2026. [Online]. Available: https://arxiv.org/abs/2506.03757
[33] E. Nikishin, J. Oh, G. Ostrovski, C. Lyle, R. Pascanu, W. Dabney, and A. Barreto, "Deep reinforcement learning with plasticity injection," in *Proc. Adv. Neural Inf. Process. Syst.*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.
[34] R. Agarwal, M. Schwarzer, P. S. Castro, A. Courville, and M. G. Bellemare, "Deep reinforcement learning at the edge of the statistical precipice," in *Proc. Adv. Neural Inf. Process. Syst.*, ser. NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2021.
[35] M. Samvelyan, T. Rashid, C. Schroeder de Witt, G. Farquhar, N. Nardelli, T. G. J. Rudner, C.-M. Hung, P. H. S. Torr, J. Foerster, and S. Whiteson, "The starcraft multi-agent challenge," in *Proc. Int. Conf. Auton. Agents Multiagent Syst.*, ser. AAMAS '19. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2019, p. 2186–2188.
[36] J. Sun, Z. Sheng, A. A. Nasir, Z. Huang, H. Yu, and Y. Fang, "Energy efficiency maximization for wpt-enabled uav-assisted emergency communication with user mobility," *Phys. Commun.*, vol. 61, no. C, Dec. 2023. [Online]. Available: https://doi.org/10.1016/j.phycom.2023.102200
[37] A. Hussain, S. Li, T. Hussain, X. Lin, F. Ali, and A. A. AlZubi, "Computing challenges of uav networks: A comprehensive survey." *Comput. Mater. Continua*, vol. 81, no. 2, 2024.
[38] W. Zhao, S. Cui, W. Qiu, Z. He, Z. Liu, X. Zheng, B. Mao, and N. Kato, "A survey on drl-based uav communications and networking: Drl fundamentals, applications and implementations," *IEEE Commun. Surveys Tuts.*, vol. 28, pp. 3911–3941, 2026.
[39] B. Badnava, T. Kim, K. Cheung, Z. Ali, and M. Hashemi, "Spectrum-aware mobile edge computing for uavs using reinforcement learning," in *Proc. IEEE/ACM Symp. Edge Comput. (SEC)*, 2021, pp. 376–380.
[40] J. N. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proc. AAAI Conf. Artif. Intell.*, ser. AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.
[41] L. Yuan, L. Li, Z. Zhang, F. Zhang, C. Guan, and Y. Yu, "Multiagent continual coordination via progressive task contextualization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 4, pp. 6326–6340, 2025.
[42] J. Foerster, N. Nardelli, G. Farquhar, T. Afouras, P. H. S. Torr, P. Kohli, and S. Whiteson, "Stabilising experience replay for deep multi-agent reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, ser. ICML'17. JMLR.org, 2017, p. 1146–1155.
[43] J. Foerster, R. Y. Chen, M. Al-Shedivat, S. Whiteson, P. Abbeel, and I. Mordatch, "Learning with opponent-learning awareness," in *Proc. Int. Conf. Auton. Agents Multiagent Syst.*, ser. AAMAS '18. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2018, p. 122–130.
[44] S. Zhao, C. Lu, R. Grosse, and J. Foerster, "Proximal learning with opponent-learning awareness," in *Proc. Adv. Neural Inf. Process. Syst.*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.
[45] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, ser. ICML'17. JMLR.org, 2017, p. 1126–1135.
[46] M. Al-Shedivat, T. Bansal, Y. Burda, I. Sutskever, I. Mordatch, and P. Abbeel, "Continuous adaptation via meta-learning in nonstationary and competitive environments," in *Proc. Int. Conf. Learn. Represent.*, 2018. [Online]. Available: https://openreview.net/forum?id=Sk2u1g-0-
[47] A. Galashov, M. Titsias, A. György, C. Lyle, R. Pascanu, Y. W. Teh, and M. Sahani, "Non-stationary learning of neural networks with automatic soft parameter reset," in *Proc. Adv. Neural Inf. Process. Syst.*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 83 197–83 234. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/978cc34c539fd26f0e8afb7e3905f34a-Paper-Conference.pdf
[48] C. Lyle, Z. Zheng, K. Khetarpal, H. van Hasselt, R. Pascanu, J. Martens, and W. Dabney, "Disentangling the causes of plasticity loss in neural networks," 2024. [Online]. Available: https://arxiv.org/abs/2402.18762
[49] A. Juliani and J. T. Ash, "A study of plasticity loss in on-policy deep reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, ser. NIPS '24. Red Hook, NY, USA: Curran Associates Inc., 2024.
[50] M. Elsayed, Q. Lan, C. Lyle, and A. R. Mahmood, "Weight clipping for deep continual and reinforcement learning," 2024. [Online]. Available: https://arxiv.org/abs/2407.01704
[51] S. Kumar, H. Marklund, and B. V. Roy, "Maintaining plasticity in continual learning via regenerative regularization," 2024. [Online]. Available: https://arxiv.org/abs/2308.11958
[52] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *Proc. Int. Conf. Learn. Represent.*, 2017. [Online]. Available: https://openreview.net/forum?id=B1ckMDqlg
[53] T. Willi, J. Obando-Ceron, J. Foerster, K. Dziugaite, and P. S. Castro, "Mixture of experts in a mixture of rl settings," 2024. [Online]. Available: https://arxiv.org/abs/2406.18420
[54] Y. Zhou, T. Lei, H. Liu, N. Du, Y. Huang, V. Y. Zhao, A. Dai, Z. Chen, Q. Le, and J. Laudon, "Mixture-of-experts with expert choice routing," in *Proc. Adv. Neural Inf. Process. Syst.*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.
[55] Z. He and Z. Liu, "Plasticity-aware mixture of experts for learning under qoe shifts in adaptive video streaming," 2025. [Online]. Available: https://arxiv.org/abs/2504.09906
[56] H. Qin, C. Ma, M. Deng, Z. Liu, S. Mei, X. Liu, C. Wang, and S. Shen, "The dormant neuron phenomenon in multi-agent reinforcement learning value factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 35 727–35 759. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/3eec5006051d9544e717067de3220198-Paper-Conference.pdf
[57] C. Lyle, Z. Zheng, K. Khetarpal, J. Martens, H. van Hasselt, R. Pascanu, and W. Dabney, "Normalization and effective learning rates in reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, ser. NIPS '24. Red Hook, NY, USA: Curran Associates Inc., 2024.

Wen Qiu received the Ph.D. degree in Co-creative Engineering from Kitami Institute of Technology, Kitami, Japan, in 2025. She is currently a Specially Appointed Researcher at Kitami Institute of Technology. Her research interests include deep reinforcement learning and emergency wireless communication networks, with a focus on intelligent resource allocation and network optimization for disaster response.

Zhiqiang He is currently pursuing a Ph.D. at the University of Electro-Communications, Tokyo, Japan. He received his M.S. degree in Control Science and Engineering from Northeastern University, Shenyang, China. His research interests include deep reinforcement learning and its applications. He previously worked at Baidu and InspirAI, where he developed a master-level AI for the card game Dou Di Zhu that outperformed professional players.

Wei Zhao (S'12-M'16) received his Ph.D. degree in the Graduate School of Information Sciences, Tohoku University. He is currently a Professor at the School of Computer Science and Technology, Anhui University of Technology. His research interests include deep reinforcement learning, edge computing, and resource allocation in wireless networks. He was the recipient of the IEEE WCSP-2014 Best Paper Award, and IEEE GLOBECOM-2014 Best Paper Award. He is a member of IEEE.

Hiroshi Masui received his Ph.D. in Science from Osaka University in 1998. He currently works at the Department of Information and Communication Engineering, and the Vice President of Kitami Institute of Technology. His research interests include theoretical nuclear physics, public transportation analysis, cloud optimization, and data-driven science.

Supplementary Material for "PRIME: Plasticity Recovery in Multi-Agent Environments for UAV-Assisted Emergency Communication Networks"

Wen Qiu, *Member, IEEE,* Zhiqiang He, *Member, IEEE,* Wei Zhao, *Member, IEEE,* and Hiroshi Masui

TABLE S1: Training Hyperparameters

Parameter	Symbol	Value
MAPPO (shared by all methods)		
Hidden layer width	H	32
Number of hidden layers	–	2
Activation function	–	ReLU
Learning rate	η	3×10^{-4}
Weight decay	–	10^{-4}
Rollout length	T_{roll}	2048
Mini-batches per update	–	32
PPO update epochs	K_{epoch}	8
Discount factor	γ	0.99
GAE parameter	λ_{GAE}	0.95
PPO clip ratio	ϵ	0.15
Value loss coefficient	c_v	2.0
Entropy coefficient	c_e	0.01
Max gradient norm	–	0.5
Optimizer	–	Adam (β_1=0.9, β_2=0.999)
Random seeds	–	42, 43
PRIME-specific		
Dormancy threshold	τ_d	0.5
Gradient-silence threshold	τ_g	0.08
Reset period	F	200 mini-batch steps
Reinitialization method	–	Kaiming uniform
Output-weight zeroing	–	Yes

This document contains the appendices of the main paper. Sections are lettered A–J as referenced from the main text; figures, tables, and equations carry the prefix S.

Appendix A
Training Hyperparameters

Table S1 lists the complete set of training hyperparameters shared by all three methods (MAPPO, PRIME-Forward, PRIME) and the PRIME-specific settings.

W. Qiu and H. Masui are with the Department of Information and Communication Engineering, Kitami Institute of Technology, Japan (e-mail: w-qiu@ieee.org; hgmasui@mail.kitami-it.ac.jp).

Z. He is with the Graduate School of Informatics and Engineering, University of Electro-Communications, Japan (e-mail: hezhiqiang@ieee.org).

W. Zhao is with the School of Computer Science and Technology, Anhui University of Technology, China (e-mail: zhaowei@ahut.edu.cn).

Corresponding authors: Wei Zhao and Hiroshi Masui.

Appendix B
Proof of Theorem 1

We first establish a lemma that quantifies the perturbation energy in the silent subspace.

Lemma 1 (Silent Subspace Noise Energy)**.** *Let Π_t be an orthogonal projection of rank d_t^*. Then $\mathbb{E}\|\Pi_t\zeta_t\|^2 = d_t^*$ for $\zeta_t \sim \mathcal{N}(0, I_d)$.*

Proof. Since $\Pi_t^2 = \Pi_t = \Pi_t^\top$, we have $\mathbb{E}\|\Pi_t\zeta_t\|^2 = \mathbb{E}[\zeta_t^\top \Pi_t \zeta_t] = \operatorname{tr}(\Pi_t) = d_t^*$. □

Proof of Theorem 1 of the main paper. Define the tracking error $e_t = w_t - w_t^*$ and the drift $\Delta_t = w_{t+1}^* - w_t^*$. Substituting the update rule the update rule of the main paper (Section IV-F) and using the first-order optimality condition $\nabla\mathcal{L}_t(w_t^*) = 0$:

$$\begin{aligned} e_{t+1} &= w_{t+1} - w_{t+1}^* \\ &= \underbrace{e_t - \eta\big(\nabla\mathcal{L}_t(w_t) - \nabla\mathcal{L}_t(w_t^*)\big)}_{A_t} - \Delta_t + \eta\gamma_{\text{r}}\Pi_t\zeta_t. \end{aligned} \tag{S1}$$

Step 1: Contraction. By μ-strong convexity and L-smoothness with $\eta \le 1/L$, the term A_t satisfies [1]:

$$\|A_t\|^2 \;\le\; (1 - \mu\eta)\|e_t\|^2. \tag{S2}$$

Step 2: Squaring and taking expectation. From (S1), since $\mathbb{E}[\zeta_t] = 0$ and ζ_t is independent of A_t and Δ_t:

$$\mathbb{E}\|e_{t+1}\|^2 = \mathbb{E}\|A_t - \Delta_t\|^2 + \eta^2\gamma_{\text{r}}^2 d_t^*, \tag{S3}$$

where we used Lemma 1. Applying the parameterized Young's inequality $\|a - b\|^2 \le (1 + \alpha)\|a\|^2 + (1 + \alpha^{-1})\|b\|^2$ with $\alpha = \mu\eta/(2 - \mu\eta)$, so that $1 + \alpha^{-1} = 2/(\mu\eta)$ exactly:

$$\begin{aligned} \mathbb{E}\|A_t - \Delta_t\|^2 &\le \tfrac{2(1-\mu\eta)}{2-\mu\eta}\,\mathbb{E}\|e_t\|^2 + \tfrac{2}{\mu\eta}\|\Delta_t\|^2 \\ &\le \big(1 - \tfrac{\mu\eta}{2}\big)\,\mathbb{E}\|e_t\|^2 + \tfrac{2}{\mu\eta}\|\Delta_t\|^2, \end{aligned} \tag{S4}$$

where the second inequality uses $\frac{2(1-\mu\eta)}{2-\mu\eta} \le 1 - \frac{\mu\eta}{2}$, which holds for all $\mu\eta \in (0, 1]$ and in particular for $\eta \le 1/L \le 1/\mu$.

Step 3: Combining. Substituting (S4) into (S3):

$$\mathbb{E}\|e_{t+1}\|^2 \;\le\; \big(1 - \tfrac{\mu\eta}{2}\big)\,\mathbb{E}\|e_t\|^2 + \tfrac{2}{\mu\eta}\|\Delta_t\|^2 + \eta^2\gamma_{\text{r}}^2 d_t^*. \tag{S5}$$

Step 4: Telescoping. Let $\rho = 1 - \mu\eta/2$. Unrolling (S5) from $t = 0$ to $T - 1$:

$$\mathbb{E}\|e_t\|^2 \;\le\; \rho^t\,\mathbb{E}\|e_0\|^2 + \tfrac{2}{\mu\eta}\sum_{k=0}^{t-1}\rho^{t-1-k}\|\Delta_k\|^2 + \eta^2\gamma_{\text{r}}^2\sum_{k=0}^{t-1}\rho^{t-1-k}d_k^*. \tag{S6}$$

Averaging from $t = 1$ to T and using $\frac{1}{T}\sum_{t=1}^{T}\rho^t \le \frac{2}{\mu\eta T}$ and $\sum_{k=0}^{t-1}\rho^{t-1-k} \le \frac{2}{\mu\eta}$:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|e_t\|^2 \le \frac{2}{\mu\eta T}\,\mathbb{E}\|e_0\|^2 + \frac{4}{\mu^2\eta^2 T}\sum_{k=0}^{T-1}\|\Delta_k\|^2 + \frac{2\eta\gamma_{\mathrm{r}}^2}{\mu}\cdot\frac{1}{T}\sum_{t=1}^{T} d_t^*. \tag{S7}$$

Because the cross terms are nonnegative, $\sum_{k=0}^{T-1}\|\Delta_k\|^2 \le \left(\sum_{k=0}^{T-1}\|\Delta_k\|\right)^2 = P_T^2$. Collecting the first two terms yields the bound of Theorem 1. □

Appendix C
Remark on Assumptions of Theorem 1

Assumptions 1 and 2 (smoothness and local strong convexity of the PPO surrogate loss) are standard in the PPO tracking-error literature [1]–[4]; they hold locally within the trust region enforced by the PPO clip constraint, not globally over the full parameter space. The additive-noise model for resets (Assumption 4) is a tractable proxy for three operations performed during each PRIME reset: (i) Kaiming reinitialization, whose per-neuron perturbation energy is $\mathbb{E}\|\Delta\mathbf{w}_{l,i}\|^2 = 1$ independent of layer width; (ii) output-weight zeroing, which makes the immediate network-output perturbation exactly zero; and (iii) Adam state clearing, which transiently raises the effective learning rate for reset neurons. Because the Gaussian proxy upper-bounds (i) and ignores the perturbation-reducing effect of (ii), the bound in Theorem 1 of the main paper is conservative: it overestimates perturbation energy and thus understates the advantage of selective over blanket resets. The operator ablation in Section V of the main paper supports the proxy directly: replacing PRIME's reset with a matched-cadence Gaussian perturbation of the same silent set leaves Change-mode performance on par (74.414 versus 72.626 IQM), consistent with the bound's dependence on the intervention subspace rather than on the specific operator. The primary value of the theorem is qualitative: it identifies d_t^* as the key quantity separating selective from blanket perturbation, a prediction the ablation experiments confirm.

Appendix D
Normal-Mode Motivation Measurements

Figures S1–S3 collect the Normal-mode counterparts of the vanilla-MAPPO measurements in Section II of the main paper.

The qualitative picture matches Change mode: Value Layer 1 saturates near 60% with DNF and LDO coinciding, the forward-dormant set exceeds the backward-silent set in the same three layers, and the false-positive lower bound occupies comparable per-layer bands, confirming that the three findings are properties of the training dynamics rather than artifacts of phase switching.

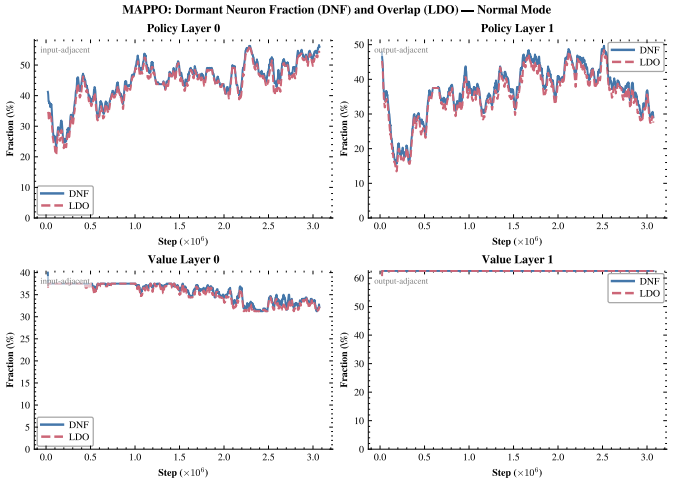


Fig. S1: Normal-mode DNF and LDO (3.07×10^6 steps); cf. Fig. 1 of the main paper.

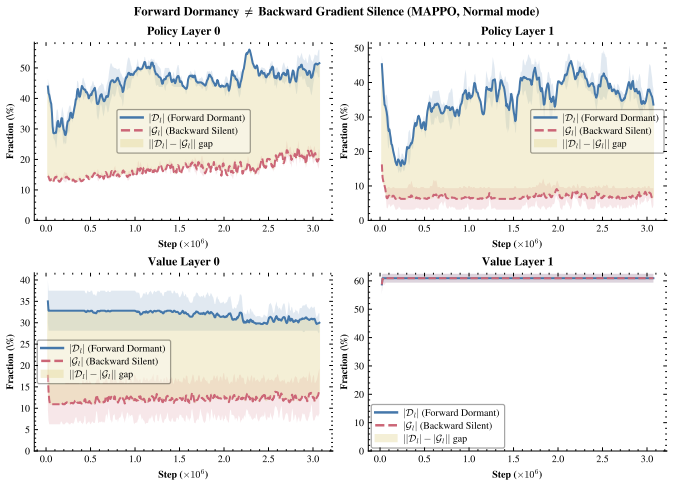


Fig. S2: Normal-mode forward–backward decoupling; cf. Fig. 2 of the main paper.

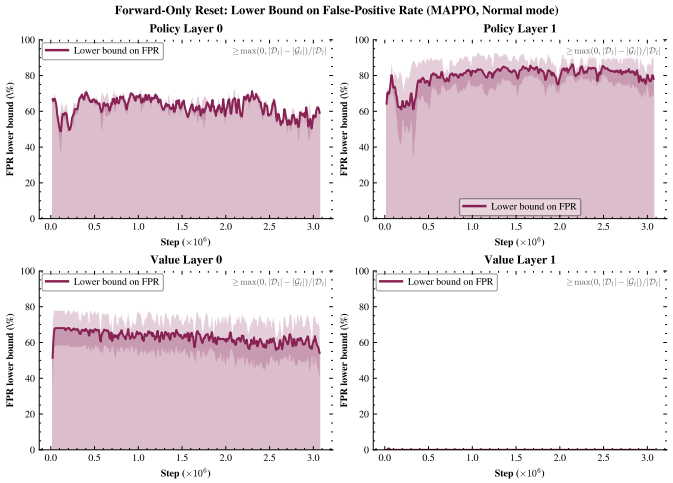
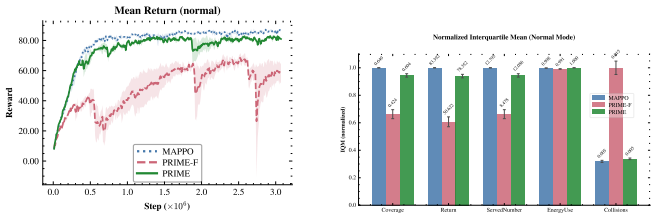
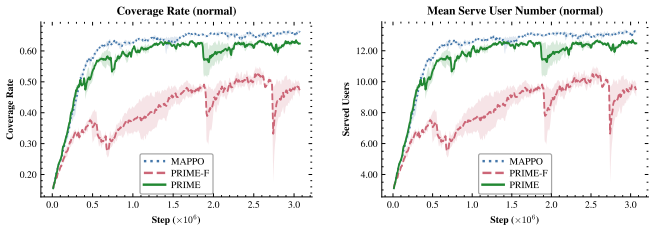


Fig. S3: Normal-mode false-positive lower bound; cf. Fig. 3 of the main paper.



(a) Training reward (b) Normalized IQM, five ECN metrics

Fig. S4: Normal-mode task performance. (a) MAPPO attains the highest asymptotic return; PRIME converges close behind; PRIME-Forward shows instability near Step 2.5×10^6. (b) MAPPO leads on return and coverage, PRIME is competitive, and PRIME-Forward trails on return.



(a) Coverage (Normal) (b) Served users (Normal)

Fig. S5: ECN task metrics in Normal mode. (a) MAPPO and PRIME converge to comparable coverage ($\sim$0.63–0.65); PRIME-Forward lags and drops near Step 2.5×10^6. (b) Served-user counts are similar for MAPPO and PRIME ($\sim$12–13).

Appendix E
Normal-Mode Task Performance

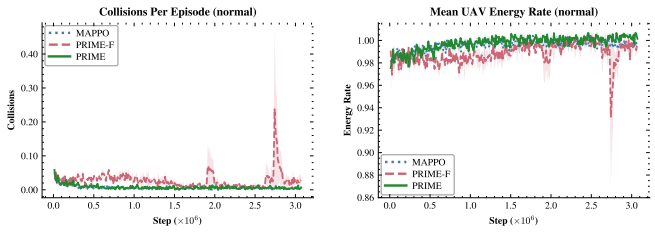
The main text focuses on Change mode, where plasticity loss is most pronounced. This appendix presents the corresponding Normal-mode results.

Fig. S4a shows the training reward curves. MAPPO attains the highest asymptotic return ($\sim$85), PRIME converges to $\sim$80, and PRIME-Forward shows instability near Step 2.5×10^6 caused by an over-resetting event that destabilizes the learned trajectory policy.

Fig. S4b reports the normalized IQM across the five ECN metrics. MAPPO leads on Return (83.392) and Coverage (0.640); PRIME is competitive (Return 78.352, Coverage 0.604); PRIME-Forward trails on Return (50.622).

Fig. S5 shows coverage rate and served-user count. MAPPO and PRIME converge to comparable coverage ($\sim$0.63–0.65) and served-user counts ($\sim$12–13), confirming that PRIME's periodic resets incur only a marginal steady-state cost. PRIME-Forward lags behind both methods, consistent with the reward instability above.

Fig. S6 shows collision rate and energy consumption. All three methods maintain near-zero collision rates throughout training, indicating that safety-critical trajectory constraints are satisfied regardless of the plasticity mechanism. Energy usage rates are nearly identical across methods ($\sim$0.99–1.00), confirming that the reinitialization procedure introduces no measurable energy overhead.



(a) Collisions (Normal) (b) Energy rate (Normal)

Fig. S6: Safety and energy in Normal mode. (a) Near-zero collisions for all methods. (b) Energy rates are nearly identical ($\sim$0.99–1.00).

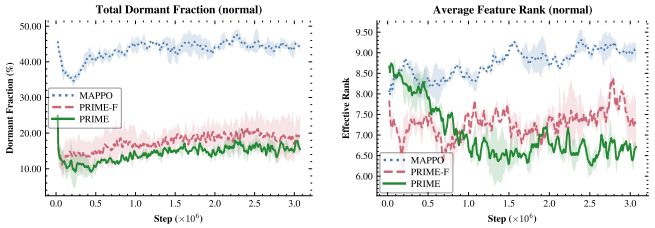
Appendix F
Normal-Mode Plasticity Diagnostics

This appendix complements the Change-mode diagnostics in Section V of the main paper with the corresponding Normal-mode indicators.

Fig. S7(a) shows the total dormant neuron fraction. MAPPO's DNF rises monotonically and plateaus at $\sim$40–45%, confirming that plasticity loss occurs even without phase switching. PRIME stabilizes at $\sim$10–18%. PRIME-Forward's DNF tracks PRIME closely ($\sim$10–18%), confirming that periodic resets effectively suppress activation-dormancy under the stationary objective regardless of whether the gradient-silence filter is applied. Note, however, that PRIME-Forward's near-identical dormancy curve in this stationary setting does *not* translate to comparable task performance: the abrupt return drop near Step 2.5×10^6 visible in Fig. S5 occurs without a corresponding dormancy spike, indicating that the resets disrupted useful representations even when the activation statistics suggested the network was healthy.

Fig. S7(b) shows the average feature rank. MAPPO maintains a higher effective rank ($\sim$9–10) than PRIME ($\sim$7) throughout training. As discussed in Section V of the main paper, this higher rank is consistent with noisy activations from dormant neurons inflating the rank estimate without a corresponding gain in policy quality.

Fig. S8 shows policy entropy. All three methods exhibit monotonically decreasing entropy as their policies converge. PRIME's entropy curve closely tracks MAPPO's, with PRIME occasionally exceeding MAPPO in the mid-training window; the three curves remain within a narrow band of each other.



(a) Dormant fraction (Normal) (b) Feature rank (Normal)

Fig. S7: Plasticity diagnostics in Normal mode. (a) MAPPO's DNF rises to $\sim$40–45%; PRIME stabilizes at $\sim$10–18%. (b) MAPPO's higher rank ($\sim$9–10) is inflated by noisy dormant-neuron activations.

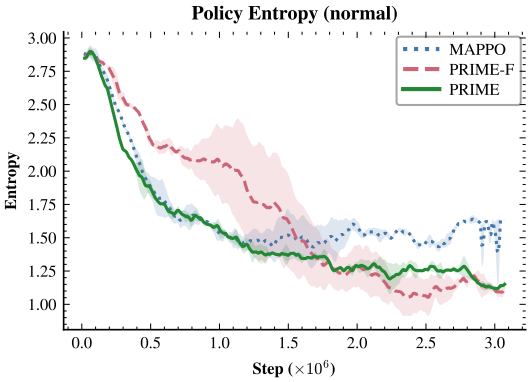


Fig. S8: Policy entropy in Normal mode. All methods show decreasing entropy; the three curves remain within a narrow band of each other, with PRIME and MAPPO largely overlapping.

Appendix G
Per-Layer Dormancy Breakdown

The main text reports the total dormant neuron fraction aggregated across all layers. This appendix provides the per-layer breakdown in Change mode (Fig. S9), which reveals that dormancy is distributed unevenly across the network.

Policy Layer 0, which directly processes the raw observation vector, exhibits the highest dormancy for MAPPO ($\sim$25–55%) and the largest absolute reduction under PRIME ($\sim$20–35%). Early layers are most exposed to input distribution shifts at phase boundaries, which explains this pattern. Policy Layer 1 shows lower absolute dormancy across all methods, with MAPPO remaining the highest.

The value network layers behave differently. MAPPO's Value Layer 1 accumulates $\sim$50–65% dormant neurons, while PRIME keeps Value Layer 0 below 10% and holds Value Layer 1 near 20%. This asymmetry arises because the critic receives stronger, more consistent gradient signals from TD-error backpropagation, allowing PRIME's resets to integrate rapidly. PRIME-Forward, by contrast, shows erratic dormancy spikes in the value layers (e.g., Value Layer 0 jumps to $\sim$55% around Step 5×10^6). These spikes reflect the destabilizing effect of indiscriminate forward-only resets on the critic, a failure mode that PRIME avoids through the gradient-silence filter.

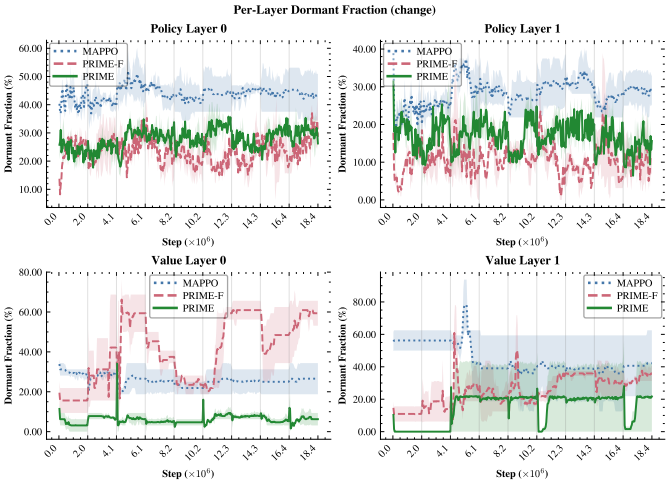


Fig. S9: Per-layer dormant fraction in Change mode. Policy Layer 0 shows the highest dormancy for MAPPO ($\sim$25–55%). PRIME keeps Value Layer 0 dormancy below 10% and Value Layer 1 near 20%, reflecting the strong TD-error gradient flow. PRIME-Forward exhibits erratic dormancy spikes in value layers from indiscriminate resets.

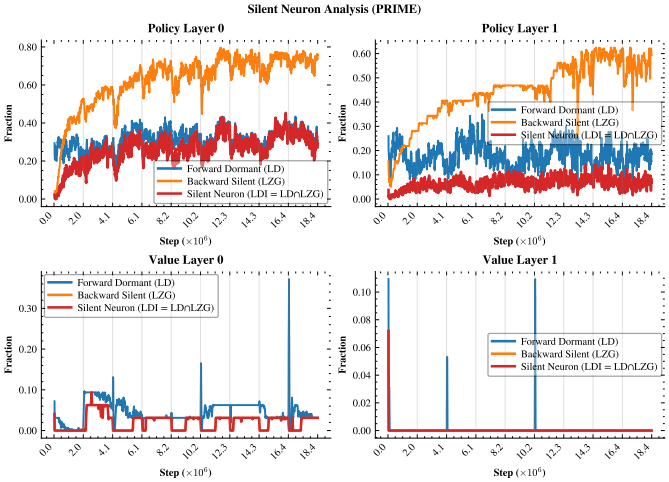


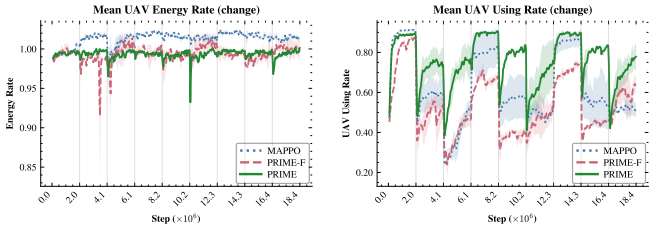
Fig. S10: Silent neuron decomposition for PRIME (Change mode). Each panel shows one network layer: forward-dormant fraction (L_D, blue), backward gradient-silent fraction (L_{ZG}, orange), and their intersection ($L_{DI} = L_D \cap L_{ZG}$, red); the legend uses the plain-text labels LD, LZG, and LDI. Policy Layer 0 exhibits 20–30% silent neurons and Policy Layer 1 a smaller fraction; value layers show minimal dormancy.

Fig. S10, referenced from Section V of the main paper, complements this view with PRIME's per-layer decomposition into the forward-dormant, gradient-silent, and intersection sets.

Appendix H
Energy and UAV Utilization in Change Mode

Fig. S11 presents two supplementary Change-mode metrics that are referenced but not plotted in the main text.

Fig. S11(a) shows the mean UAV energy usage rate, which stays within 0.96–1.01 for all three methods apart from brief transient dips at phase boundaries. MAPPO dips furthest,



(a) Energy rate (Change) (b) UAV using rate (Change)

Fig. S11: Energy and utilization in Change mode. (a) Energy rates are comparable (~ 0.96–1.01 apart from brief boundary dips); performance differences are not attributable to energy trade-offs. (b) UAV utilization rate tracks coverage: PRIME recovers fastest to ~ 0.90.

briefly reaching ~ 0.86 around Step 10.2×10^6, but the differences are small relative to the inter-method reward and coverage gaps. PRIME's performance advantages are therefore not attributable to differential energy expenditure.

Fig. S11(b) shows the UAV utilization rate, defined as the fraction of time slots during which a UAV is actively serving at least one user. This metric qualitatively tracks the coverage rate reported in Fig. 7(a) of the main paper: PRIME achieves the fastest post-switch recovery to ~ 0.90, while MAPPO's recovery speed degrades across later phases. The correlation between utilization and coverage confirms that PRIME's coverage advantage stems from more effective trajectory repositioning rather than from differences in power allocation or channel conditions.

Appendix I
Per-Layer Gradient Analysis

The main text employs the backward-silence criterion $\hat{g}_{l,i} \leq \tau_g$ for reset decisions (Section V of the main paper) but does not visualize the per-layer zero-gradient fraction directly. Fig. S12 provides the full per-layer LZG (the layer-wise zero-gradient fraction, denoted L_{ZG} in Fig. S10) visualization in Change mode.

In the policy layers (top row), MAPPO's LZG fluctuates between 0.6 and 1.0, reflecting transient gradient bursts that do not translate into sustained learning. PRIME-Forward's LZG increases over training as reset neurons begin with near-zero weights that produce small gradients. PRIME's LZG stabilizes at ~ 0.5–0.8 in Policy Layer 0 and ~ 0.3–0.55 in Policy Layer 1, indicating that a meaningful fraction of policy neurons maintain active gradient flow while the silent subset is periodically refreshed.

In the value layers (bottom row), the difference is pronounced. PRIME maintains near-zero LZG (< 0.05), confirming that virtually all value-network neurons receive strong gradient signals from TD-error backpropagation. MAPPO's value LZG approaches 1.0 in later phases, consistent with the high value-network dormancy documented in Appendix G. This pattern explains the asymmetric reset counts in Fig. S13: the policy network requires far more resets than the critic because the critic's gradient flow naturally resists dormancy, with Policy Layer 0 receiving ~ 5–20 resets per period.

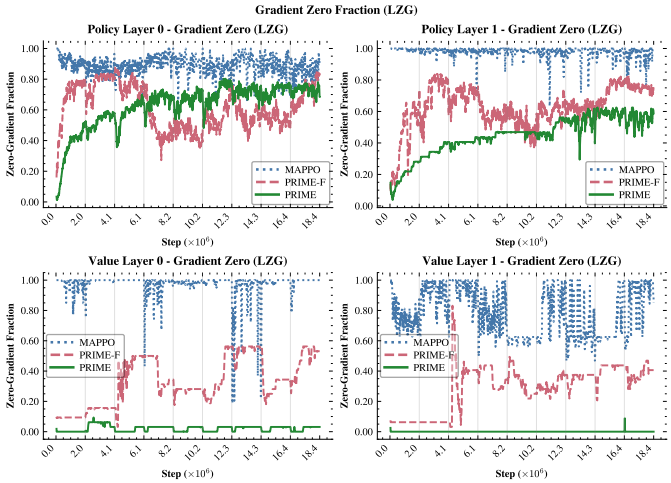


Fig. S12: Per-layer zero-gradient fraction (LZG) in Change mode. Policy layers: MAPPO's LZG fluctuates (0.6–1.0); PRIME stabilizes at ~ 0.5–0.8 (Layer 0) and ~ 0.3–0.55 (Layer 1). Value layers: PRIME maintains near-zero LZG ($<$ 0.05), confirming strong gradient flow; MAPPO approaches 1.0, consistent with high value-network dormancy.

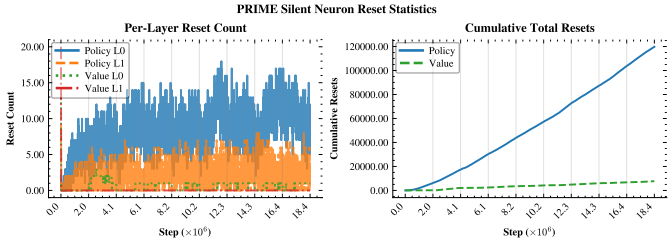


Fig. S13: PRIME reset statistics (Change mode). Left: per-layer reset counts (Policy Layer 0 receives the most resets). Right: cumulative totals show that the policy backbone accumulates $\sim 1.2\times10^5$ resets versus $\sim 5\times10^3$ for the critic.

Appendix J
Multi-Metric Reset Period Ablation

The main text reports reset period sensitivity using IQM return as the sole metric (Fig. 10(a) of the main paper(a)). This appendix extends the ablation to coverage rate and dormant neuron fraction to verify that PRIME's robustness to F holds across the evaluation suite.

As shown in Fig. S14, PRIME's IQM return and coverage are stable across $F \in \{50, 200, 300\}$, varying by less than 2% on return. The dormant fraction is more sensitive: it stays near 17–22% for $F \in \{50, 200\}$ but rises to $\sim 38\%$ at F=300, where the inter-reset interval is long enough for dormant neurons to accumulate before the next detection sweep. This elevated dormancy at F=300 nevertheless produces no measurable performance loss, confirming that the bidirectional silent set is selective enough that even temporarily inflated dormancy between resets remains harmless. PRIME-Forward, in contrast, is sensitive on all three metrics: its return and coverage improve with larger F (less frequent over-resetting causes less disruption), while its dormant fraction increases monotonically with F (more frequent resets mechanically

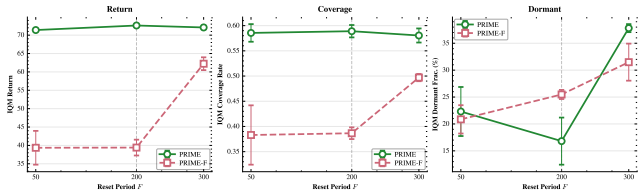


Fig. S14: Multi-metric reset period ablation (Change mode, τ_d=0.5, τ_g=0.08). Left: IQM return; Center: coverage rate; Right: dormant fraction. PRIME is stable across all F values and metrics; PRIME-Forward varies with reset frequency.

reduce dormancy but destroy useful representations in the process).

These results reinforce the conclusion from Section V of the main paper: the bidirectional intersection criterion, not the specific choice of F, is the primary factor behind PRIME's task-level robustness.

REFERENCES

[1] Z. He and Z. Liu, "Silent neuron theory and plasticity preservation for deep reinforcement learning in adaptive video streaming," *IEEE Trans. Mobile Comput.*, 2025, early access.

[2] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017. [Online]. Available: https://arxiv.org/abs/1707.06347

[3] R.-A. Lascu, D. Šiška, and Łukasz Szpruch, "Ppo in the fisher-rao geometry," 2026. [Online]. Available: https://arxiv.org/abs/2506.03757

[4] W. Qiu, Z. He, W. Zhao, and H. Masui, "Plasticity-enhanced multi-agent mixture of experts for dynamic objective adaptation in uavs-assisted emergency communication networks," 2026. [Online]. Available: https://arxiv.org/abs/2604.09028