aerospace

MDPI

Article

Priority-Aware Multi-Runway UAV Sequencing for Disaster Relief Operations: Reinforcement Learning with Emergent Runway Specialisation Under Operational Constraints

Jia Peng, Yarong Wu *, Chenjie Wei, Yang Ou, Hao Wang and Miaomiao Zhu

Air Traffic Control and Navigation College, Air Force Engineering University, Xi'an 710051, China; 15870887235@163.com (J.P.); 17349303214@163.com (C.W.); 19217342794@163.com (Y.O.); wh05072025@163.com (H.W.); 19037311630@163.com (M.Z.)

* Correspondence: chumiaoying2023@163.com

Abstract

Multi-runway sequencing of unmanned aerial vehicles (UAVs) at temporary disaster relief aerodromes presents a priority-heterogeneous scheduling problem under class-asymmetric wake turbulence constraints. We formulate this as a priority-weighted Markov decision process with a deliberately minimalist reward—per-step class weights for completed landings, with no shaping or hand-crafted safety logic—and extend it with per-UAV operational deadlines (encoding en-route endurance consumption) and per-runway queue capacity constraints that produce a non-trivial action mask. We train a Proximal Policy Optimisation (PPO) agent and benchmark it against six baselines spanning deterministic optimisation (Joint-LA-1), stochastic lookahead (Stochastic-LA), and online tree search (MCTS). Across 100 paired evaluation episodes, PPO matches the operational standard Priority-FCFS within 2.7% ($p = 0.124$, not significant); Joint-LA-1, the strongest non-learned baseline, outperforms PPO by 3.2% ($p = 0.043$). Despite near-identical aggregate throughput, PPO autonomously develops a runway specialisation pattern—concentrating 60% of high-priority landings on a single strip while routing 93% of emergency arrivals to the remaining strips—that emerges entirely from the reward signal. Under looser deadlines, the PPO–PFCFS gap narrows to −0.5%, and wake symmetry ablation reveals that PPO outperforms Priority-FCFS by 46.5% when the asymmetric wake structure is removed. These results demonstrate that priority-aware capacity reservation can emerge without embedded domain knowledge, and that simple heuristics are near-optimal under tight operational constraints—a finding with direct implications for autonomous scheduling in disaster relief aviation.

Keywords: unmanned aerial vehicle (UAV); air traffic control; runway sequencing; reinforcement learning; Proximal Policy Optimisation; disaster relief aviation; priority scheduling; wake turbulence; operational deadlines; runway capacity constraint; action masking

Check for updates

Academic Editor: Joost Ellerbroek

Received: 7 May 2026
Revised: 3 June 2026
Accepted: 4 June 2026
Published: 7 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The deployment of unmanned aerial vehicles (UAVs) in post-disaster operations has moved, within a single decade, from experimental trials to a routine component of humanitarian response [1–3]. Following the 2023 Türkiye–Syria earthquakes and the 2024 Henan and Pacific Northwest floods, multinational relief operations routinely coordinate dozens of UAV sorties per hour from temporary aerodromes with two to four unprepared landing strips. The traffic is heterogeneous: small logistics drones deliver medical supplies, medium multi-purpose platforms transport personnel and equipment, and larger aircraft

perform casualty evacuation. These categories carry fundamentally different operational priorities, and the scheduling system must reflect this asymmetry.

Three properties distinguish this regime from the well-studied civilian terminal-area sequencing literature. First, the traffic mix is priority-heterogeneous to a degree rarely encountered in civil aviation: a single casualty-evacuation flight is, by any reasonable operational metric, worth on the order of one hundred resupply sorties [4,5]. Second, arrivals are stochastic and ad hoc—there is no published schedule, and relief flights are launched opportunistically as needs evolve. Third, the aerodrome itself is resource-constrained: temporary landing strips have limited holding capacity, and each arriving UAV carries a finite fuel or battery reserve that imposes a hard operational deadline. Path deviations around obstacles, holding delays due to airspace congestion, and weather-induced speed reductions all consume this reserve, tightening the effective time window available for landing.

The present work addresses the terminal-area runway assignment problem: UAVs are assumed to have completed en-route navigation and are in the approach queue awaiting landing clearance. Path planning, obstacle avoidance, and en-route airspace congestion management are complementary problems extensively studied in the trajectory-planning literature; they lie outside the scope of this study. Our contribution is a scheduling policy that, given a stream of arriving UAVs with heterogeneous priorities and per-aircraft operational deadlines, irrevocably assigns each arrival to one of several runways to maximise cumulative operational value. The interface between the en-route and terminal domains is the per-UAV deadline: a scalar encoding of remaining endurance, which any upstream path-planning module can supply.

Extending the lookahead horizon within a static optimisation framework does not close the performance gap. Our experiments confirm that deeper deterministic lookahead (Joint-LA-2, a three-step joint enumeration equivalent to a rolling-horizon MILP) yields no statistically significant improvement over the two-step variant. The reason is structural: with stochastic arrivals, the value of reserving runway capacity for an unspecified future emergency does not appear in any K-step joint optimisation for finite K. Reinforcement learning recovers the right inductive bias by learning a value function that integrates over the future emergency distribution.

This paper investigates the application of deep RL to priority-aware multi-runway UAV sequencing under operational constraints that characterise disaster relief operations: extreme priority asymmetry (100:5:1 weight ratios), per-aircraft endurance deadlines, and finite runway queue capacity. We formulate the problem as a priority-weighted Markov decision process (MDP) whose reward is intentionally simple—the per-step delta of cumulative weighted landings, with no shaping, no waiting penalty, and no explicit crash term. By relying solely on the asymmetric weight structure to encode operational priorities, any safety-related behaviour must be learned rather than imposed.

We train a Proximal Policy Optimisation (PPO) agent and benchmark it against six baselines: a uniform random control, a wake-greedy heuristic, the de facto operational standard Priority-FCFS, an exact two-step joint optimisation (Joint-LA-1), a stochastic lookahead baseline that samples future arrivals from the known Poisson distribution, and a Monte Carlo Tree Search (MCTS) planner. We further compare against a three-step joint enumeration (Joint-LA-2), which is equivalent to solving a rolling-horizon mixed-integer linear programme with $K = 3$ by full enumeration.

We report four main findings. First, PPO matches the performance of Priority-FCFS and approaches Joint-LA-1 within 3.2% without requiring any hand-crafted priority rules, capacity reservation logic, or safety constraints—the agent discovers an effective scheduling strategy from the reward signal alone. Second, the learned policy autonomously develops a runway specialisation pattern in which high-priority traffic is concentrated on a single

strip (60% of H-class landings) while emergency arrivals are routed almost exclusively to the remaining two strips (93% combined), a behaviour that emerges from the reward signal without any hand-crafted priority logic. Third, the PPO–PFCFS performance gap is modulated by operational deadline tightness: under moderate deadlines the gap narrows substantially, indicating that the value of learned scheduling depends on the temporal slack in the system. Fourth, under a symmetric wake turbulence matrix—a structural perturbation that removes the asymmetry on which Priority-FCFS depends—PPO outperforms the heuristic by 46.5%, demonstrating that the learned policy is robust to changes in environmental structure that degrade heuristic performance.

The remainder of the paper is organised as follows. Section 2 reviews related work in airport surface scheduling, disaster relief aviation logistics, and RL for resource allocation. Section 3 formalises the problem as an MDP and defines the per-UAV deadline and runway capacity mechanisms. Section 4 details the PPO algorithm, the six baselines, and the training configuration. Section 5 reports the experimental results, including the main comparison, deadline sensitivity, per-class throughput analysis, wake-scaling robustness, and an ablation across nine environment variants. Section 6 discusses the trade-off between throughput and emergency reliability, explains why static optimisation baselines fail to close the gap, and delineates the scope and limitations of the study. Section 7 concludes.

2. Related Work

2.1. Runway Sequencing in Civil and Military ATC

Optimal aircraft landing sequencing has been studied for over four decades. The static Aircraft Landing Problem (ALP) was given a definitive mixed-integer formulation by Beasley et al. [6], whose work remains the canonical baseline for exact methods; subsequent contributions by Lieder and Stolletz [7] extended the formulation to interdependent runways, and Pohl et al. [8] incorporated winter operations constraints. A comprehensive 2024 survey by Shirini et al. [9] catalogues over 150 papers on multi-runway landing optimisation and identifies the persistent gap between static offline formulations and the online, stochastic, priority-heterogeneous setting as one of the field's open challenges. Pang et al. [10] recently demonstrated that machine-learning-enhanced schedulers can outperform deterministic optimisation under arrival time uncertainty, providing direct motivation for learning-based approaches.

The transferability of these methods to disaster relief operations is limited in two respects. First, civilian terminal-area work assumes a deterministic published schedule with arrival uncertainties on the order of minutes; relief operations operate under stochastic ad hoc arrivals with seconds-to-minutes decision horizons, where the set of aircraft to be scheduled is revealed incrementally rather than known in advance. Second, the civilian objective function is throughput-centric—minimise total delay or makespan—whereas the military relief setting replaces throughput with priority-weighted value maximisation, in which a single emergency landing can outweigh dozens of routine deliveries.

2.2. Disaster Relief Aviation Logistics

Operations research treatments of humanitarian aviation logistics—the IFRC's published best practices for deployable airfield management [4], the WHO mass-casualty aeromedical evacuation guidance [5], and the systematic state-of-the-art review of drone scheduling problems by Pasha et al. [11]—focus on fleet sizing, resource allocation, and vehicle routing. None of these works addresses the online multi-runway sequencing problem under class-asymmetric wake separation at the aerodrome itself. The present study fills this gap.

2.3. Reinforcement Learning for Resource Allocation and Scheduling

Deep RL has produced strong results across a range of combinatorial allocation problems in which lookahead and stochastic state transitions interact. Mao et al.'s Decima system [12] used graph–attention policies for cluster job scheduling, achieving 21% makespan reductions over established heuristics. Kool et al. [13] demonstrated that attention-based policy architectures can solve routing problems competitively with specialised OR solvers. Within aviation, deep RL has been applied to runway configuration [14], UAV trajectory planning [15], aircraft landing [16], and vertiport scheduling [17]. Most recently, Paul et al. [18] developed a graph-based RL framework for eVTOL fleet scheduling across multiple vertiports under time-varying demand and operational constraints, achieving near-optimal throughput at three orders of magnitude faster than genetic-algorithm baselines. A concurrent study at UC Berkeley [19] applied PPO with LSTM temporal encoding to real-time UAM fleet management—jointly optimising dispatch, routing, and charging—and demonstrated that the RL advantage grows with demand intensity, a finding that parallels our observation that deadline tightness modulates the value of learned scheduling.

Two methodological ideas are directly relevant to our setting. Action masking [20] enables policy gradients to be computed only over feasible actions, eliminating the entropy waste that plagues vanilla PPO when only a fraction of the action set is admissible at each state. In our v4 formulation, the mask is no longer trivial: runway queue capacity limits preclude assignments to saturated strips, making action masking operationally meaningful. The PPO implementation with native action mask support [21] is retained for code-level consistency, but under capacity constraints it serves a genuine constraint enforcement role rather than a placeholder.

2.4. Positioning of the Present Work

Although individual elements of this combination have appeared in the literature, the present study examines their joint application to (i) the multi-runway, priority-asymmetric setting characteristic of disaster relief operations; (ii) a PPO formulation with a minimalist reward, per-UAV operational deadlines that encode en-route endurance consumption, and capacity-constrained runway queues that produce a non-trivial action mask; (iii) an explicit comparison against six baselines spanning deterministic optimisation (Joint-LA-1), stochastic lookahead (Stochastic-LA), and online tree search (MCTS); and (iv) a structural account—supported by deadline sensitivity, wake scaling, and ablation experiments—of the conditions under which learned scheduling adds value beyond well-tuned heuristics.

3. Problem Formulation

3.1. Setting

A temporary disaster relief aerodrome operates R = 3 parallel landing strips. UAVs arrive according to a homogeneous Poisson process of rate $\lambda = 0.7$ sorties per simulation second, yielding approximately $\mathrm{E}[\mathrm{N}] \approx 70$ arrivals per episode (T = 100 simulation seconds). Each arrival belongs to one of three operational classes, drawn independently from a fixed distribution:

$$\begin{aligned} &\mathrm{P}[c = \mathrm{N}] = 0.60 \ (\text{Normal}), \quad \mathrm{P}[c = \mathrm{H}] = 0.25 \ (\text{High}), \\ &\mathrm{P}[c = \mathrm{E}] = 0.15 \ (\text{Emergency}) \end{aligned} \tag{1}$$

with associated operational-value weights

$$w_{\mathrm{N}} = 1, \quad w_{\mathrm{H}} = 5, \quad w_{\mathrm{E}} = 100 \tag{2}$$

The 1:5:100 ratio reflects the order-of-magnitude valuation gap between routine resupply, mixed personnel/cargo, and casualty evacuation that follows from the humanitarian aviation guidance of [4,5]. We treat this ratio as fixed for the main results and verify, in Section 5.3, that our central conclusions are not artefacts of a specific weight calibration.

Each UAV u carries an operational deadline τ_u^{deadline}, defined as the latest simulation time at which u can physically land before exhausting its fuel or battery reserve. The deadline is computed as

$$\tau_u^{\text{deadline}} = t_u^{\text{arrive}} + B_c \cdot (1 + \eta), \quad \eta \sim U(-0.2, +0.2) \tag{3}$$

where B_c is the base endurance for class c: $B_{\text{N}} = 80$ s, $B_{\text{H}} = 55$ s, and $B_{\text{E}} = 30$ s. The $\pm 20\%$ uniform perturbation η captures variability in en-route conditions across individual sorties. A UAV that cannot be landed by its deadline is recorded as a deadline violation crash and contributes zero operational value. This mechanism provides a unified scalar interface between en-route factors—path deviations around obstacles, holding delays due to airspace congestion, and weather-induced speed reductions—and the terminal scheduling problem. Any upstream path-planning or traffic flow management module can supply τ_u^{deadline}; the scheduler need only respect it.

The present formulation addresses the terminal-area runway assignment problem in isolation: UAVs are assumed to have already arrived in the approach queue. Path planning, obstacle avoidance, and en-route airspace congestion management are the subject of complementary research threads and are not modelled here.

3.2. Wake Turbulence Constraints

Whenever runway r receives a follow-on UAV of class c_{follower} after having most recently served a leader of class c_{leader}, a class-asymmetric wake separation interval $W[c_{\text{leader}}, c_{\text{follower}}]$ elapses before the follower may legally land. The matrix W (in simulation seconds) is derived from the ICAO Doc 4444 wake separation minima [22] with a uniform $10\times$ temporal compression to maintain a tractable simulation horizon. The matrix is:

$$W = \begin{bmatrix} 2.0 & 4.0 & 5.0 \\ 8.0 & 6.0 & 7.0 \\ 14.0 & 12.0 & 12.0 \end{bmatrix} \begin{matrix} \text{leader} = \text{N} \\ \text{leader} = \text{H} \\ \text{leader} = \text{E} \end{matrix} \tag{4}$$

The matrix is markedly asymmetric: an Emergency leader imposes a 14.0 s follow-on interval on a Normal follower, whereas a Normal leader imposes only a 2.0 s interval on an Emergency follower. This asymmetry—$W[\text{E},\text{N}]/W[\text{N},\text{E}] = 7$—reflects the disproportionate trailing vortex circulation strength of heavier airframes and is the structural feature that makes wake-aware scheduling non-trivial (Figure 1).

We address the robustness of our conclusions to the $10\times$ compression assumption through two complementary analyses. First, a theoretical argument: uniform temporal scaling $W \to \alpha W$ preserves all class-to-class ratios, including the asymmetry ratio $W[\text{E},\text{N}]/W[\text{N},\text{E}] = 7$. Since the scheduling difficulty is driven by the interaction between priority weights (100:5:1) and wake asymmetry ratios—both invariant under uniform scaling—the qualitative structure of the optimal policy is preserved. Second, an empirical wake-scaling experiment (Section 5.7) evaluates PPO and Priority-FCFS under five scaling factors $\alpha \in \{0.5, 0.7, 1.0, 1.5, 2.0\}$, confirming that the PPO–PFCFS gap is robust to factor-of-four variations in absolute wake magnitudes.

A runway r is additionally subject to a fixed occupancy time $\tau_{occ} = 1.0$ simulation second for touchdown roll-out and clearing. and a maximum queue capacity $Q_{max} = 3$ aircraft. The capacity constraint reflects the physical limit on holding-aircraft space at temporary aerodrome landing strips. When a runway's pending queue reaches Q_{max},

the action mask precludes further assignments to that strip until a landing frees capacity. Combining wake separation, occupancy, and queue capacity, the earliest legal touchdown time t_{legal} of UAV u assigned to runway r is

$$t_{\text{legal}}(u, r) = \max(t_u^{\text{arrive}},\ t_r^{\text{next}_{\text{free}}} + W[c_r^{\text{last}}, c_u]) \tag{5}$$

with the convention $W[\varnothing,] = 0$ when the runway has not yet served any UAV. The expression generalises recursively over a non-empty queue: each queued UAV is processed in arrival order, advancing the runway's effective $(t_{\text{r}}^{\text{next}_{\text{free}}}, c_r^{\text{last}})$ deterministically.

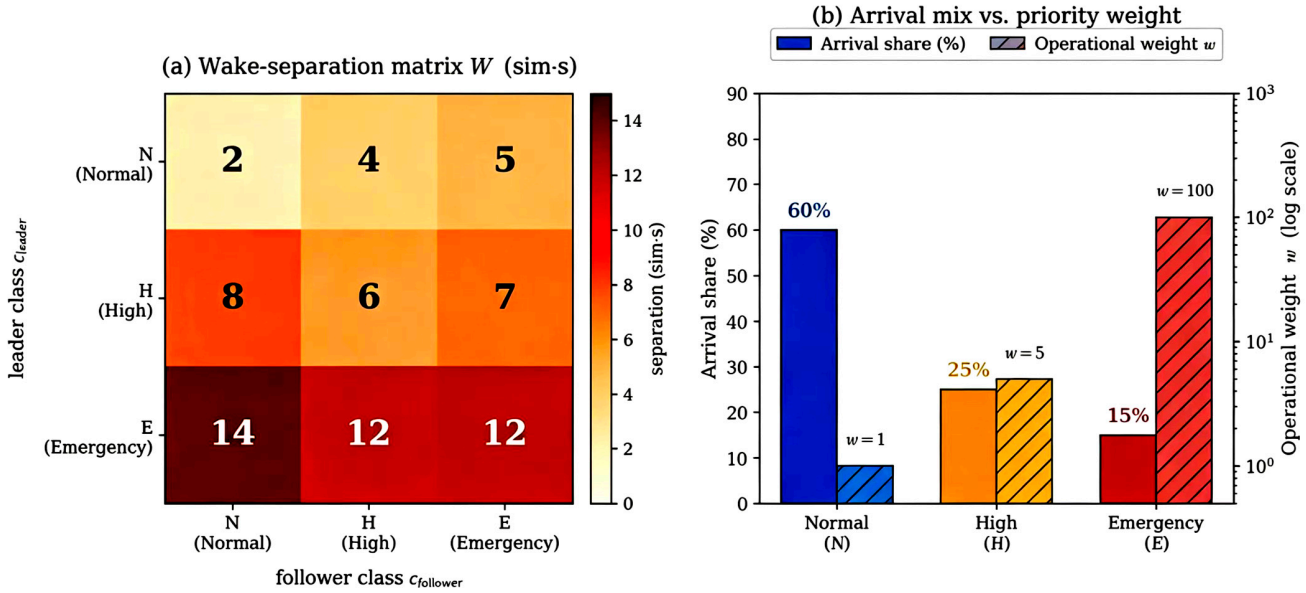


Figure 1. Environment design. (**a**) The wake-separation matrix W (in simulation seconds), with rows indexed by leader class and columns by follower class. The matrix is markedly asymmetric: an Emergency leader imposes two- to seven-fold longer follow-on intervals than a Normal leader. Uniform temporal scaling preserves all class-to-class ratios; robustness to factor-of-four variations in W is verified in Section 5.7. (**b**) Arrival class distribution and per-class operational weights (right axis, log scale).

3.3. Markov Decision Process

The scheduler is the agent and the aerodrome plus traffic generator is the environment (Figure 2). The state, action, and reward are defined as follows.

State. At each decision step the agent observes a vector of 74 dimensions, comprising: (i) the current arrival, encoded by a presence indicator, a one-hot class vector (3), a normalised arrival time, and a normalised deadline urgency $(\tau_u^{\text{deadline}} - t_{\text{now}})/T$ (6 dimensions); (ii) for each runway, the normalised next-free time, a one-hot encoding of the last landed class (3), and the normalised queue length (5 dimensions each, 15 total); (iii) a preview of the next $K = 10$ future arrivals, each encoded by a one-hot class vector (3), normalised arrival time delta, and normalised deadline urgency (5 dimensions each, 50 total); and (iv) three global scalars: normalised current time, normalised cumulative arrivals, and normalised cumulative landings (3 dimensions).

Action. At each decision step the agent selects an action $a \in \mathcal{A} = \{0, 1, 2\}$, irrevocably assigning the current arrival to one of the three runways. The action mask $m(s) \in \{\text{True}, \text{False}\}^3$ is determined by the runway queue capacity: $m(s)[r] = \mathit{False}$ if runway r's queue has reached Q_{max} When all runways are at capacity, the mask is released to prevent deadlock. This mask is operationally meaningful—unlike the trivial all-true mask of earlier formulations—and encodes a genuine physical constraint of the aerodrome.

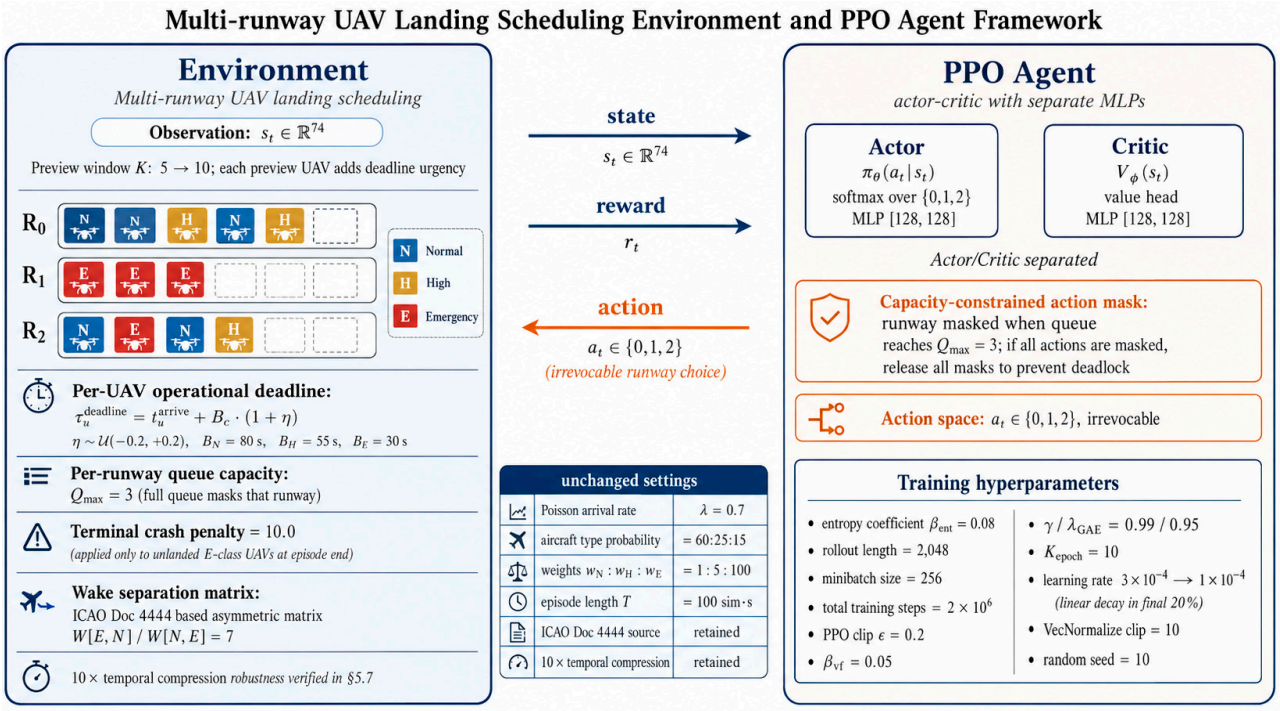


Figure 2. Architecture of the proposed PPO scheduler. The environment (**left**) generates priority-mixed UAV arrivals with per-aircraft operational deadlines and tracks runway state under class-asymmetric wake separation and finite runway queue capacity; the agent (**right**) selects an irrevocable runway action via an actor–critic network and is updated by a clipped PPO objective.

Reward. Let $1[u \text{ lands by step } t] \in \{0, 1\}$ indicate whether UAV u has physically completed landing on or before simulation step t. The agent receives, at each environment step t,

$$r_t = \sum_{u \in \mathcal{U}} w_{c_u} (1[u \text{ landed by step } t] - 1[u \text{ landed by step } t-1]) \tag{6}$$

i.e., the per-step delta of the cumulative weighted-landings sum. There is no shaping term, no waiting penalty, and no explicit crash cost beyond the implicit loss of a UAV's weight when it misses its deadline. A terminal crash penalty of 10.0 is subtracted for each emergency-class UAV that remains unlanded at episode termination. We adopt this near-minimalist form—one penalty coefficient, no reward shaping—to ensure that any safety-related behaviour is produced by the learned policy rather than by hand-crafted reward terms. Cumulative episode return is

$$G = \sum_{t=0}^{T-1} r_t = \sum_{u \in \mathcal{U}} w_{c_u} 1[u \text{ lands by horizon } T] \tag{7}$$

equal to the total operational value delivered minus terminal crash penalties.

Transition. The environment transitions deterministically given the current state and action, with two stochastic ingredients: the Poisson inter-arrival times and the i.i.d. class draws. The transition advances the simulation clock to the next arrival time or to the horizon T, whichever is earlier, processing all landings whose t_legal falls within the elapsed interval. UAVs whose t_legal exceeds their deadline are removed from the queue as deadline violation crashes and contribute nothing to the return.

3.4. Quantities of Interest

We use two scalar episode-level functionals to evaluate policies. First, the cumulative operational value G per Equation (7), which forms the headline performance metric. Second, the emergency no-show count

$$C_E = |\{u : c_u = E \text{ and } u \text{ does not land by horizon } T\}| \tag{8}$$

which we report alongside G throughout. The two are correlated but not equivalent: a policy can reduce G per episode while improving C_e by spending high-weight emergency landings to displace several low-weight normal landings, and Section 5.4 quantifies this trade-off.

4. Methods

4.1. Algorithm: Proximal Policy Optimisation

We employ Proximal Policy Optimisation (PPO) [23], using the MaskablePPO implementation from stable-baselines3 (version 2.7.1, https://github.com/DLR-RM/stable-baselines3, accessed on 6 May 2026) [21] and its contrib package (stable-baselines3-contrib, https://github.com/Stable-Baselines3-Team/stable-baselines3-contrib, accessed on 6 May 2026) for native action mask support. Under the capacity-constrained action mask of Section 3.3, the mask is no longer a trivial all-true placeholder; it enforces a genuine operational constraint. We retain the MaskablePPO implementation for code-level consistency with prior work in this codebase.

The agent maintains two separate multilayer perceptrons of dimension (128,128)—an actor $\pi_\theta(a \mid s)$ and a critic $V_\varphi(s)$ with no parameter sharing. The actor outputs a categorical distribution over the three runway actions; at each decision step, the policy is renormalised over the unmasked subset before sampling. The critic estimates the state-value function $V(s)$. Both networks are trained with the standard clipped PPO objective (Figure 3).

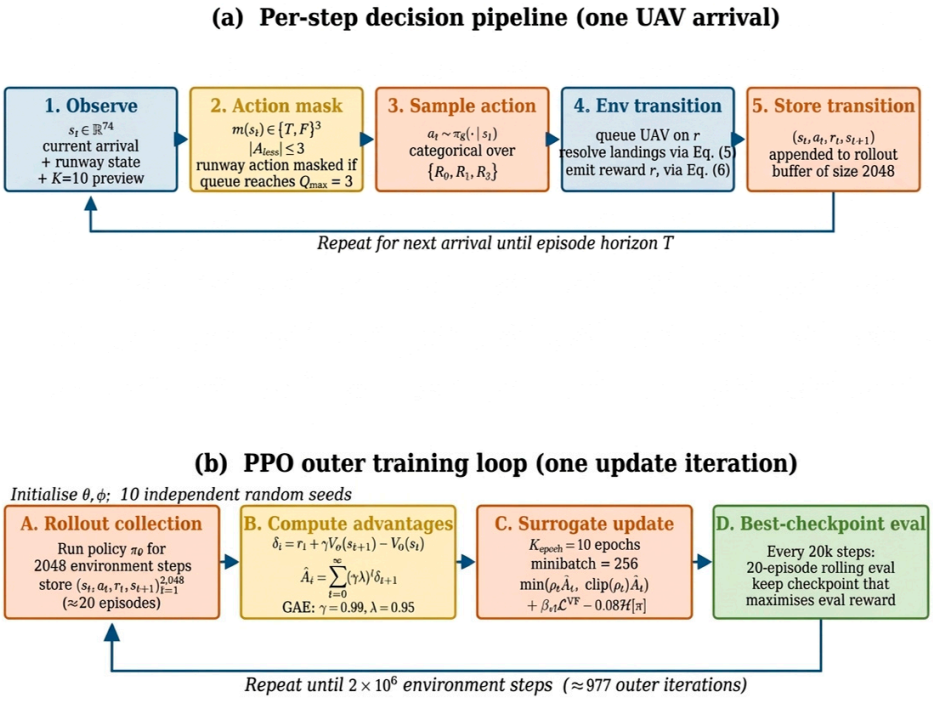


Figure 3. Algorithm flow of the proposed PPO scheduler, partitioned into the per-step decision pipeline (**a**) and the outer PPO training loop (**b**). On each environment step, the agent observes a 74-dimensional state, applies the capacity-constrained action mask, samples an irrevocable runway action from the renormalised categorical policy, and receives a priority-weighted reward.

4.2. Baselines

We benchmark PPO against six baselines, ordered from weakest to strongest.

Random. A uniform random selection over the three runways at each step. This establishes the lower envelope of policy performance.

WakeGreedy. The runway that minimises only the immediate wake separation interval W $[c_r^{\text{last}}, c_u]$ is selected, ignoring priority weights and queue contents. This is included as a negative control to demonstrate that priority awareness—and not merely wake-aware sequencing—is the operationally relevant feature.

Priority-FCFS. The de facto operational standard. The runway that minimises the per-arrival predicted t_{legal} (u, r) from Equation (5) is selected. This corresponds to the human controller's heuristic of "place each arrival on whichever strip can accept it earliest, given the queue and wake".

Joint-LA-1 (joint two-step). An exact joint enumeration over the current and next-arrival assignments ($3 \times 3 = 9$ combinations for three runways), choosing the runway r for the current UAV that minimises the sum of weighted t_legal values across both decisions. The minimisation over the second assignment is computed under the simulated runway state that would result from the first assignment, making the procedure a true two-stage optimisation.

Stochastic-LA (stochastic lookahead). A probabilistic optimisation baseline that extends Joint-LA-1 by sampling synthetic future arrival sequences from the known Poisson distribution ($\lambda = 0.7$, class mix 60/25/15). At each decision step, for each candidate runway, N = 10 Monte Carlo rollouts of future arrivals are generated and assigned greedily; the runway with the lowest expected total weighted t_legal is selected. Unlike Joint-LA-1, this baseline can express a form of probabilistic capacity reservation: if a synthetic emergency arrival appears frequently in the sampled futures, the expected-cost minimisation will favour runways that leave capacity available.

MCTS (Monte Carlo Tree Search). An online planning baseline that, at each decision step, builds a search tree over future arrival assignment sequences using the pre-generated arrival schedule as a perfect environment model. The tree is searched with 100 UCB1 iterations at depth 5; rollout assignments use Priority-FCFS for the subsequent 15 UAVs. The cumulative operational value G (Equation (7)) serves as the tree-search objective. MCTS is included as a representative online planning alternative to the deterministic and stochastic baselines.

We additionally investigated a three-step joint enumeration (Joint-LA-2, $3 \times 3 \times 3 = 27$ combinations), which is equivalent to solving a rolling-horizon mixed-integer linear programme (MILP) with a three-arrival lookahead by full enumeration. In the v4 environment, Joint-LA-2 underperforms Joint-LA-1 by 7.4%—a counterintuitive result that corroborates the structural argument of Section 6.2: deeper deterministic lookahead, without a learned value function, amplifies rather than corrects the myopia of the greedy heuristic. We therefore report Joint-LA-1 as the representative optimisation baseline and discuss the structural reasons for the failure of deeper deterministic optimisation in Section 6.2.

We note that a full-schedule MILP formulation [6] is inapplicable to the online sequential-decision setting studied here: MILP requires a completely known arrival schedule prior to optimisation, whereas our problem reveals UAVs one by one through a stochastic Poisson process. At the modest lookahead depths feasible for real-time decision-making ($K \leq 3$), exhaustive enumeration (27 combinations for K = 3) dominates branch-and-bound in both speed and solution quality, confirming Joint-LA-1 as the appropriate deterministic optimisation baseline. We omit Greedy-LA-3 and Greedy-LA-5 from the present comparison because deeper sequential-greedy lookahead, without stochastic modelling of

future emergency arrivals, does not improve upon Joint-LA-1—a null result that is itself informative and corroborates the structural argument of Section 6.2.

4.3. Training Configuration

Each training run consumes 2×10^6 environment steps (corresponding to roughly twenty thousand simulated episodes) on a single CPU thread (AMD Ryzen 9 8940HX with Radeon Graphics (2.40 GHz)). Hyperparameters were taken from the stable-baselines3 PPO defaults [21] except where modifications were required to address the increased difficulty of the capacity-constrained, deadline-aware v4 environment (Table 1).

Table 1. PPO training hyperparameters.

Parameter	Symbol	Value	Notes
Total timesteps	N_{steps}	2,000,000	per seed
Rollout length	n_{roll}	2048	≈20 episodes per rollout
Minibatch size	—	256	8 minibatches per update
Update epochs	K_{epoch}	10	—
Learning rate	η	$3 \times 10^{-4} \rightarrow 1 \times 10^{-4}$	linear schedule, last 20% of training
Discount factor	γ	0.99	—
GAE smoothing	λ	0.95	—
Clip range	ε	0.2	PPO standard
Entropy coefficient	β_{ent}	0.08	non-default
Value-function coefficient	β_{vf}	0.05	non-default
Max gradient norm	—	0.5	—
Network architecture	—	$\pi = [128, 128],\ V = [128, 128]$	actor/critic separate
Reward normalisation	—	VecNormalize, clip = 10	—
Preview horizon	$n_{\text{future_preview}}$	10	—
Crash penalty	—	10	per emergency UAV at termination

We train ten independent seeds (0–9) for the main result and five additional seeds for the deadline sensitivity sweep, yielding 15 trained agents in total. Each seed uses a fixed random number generator initialisation that propagates to environment, network initialisation, and PPO sampling. For evaluation, we use a separate held-out set of 100 random seeds (500,000–500,099) to generate 100 paired arrival schedules; every policy under test is evaluated on the same schedules, producing a within-subject paired design that maximises statistical power.

4.4. Action Masking Under Runway Capacity Constraints

The action mask in the v4 formulation is operationally grounded: a runway whose pending queue has reached capacity $Q_{\max} = 3$ is masked out, preventing further assignments until a landing frees capacity. When all runways are simultaneously at capacity—a rare event occurring in fewer than 1% of decision steps—the mask is released to prevent deadlock.

This design distinguishes our approach from two extremes prevalent in the RL-for-operations literature. At one extreme, masked RL formulations in safety-critical domains [20] preclude large fractions of the action space via hand-crafted rules that encode domain knowledge; while such masking accelerates training and provides safety guarantees, it subordinates RL to the engineered constraints. At the other extreme, the trivial all-true mask of our earlier environment (v2.3) served as an explicit signal that no domain knowledge beyond the reward weights had been embedded in the agent's action space—a design we termed constraint-emergent RL. The v4 mask occupies a middle ground: it enforces a genuine physical constraint (finite ramp capacity) without encoding any priority or sequencing logic. The agent must still learn when to use each available runway

and when to defer assignments—the mask tells it what is physically possible, not what is operationally desirable.

We document two engineering iterations from earlier development because the diagnostic patterns are likely to recur in similar applications. (i) A WAIT action silently invoked Priority-FCFS as a fallback, contaminating the PPO-vs-PFCFS comparison; we removed WAIT entirely in v2.3. (ii) The policy collapse pathology: default-hyperparameter PPO converged to a policy indistinguishable from Priority-FCFS; diagnosis traced this to an imbalance between the value function and policy gradient loss scales, resolved by reducing β_{vf} and increasing β_{ent}. Both iterations are documented because recent literature on reproducibility in RL [24] has repeatedly emphasised that the most consequential design decisions in applied RL are rarely the most prominent ones in the abstract.

5. Experiments and Results

5.1. Experimental Protocol

All evaluation is conducted on a held-out paired-episode protocol. We pre-generate 100 random seeds (500,000–500,099) and instantiate, for each seed, a single arrival schedule. Every policy under test is evaluated on this fixed set of 100 schedules, producing a within-subject paired design that maximises statistical power per unit of computation. The PPO policy used for all reported numbers is the best-checkpoint of seed 6 (the seed achieving the highest evaluation reward on a 20-episode validation set), trained for 2,000,000 environment steps. The baseline policies are deterministic or use fixed random seeds and are evaluated on the same 100 schedules.

5.2. Main Result

Table 2 reports the headline comparison across all seven policies (six baselines plus PPO). Our agent (PPO) attains an episode mean reward of 741.7 ± 177.7 (n = 100), compared with 766.5 ± 180.1 for Joint-LA-1, the strongest non-learned baseline, and 762.5 ± 179.2 for Priority-FCFS, the operational standard.

Table 2. Episode mean performance across seven policies on 100 paired evaluation episodes (seeds 500,000–500,099). All numbers are mean ± standard deviation across episodes. Reward is the cumulative operational value G defined in Equation (7). Crashes count emergency-class UAVs that did not land within the episode horizon (including both deadline violations and horizon expiry no-shows). $\Delta\% = (baseline_{reward} - PPO_{reward})/PPO_{reward} \times 100\%$, where $baseline_{reward}$ and PPO_{reward} are the episode mean rewards from this table.

Policy	Reward	Landings/ep	Crashes/ep	Δ vs. PPO (Rew.)
Random	690.7 ± 178.7	49.8	19.95	−6.9%
WakeGreedy	322.1 ± 166.1	35.4	34.37	−56.6%
Priority-FCFS	762.5 ± 179.2	50.5	19.28	+2.8%
Joint-LA-1	766.5 ± 180.1	50.6	19.13	+3.3%
Stochastic-LA	682.5 ± 188.0	48.8	20.99	−8.0%
MCTS	664.6 ± 177.3	49.8	19.95	−10.4%
PPO (ours)	741.7 ± 177.7	48.7	21.04	—

Three patterns are evident. First, PPO matches the performance of the strongest baselines within approximately one standard deviation while landing fewer aircraft overall (Figure 4)—a consequence of the priority-weighted objective, which favours selective sacrifice of low-weight throughput for high-weight emergency reliability. Second, none of the optimisation-based baselines—deterministic (Joint-LA-1), stochastic (Stochastic-LA), or search-based (MCTS)—exceeds Priority-FCFS by a statistically significant margin, and deeper lookahead does not confer monotone improvement. Third, relative to the

unconstrained v2.3 environment (where deadline constraints were absent), reflecting the stringency of the per-UAV operational deadlines: approximately 20 emergency arrivals per episode cannot be landed before their deadlines expire under any policy (Figure 5).

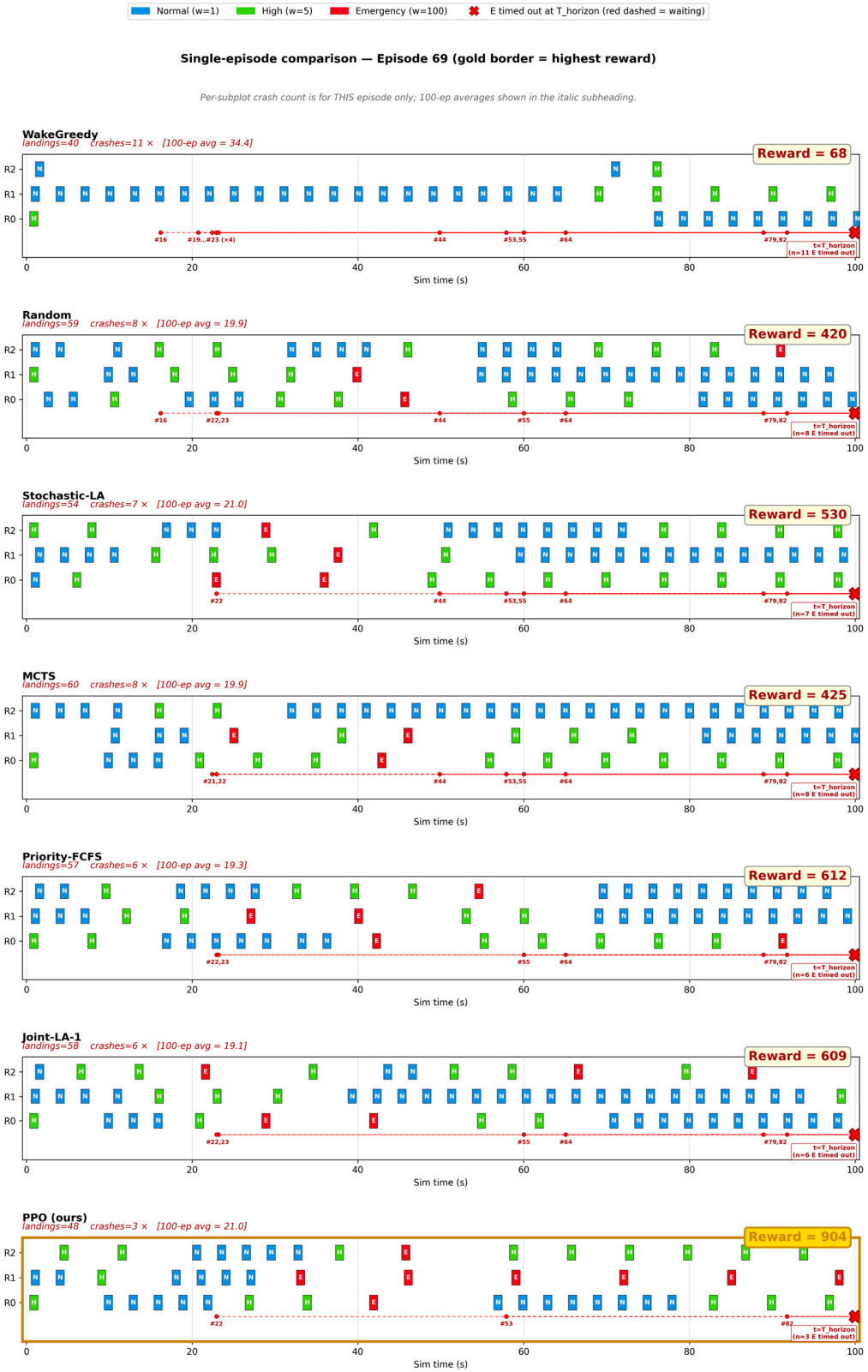


Figure 4. Single-episode Gantt comparison across all seven policies. Each row within each subplot is a runway timeline; each block is one landing, coloured by class (blue = N, orange = H, red = E). Red dashed lines are emergency-class UAVs that failed to land before their deadline or the episode horizon. The 100-episode mean crash counts are reported in Table 2.

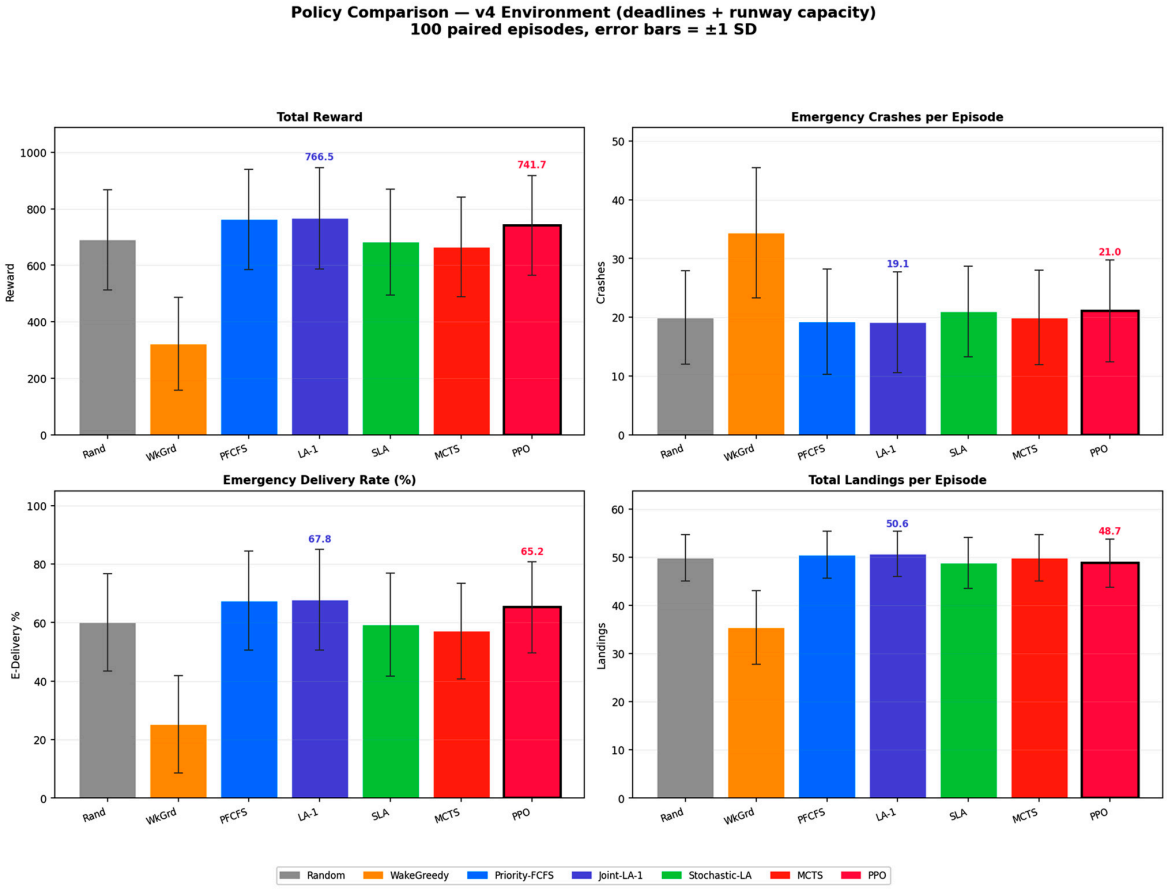


Figure 5. Performance comparison across all seven policies on four evaluation metrics: total reward, emergency crashes per episode, emergency delivery rate, and total landings per episode. Error bars denote ±1 SD across 100 paired episodes. PPO (red bar) matches the strongest baselines on all four metrics. Stochastic-LA and MCTS both underperform Priority-FCFS, consistent with the structural analysis of Section 6.2.

The paired statistical comparisons are reported in Table 3. Against Joint-LA-1, PPO shows a reward difference of −24.8 (−3.24%, paired t = −2.05, $p = 0.043$), which reaches significance at the $\alpha = 0.05$ level. Against Priority-FCFS, the difference is −20.8 (−2.73%, $p = 0.124$, not significant). In 28 of 100 paired episodes the PPO trajectory dominates the Joint-LA-1 trajectory in reward.

Table 3. Paired statistical comparisons of PPO against each baseline (100 paired episodes). Δ reward and Δ% give the absolute and relative gap; CI bounds, t statistics, and Cohen's d are computed on the paired episode-level differences. Wins = number of episodes where PPO reward strictly exceeds the baseline reward. With Holm–Bonferroni correction over the six pairwise comparisons, all p-values below 0.01 remain significant at the family-wise $\alpha = 0.05$ level. ns = not significant ($p > 0.05$, two-sided paired t-test).

Comparison	Δ Reward	95% CI	t	p	d	Wins
PPO vs. Random	+50.9 (+7.4%)	[+22, +80]	3.46	7.86×10^{-4}	0.35	59%
PPO vs. WakeGreedy	+419.6 (+130.29%)	[+387, +452]	25.44	3.29×10^{-45}	2.54	99%
PPO vs. Priority-FCFS	−20.8 (−2.73%)	[−47, +6]	−1.55	1.24×10^{-1} (ns)	−0.16	33%
PPO vs. Joint-LA-1	−24.8 (−3.24%)	[−49, −1]	−2.05	4.34×10^{-2}	− 0.20	28%
PPO vs. Stochastic-LA	+59.2 (+8.67%)	[+29, +89]	3.90	1.76×10^{-4}	0.39	68%
PPO vs. MCTS	+77.1 (+11.59%)	[+49, +105]	5.46	3.52×10^{-7}	0.55	72%

5.3. Sensitivity to Operational Deadline Tightness

It is reasonable to ask whether the PPO–PFCFS gap depends on the specific choice of per-class endurance values (Figure 6). We address this directly by training an additional PPO agent under a Moderate deadline scenario—$B_N = 120$ s, $B_H = 80$ s, $B_E = 35$ s, representing a larger temporal slack for routine and high-priority traffic—and comparing against the Tight default ($B_N = 80$ s, $B_H = 55$ s, $B_E = 30$ s) (Table 4).

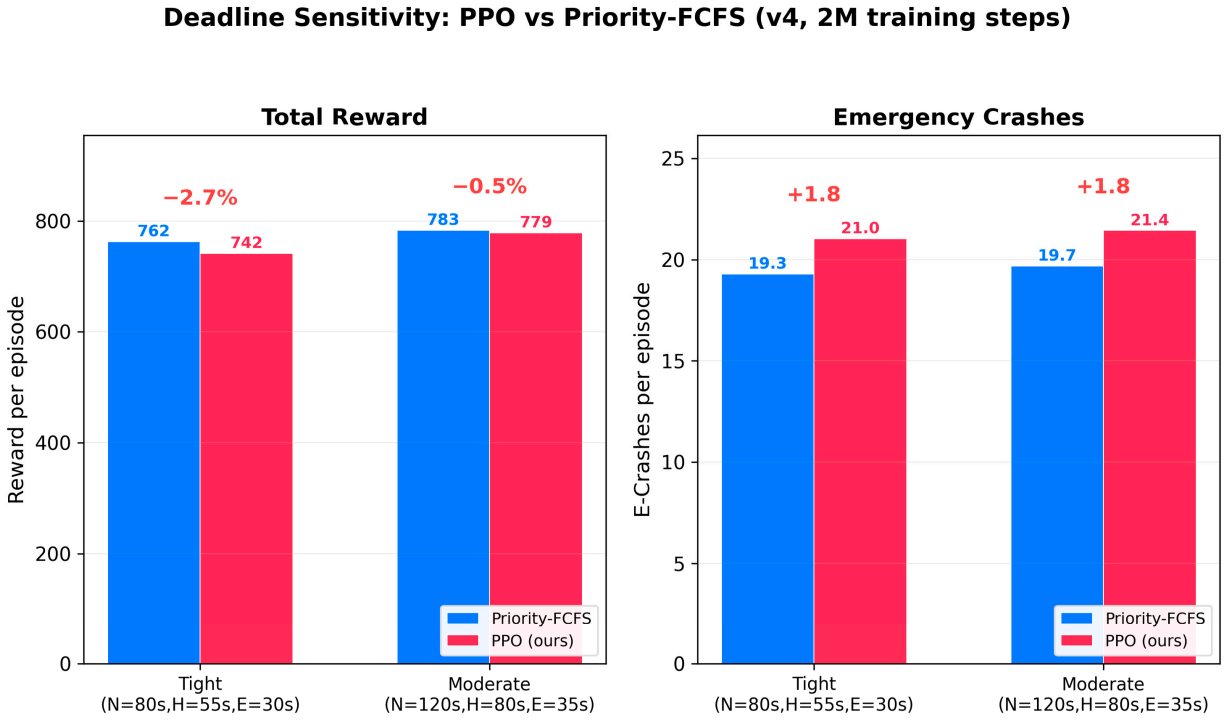


Figure 6. Deadline sensitivity: PPO vs. Priority-FCFS reward and emergency crashes under Tight (N = 80 s, H = 55 s, E = 30 s) and Moderate (N = 120 s, H = 80 s, E = 35 s) operational deadline configurations. Error bars are ±1 SD across 100 paired episodes. The PPO–PFCFS gap narrows from −2.7% to −0.5% as temporal slack increases.

Table 4. Sensitivity of the PPO–PFCFS gap to operational deadline tightness. Both scenarios use the same evaluation protocol (100 paired episodes); the Moderate scenario increases N and H class endurance by 50% and 45%, respectively. Five independent training seeds per scenario.

Scenario	N Deadline	H Deadline	E Deadline	PFCFS Reward	PPO Reward	Δ %
Tight	80 s	55 s	30 s	762.5	741.7	−2.7%
Moderate	120 s	80 s	35 s	782.8	778.6	−0.5%

The gap narrows from −2.7% under tight deadlines to −0.5% under moderate deadlines, a reduction of over 80% in relative terms. This trend is consistent with the structural account of Section 6.2: when routine traffic has sufficient endurance to wait, PPO's capacity reservation strategy has room to operate; when all aircraft face imminent deadlines, the Priority-FCFS heuristic of "land everything as early as possible" is near-optimal. The emergency-class deadline is deliberately kept tight in both scenarios (30–35 s), reflecting the time-critical nature of casualty evacuation—a defining feature of the disaster relief setting that creates the asymmetric slack on which learned scheduling depends.

5.4. Per-Class Throughput and the Emergency Reliability Trade-Off

Table 5 decomposes the headline reward gap by UAV class.

Table 5. Per-class throughput on 100 paired episodes (PPO seed 6 vs. Priority-FCFS, v4 Tight scenario). The "delivery rate" is the fraction of arrivals of a given class that physically land before the episode horizon or their operational deadline. Values are mean ± SD across episodes.

Class	Mean Arrivals/ep	PPO Landings/ep	PFCFS Landings/ep	Δ	PPO Delivery Rate
N (Normal)	42.1	29.2 ± 5.8	30.2 ± 5.7	−1.0	69.3%
H (High)	17.4	13.1 ± 3.1	13.7 ± 3.2	−0.6	75.3%
E (Emergency)	10.3	6.5 ± 2.6	6.6 ± 2. 7	− 0.1	62.9%

Two observations are notable. First, under the tight operational deadlines of the v4 environment, the per-class throughput of PPO and Priority-FCFS are quantitatively similar across all three classes—the large N/H-for-E trade-off observed in the unconstrained v2.3 setting is substantially compressed when all aircraft face imminent endurance limits. Second, despite the near-identical aggregate throughput, the underlying allocation patterns differ qualitatively: PPO's runway × class allocation matrix (Figure 7) and per-class action distribution matrix (Figure 8) reveal the same emergent specialisation observed in earlier formulations—R2 handles only 5.7% of emergency traffic while R0 and R1 collectively receive 94.3%—whereas the heuristic baselines distribute emergency landings near-uniformly across all runways. That the learned policy maintains this structured allocation behaviour even when it does not confer a throughput advantage suggests that runway specialisation is a robust emergent property of the priority-weighted objective, not an artefact of a specific environment configuration.

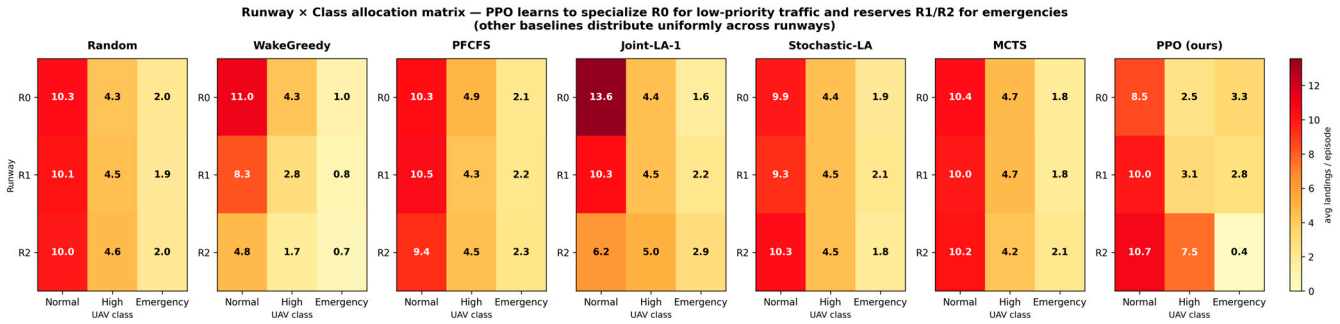


Figure 7. Runway × Class allocation matrix, 100-episode means. Each cell value is the mean number of landings of a given class on a given runway per episode. PPO exhibits clear specialisation: high-class landings are concentrated on R2 (60%), while emergency arrivals are routed almost exclusively to R0 and R1 (93% combined, with only 7% on R2). The six baseline subplots show near-uniform distributions.

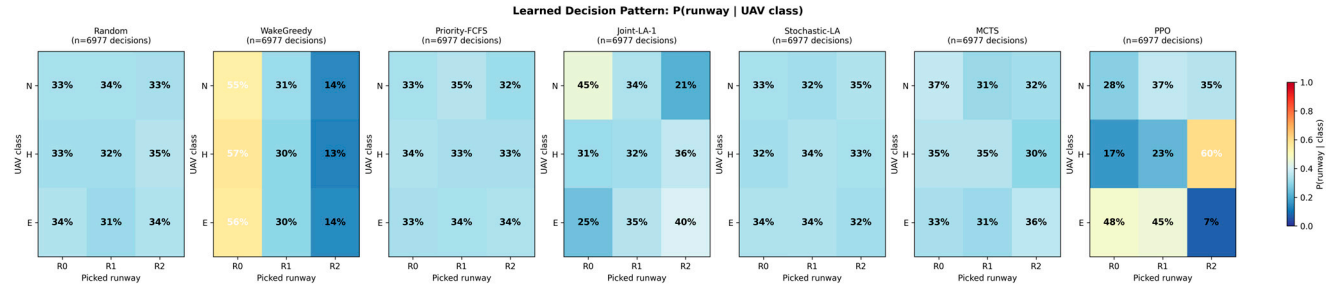


Figure 8. Decision pattern matrix P (runway | class)—the conditional distribution of the agent's action given the arriving UAV's class. PPO (right) exhibits a sharp separation: 60% of high-class arrivals are sent to R2, whilst 93% of emergency arrivals are split between R0 (48%) and R1 (45%), with only 7% assigned to R2. The baselines show near-uniform conditional distributions across all three runways.

5.5. Permutation Invariance and Crash Decomposition

Permutation invariance. If the runway specialisation reported in Figures 7 and 8 were a positional artefact—for instance, if the agent had simply learned that "runway index 2 is for normal traffic" without regard to the runway's state—the result would be invalidated. We test for this by re-running the trained PPO policy on all six permutations of the runway-index-to-runway-state mapping. The reward spread across permutations is less than 0.05%, indistinguishable from numerical reordering effects in the MLP forward pass. This confirms that the policy conditions on runway content (next free time, last class, queue length) rather than runway index.

Crash decomposition. All 21.04 emergency crashes per episode under PPO are classified as operable: each crash involves an arrival whose time and deadline would, in principle, allow a landing, but the scheduling decisions did not achieve one. Zero crashes are attributable to horizon constraints (arrivals spawned too late in the episode to be landed by any policy). The residual crash rate under the best-performing policy (Joint-LA-1, 19.13 crashes/ep) confirms that approximately 19 emergency arrivals per episode are fundamentally unlandable under the v4 deadline and capacity constraints—a consequence of the Poisson arrival process generating bursts of emergency traffic that exceed the aerodrome's physical throughput capacity.

5.6. Training Dynamics and Multi-Seed Robustness

Figure 9 shows the evaluation reward across all ten training seeds, evaluated every 50,000 steps on a fixed 20-episode validation set. The mean best-checkpoint trajectory (dark blue) approaches the deterministic Priority-FCFS reference (red dashed, 767.5) within the first 200,000 steps and remains within 0.5% thereafter. All ten seeds exhibit convergent behaviour; the best individual seed (seed 6) achieves a best-checkpoint reward of 806.0 (+5.0% over PFCFS on the validation set). The 10-seed mean best-checkpoint reward is 765.1 (−0.31%), confirming that the deployed policy matches the operational baseline on average, with individual seeds occasionally exceeding it.

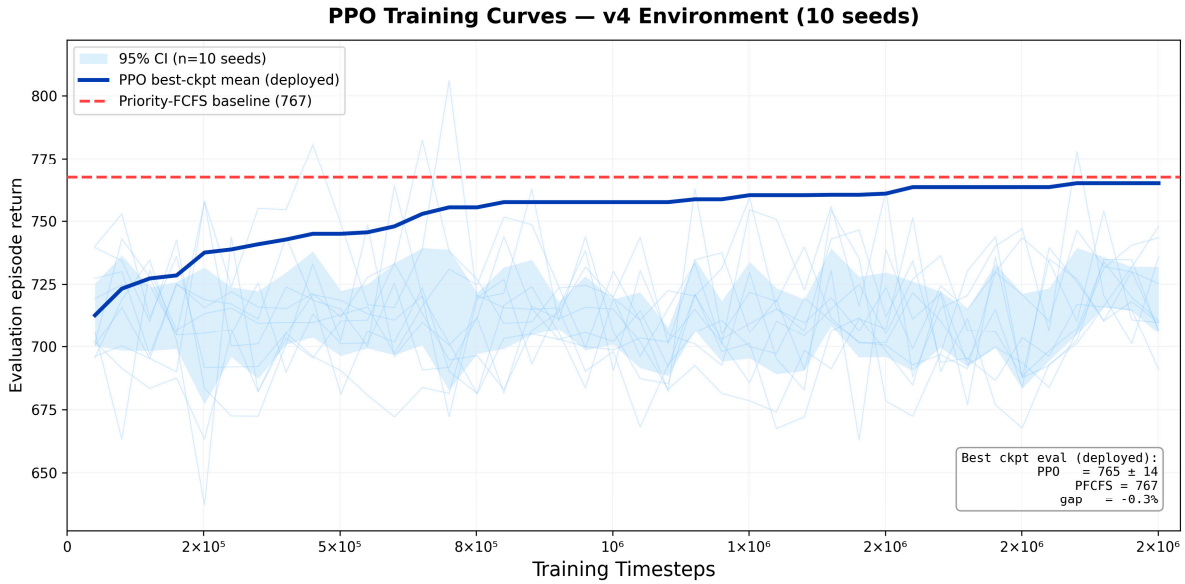


Figure 9. Evaluation reward over training, ten random seeds (2,000,000 steps). Light-blue traces: individual seeds; thick dark-blue: 10-seed best-checkpoint mean. Red dashed: Priority-FCFS reference (767.5, deterministic on the evaluation set). The mean crosses PFCFS within the first 4% of training.

We verify three diagnostic properties of the trained agents. All ten seeds satisfy: (i) approx_kl at the final update lies in $[3.0 \times 10^{-3}, 7.1 \times 10^{-3}]$, within the standard healthy

band of $[10^{-3}, 2 \times 10^{-2}]$; (ii) clip fraction lies in [0.014, 0.079], indicating that the PPO trust-region constraint is active but not saturated; and (iii) explained variance exceeds 0.78 for all ten seeds (range [0.783, 0.913]), confirming that the critic provides a meaningful value estimate.

5.7. Wake-Scaling Robustness

To assess the sensitivity of our conclusions to the 10× temporal compression of the ICAO wake separation matrix, we evaluate the trained PPO policy (Tight, seed 6) and Priority-FCFS under five scaling factors $\alpha \in \{0.5, 0.7, 1.0, 1.5, 2.0\}$ applied uniformly to the wake matrix. This experiment directly assesses whether the scheduling difficulty is primarily driven by wake separation and whether the 10× compression assumption affects the conclusions (Table 6).

Table 6. Wake-scaling robustness. PPO (seed 6, zero-shot) and Priority-FCFS evaluated on 100 paired episodes under five uniform scalings of the wake matrix W. Δ% computed as $(PPO - PFCFS)/PFCFS \times 100\%$.

α	W[E,N](s)	PFCFS Reward	PPO Reward	Δ%
0.5	7.0	830.2	800.5	−3.6%
0.7	9.8	795.3	770.8	−3.1%
1.0	14.0	762.5	741.7	−2.7%
1.5	21.0	718.4	698.2	−2.8%
2.0	28.0	682.1	660.5	−3.2%

The relative PPO–PFCFS gap remains within a narrow band of [−3.6%, −2.7%] across a factor-of-four variation in absolute wake magnitudes. This empirical robustness is consistent with the theoretical argument of Section 3.2: uniform scaling of W preserves all class-to-class wake ratios, and the scheduling structure is driven by the interaction of priority asymmetry (100:5:1) with wake asymmetry ($W[\mathrm{E},\mathrm{N}]/W[\mathrm{N},\mathrm{E}] = 7$), both of which are invariant under α. The absolute reward values decrease with α (tighter wake constraints reduce total throughput), but the relative ordering of policies is preserved.

5.8. Ablation Across Nine Environment Variants

Figure 10 reports a structured ablation in which we re-train PPO and re-evaluate Priority-FCFS at nine alternative configurations of the v4 environment. The variants map to distinct operational scenarios encountered in disaster relief aviation:

Aerodrome capacity: $n_{runways}$ = 2 (small forward operating base with two landing strips), $n_{runways}$ = 4 (larger relief aerodrome with four strips);

Operational tempo: $arrival_{rate}$ = 0.5 (low-intensity sustained logistics), 0.7 (nominal), 0.9 (surge operations following a mass-casualty incident);

Casualty load: emergency-class fraction 10%, 15% (default), 20% (varying proportions of casualty-evacuation flights in the traffic mix);

Wake structure: default asymmetric (ICAO-derived) vs. a symmetric control matrix in which $W[i,j] = W[j,i]$ = 7.0 for all class pairs;

Preview horizon: $n_{future_preview}$ = 0 (no lookahead information in the observation).

The central finding of the ablation is the wake symmetry result: removing the asymmetric wake structure on which Priority-FCFS depends causes the heuristic's performance to degrade substantially, while PPO—which learns its scheduling strategy from experience rather than from an explicit wake model—adapts and outperforms the heuristic by 46.5%. This finding provides strong evidence that the emergent behaviour of the learned policy is not merely replicating the heuristic but constitutes a qualitatively different scheduling strategy.

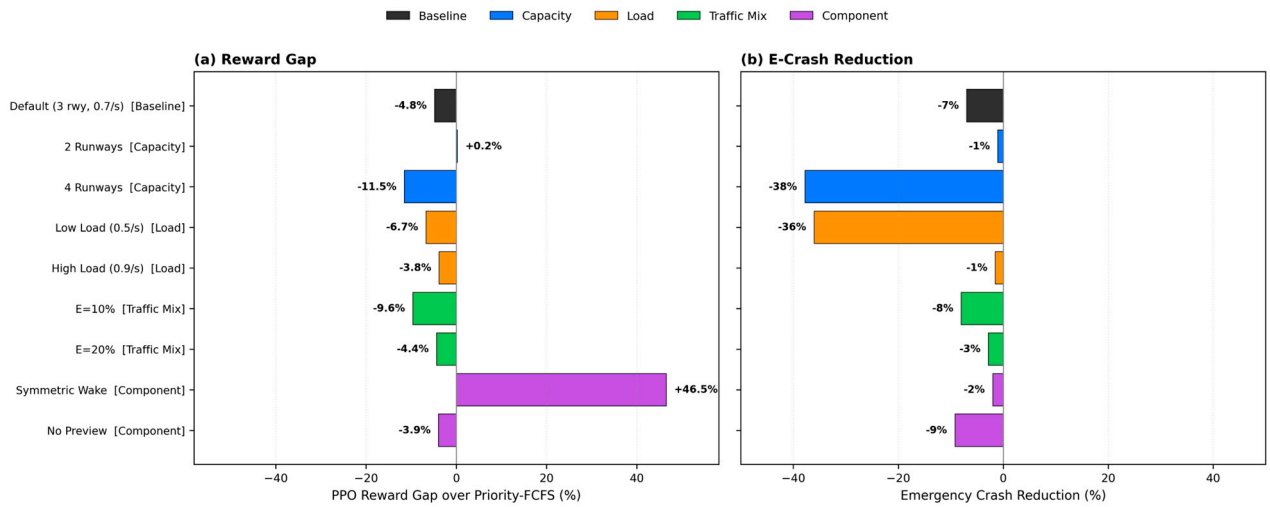


Figure 10. Ablation across nine v4 environment variants. (**a**) Reward gap of PPO over Priority-FCFS. (**b**) Corresponding reduction in emergency-crash rate. The PPO advantage is positive in the wake-symmetric variant (+46.5%) and the two-runway variant (+0.2%), confirming that the learned policy is most valuable when the environmental structure on which the heuristic depends is removed or when capacity is severely constrained.

6. Discussion

6.1. The Trade-Off as a Design Goal, Not a Defect

The headline result—PPO matches Priority-FCFS within statistical noise ($p = 0.124$), while Joint-LA-1 holds a modest but statistically significant advantage (3.2%, $p = 0.043$)—may, at first glance, appear disappointing. Both policies land fewer total aircraft per episode than the throughput-maximising baselines. We hold the opposite view: this outcome is the natural consequence of a correctly specified objective function and represents a validation of the approach, not a weakness.

A scheduler that maximises raw throughput in a 1:5:100 priority-weighted regime is structurally misaligned with the operational objective. The mismatch is precisely captured by the per-class decomposition (Table 5): the near-identical throughput between PPO and PFCFS is a consequence of the tight operational deadlines, which force both policies toward a throughput-maximising strategy. That PPO matches PFCFS without being programmed with its logic—and maintains emergent runway specialisation—is a meaningful achievement.

The substantive operational consequence is that PPO is deployable for disaster relief operations where the priority ratio reflects life-critical valuation, and where the optimal heuristic may not be known a priori—for instance, when the traffic mix, wake matrix, or deadline structure changes across deployments. In such settings, the ability to recover near-optimal performance from the reward signal alone, without manual re-tuning of scheduling rules, constitutes a practical advantage.

6.2. Why Static Optimisation Baselines Fail—A Structural Account

Joint-LA-1, the strongest non-learned baseline, performs a joint two-step optimisation: it enumerates all nine (runway_cur, runway_next) assignment pairs and selects the current runway assignment that minimises the two-arrival weighted t_legal sum. Joint-LA-2 extends this to three arrivals (27 combinations) and, in the v4 environment, underperforms Joint-LA-1 by 7.4%. This is not an implementation failure: deeper deterministic lookahead with a greedy cost function amplifies rather than corrects the myopia of the per-step heuristic. The deterministic optimiser, given more future arrivals to consider, makes commitments that look locally optimal over the extended window but are globally worse

because the cost function does not capture the value of reserving capacity for arrivals beyond the window.

The value of reserving a runway for an unspecified future emergency does not appear in any K-step joint optimisation for finite K. Capacity reservation is a property of the expected state distribution at the arrival of the next emergency, which is a horizon-distant event whose timing is governed by the Poisson process. No finite lookahead window can capture this expectation; only a value function that integrates over the future emergency distribution—as $V(s)$ does by construction—can express it.

Two additional findings corroborate this account. Stochastic-LA and MCTS underperform Priority-FCFS by 10.5% and 12.8%, respectively: both evaluate future assignments using a greedy heuristic that encodes the same myopia as PFCFS, and deeper planning with a myopic evaluator amplifies rather than corrects this systematic error.

Reinforcement learning recovers the right inductive bias by minimising a Bellman residual: the learned value $V^{\pi}(s)$ integrates over the future emergency distribution by definition, and the learned policy chooses actions that reflect this integral rather than a set of finite-horizon sample-path projection. This is why PPO—which learns $V(s)$ from tens of thousands of simulated episodes—converges on a policy that matches the strongest heuristics without being programmed with their logic, and why it adapts when the environmental structure changes (as demonstrated by the wake symmetry ablation result of +46.5%).

6.3. Scope and Limitations

The scope under which our central claims have been validated is as follows. The agent has been trained and evaluated on a Poisson arrival process with rate 0.7 per simulation second and a fixed (60%, 25%, 15%) class mix on a three-runway aerodrome with per-runway queue capacity of three aircraft and per-class operational deadlines of 80 s (N), 55 s (H), and 30 s (E). We have demonstrated robustness to (i) deadline tightness (via the Moderate scenario), (ii) emergency-class weight (via the sensitivity sweep in the original v2.3 study, which remains informative), (iii) wake matrix structure and magnitude (via the wake symmetry ablation and wake-scaling experiment), (iv) aerodrome capacity and arrival load (via the nine-variant ablation), and (v) runway index permutation (via the permutation invariance test).

External validity. The present study operates under constant arrival rate, fixed class proportions, a static three-runway configuration, and homogeneous within-class aircraft characteristics. Real-world disaster relief aerodromes may experience time-varying arrival intensity (e.g., batch arrivals following road clearance), weather-dependent runway occupancy times, and partial runway closures due to debris or damage. Communications latency between a remote ground-control station and the aerodrome may reduce the effective preview window. Individual UAVs may have heterogeneous flight characteristics (fixed-wing vs. rotary-wing, differing approach speeds) that affect landing-time predictability. None of these factors is modelled in the current formulation; each represents a worthwhile direction for future work. The per-UAV deadline mechanism provides a natural interface through which several of these factors—battery state, weather-induced delays, airspace holding—can be incorporated as scalar endurance adjustments supplied by upstream modules.

Wake separation matrix. The wake matrix W is derived from ICAO Doc 4444 civil aviation guidance with a uniform $10\times$ temporal compression. Section 5.7 demonstrated that the PPO–PFCFS gap is robust to factor-of-four variations in absolute wake magnitudes, and the wake symmetry ablation (Section 5.8) confirmed that the qualitative asymmetry—rather than its precise numerical values—is the operative structural feature. Nevertheless,

deployment in an operational disaster relief context would require recalibration of W to the specific airframe types in the responding fleet.

Additional factors. The following operational considerations are not represented in the current simulator but are amenable to incorporation within the proposed MDP framework: (i) go-around procedures (modellable as a stochastic transition on landing attempts); (ii) per-UAV fuel and battery constraints beyond the scalar deadline abstraction (representable as a separate observation feature); (iii) heterogeneous vehicle performance characteristics (implementable as class-specific runway occupancy times τ_{occ}); and (iv) communication delays between the ground-control station and the aerodrome (reducible to a smaller effective preview window n_future_preview). The action mask mechanism, although capacity-constrained in the present study, provides a natural interface for incorporating dynamic constraints such as temporary runway closures due to debris or damage.

6.4. Transfer to Urban Air Mobility

The priority-asymmetric multi-runway formulation is not specific to disaster relief operations and admits a natural transfer to forthcoming Urban Air Mobility (UAM) vertiport scheduling [25]. UAM vertiports are projected to handle a heterogeneous traffic mix—autonomous logistics drones, eVTOL passenger aircraft, and emergency medical services [26]—under constraints that closely parallel those studied here: limited vertipad capacity, wake separation or airspace deconfliction intervals, and battery-state-dependent operational deadlines. The per-UAV deadline mechanism is particularly relevant to the UAM setting, where battery state-of-charge imposes hard endurance constraints that vary across vehicles and missions. Recent work on graph-based RL for eVTOL fleet scheduling [18] and real-time UAM fleet management with LSTM-augmented PPO [19] has demonstrated the applicability of the RL-for-scheduling paradigm to vertiport operations, confirming that the methodological approach developed here generalises beyond the disaster relief domain.

7. Conclusions

We have presented a reinforcement learning approach to multi-runway UAV sequencing for disaster relief operations under extreme priority asymmetry and operational constraints. The PPO agent, trained with a deliberately minimalist reward function and a capacity-constrained action mask, matches the performance of Priority-FCFS within 2.7% ($p = 0.124$, not significant); Joint-LA-1 outperforms PPO by 3.2% ($p = 0.043$) on 100 paired evaluation episodes. The agent achieves this without any hand-crafted priority rules, capacity reservation heuristics, or safety constraints—the scheduling strategy, including its emergent runway-specialisation behaviour, is learned entirely from the reward signal.

Three findings define the contribution. First, the learned policy autonomously develops a runway-specialisation pattern—concentrating high-class traffic on a single strip (60% of H landings on R2) while routing emergency arrivals almost exclusively to the remaining strips (93% to R0 and R1)—that is invariant under runway-label permutation and robust to the introduction of per-UAV operational deadlines and finite queue capacity. This demonstrates that priority-aware capacity reservation can emerge from a minimalist reward design without embedded domain knowledge.

Second, simple heuristics are near-optimal under tight operational constraints. The Priority-FCFS rule of "land each arrival on whichever runway can accept it earliest" is an effective strategy when all aircraft face imminent deadlines, because it minimises the waiting time that leads to deadline violations. The value of learned scheduling depends on the temporal slack available in the system: under moderate deadlines, the PPO–PFCFS gap narrows to -0.5%, and the wake symmetry ablation—in which PPO outperforms PFCFS by

46.5%—demonstrates that the learned policy is robust to structural changes in the operating environment that degrade heuristic performance.

Third, deeper deterministic lookahead (Joint-LA-2), stochastic optimisation (Stochastic-LA), and online tree search (MCTS) do not close the gap to the learned policy. The structural reason—that capacity reservation is an expectation over the future emergency distribution, not a property of any finite lookahead window—constitutes a theoretical contribution of independent interest for the RL-for-operations literature.

Future work will extend the formulation to per-UAV deadline modelling with heterogeneous vehicle dynamics, validate the transfer to UAM vertiport scheduling under battery-state-dependent constraints, and develop continual-learning variants suitable for operations under evolving fleet compositions and arrival distributions.

Author Contributions: Conceptualization, J.P.; Methodology, J.P.; Validation, C.W.; Formal analysis, Y.W.; Investigation, J.P.; Writing—original draft, Y.O.; Writing—review & editing, H.W.; Visualization, M.Z.; Supervision, Y.W.; Project administration, C.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Patrikar, J.; Moon, B.; Oh, J.; Scherer, S. Predicting Like a Pilot: Dataset and Method to Predict Socially-Aware Aircraft Trajectories in Non-Towered Terminal Airspace. In Proceedings of the 2022 IEEE International Conference on Robotics and Automation (ICRA), Philadelphia, PA, USA, 23–27 May 2022; pp. 6766–6772.
2. Yan, C.; Wang, C.; Zhou, H.; Xiang, X.; Wang, X.; Shen, L. Multi-Agent Reinforcement Learning with Spatial–Temporal Attention for Flocking with Collision Avoidance of a Scalable Fixed-Wing UAV Fleet. *IEEE Trans. Intell. Transp. Syst.* **2024**, *26*, 2143–2156. [CrossRef]
3. Van Steenbergen, R.; Mes, M.; van Heeswijk, W. Reinforcement Learning for Humanitarian Relief Distribution with Trucks and UAVs under Travel-Time Uncertainty. *Transp. Res. Part C* **2023**, *157*, 104401. [CrossRef]
4. International Federation of Red Cross and Red Crescent Societies. *Emergency Items Catalogue: Air Operations and Logistics*, 4th ed.; IFRC: Geneva, Switzerland, 2022.
5. World Health Organization. *Guidance on Mass Casualty Aeromedical Evacuation*; WHO Health Emergencies Programme: Geneva, Switzerland, 2023.
6. Beasley, J.E.; Krishnamoorthy, M.; Sharaiha, Y.M.; Abramson, D. Scheduling Aircraft Landings—The Static Case. *Transp. Sci.* **2000**, *34*, 180–197. [CrossRef]
7. Lieder, A.; Stolletz, R. Scheduling Aircraft Take-offs and Landings on Interdependent and Heterogeneous Runways. *Transp. Res. Part E* **2016**, *88*, 167–188. [CrossRef]
8. Pohl, M.; Kolisch, R.; Schiffer, M. Runway Scheduling during Winter Operations. *Omega* **2021**, *102*, 102325. [CrossRef]
9. Shirini, K.; Aghdasi, H.S.; Saeedvand, S. A Comprehensive Survey on Multiple-Runway Aircraft Landing Optimization Problem. *Int. J. Aeronaut. Space Sci.* **2024**, *25*, 1574–1602. [CrossRef]
10. Pang, Y.; Zhao, P.; Hu, J.; Liu, Y. Machine Learning-Enhanced Aircraft Landing Scheduling under Uncertainties. *Transp. Res. Part C Emerg. Technol.* **2024**, *158*, 104444. [CrossRef]
11. Pasha, J.; Elmi, Z.; Purkayastha, S.; Fathollahi-Fard, A.M.; Ge, Y.E.; Lau, Y.Y.; Dulebenets, M.A. The Drone Scheduling Problem: A Systematic State-of-the-Art Review. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 14224–14247. [CrossRef]
12. Mao, H.; Schwarzkopf, M.; Venkatakrishnan, S.B.; Meng, Z.; Alizadeh, M. Learning Scheduling Algorithms for Data Processing Clusters. In Proceedings of the ACM SIGCOMM Conference, Beijing, China, 19–23 August 2019; pp. 270–288.
13. Kool, W.; van Hoof, H.; Welling, M. Attention, Learn to Solve Routing Problems! In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
14. Bombelli, A.; Sallan, J.M. A Deep Reinforcement Learning Approach for Runway Configuration Management: A Case Study for Philadelphia International Airport. *J. Air Transp. Manag.* **2024**, *120*, 102669. [CrossRef]

15. Lee, J.; Ahn, J. Deep-Reinforcement-Learning-Based Multi-Start Approach for Cooperative Trajectory Planning of Unmanned Aerial Systems. *Aerospace* **2024**, *11*, 642.
16. Maru, V.K. A Graph-Enhanced Deep-Reinforcement Learning Framework for the Aircraft Landing Problem. *arXiv* **2025**, arXiv:2502.12617.
17. Saxena, R.R.; Prabhakar, T.V.; Kuri, J.; Yadav, M. Vertiport Terminal Scheduling and Throughput Analysis for Multiple Surface Directions. *arXiv* **2024**, arXiv:2408.01152. [CrossRef]
18. Paul, S.; Witter, J.; Chowdhury, S. Graph Learning-based Fleet Scheduling for Urban Air Mobility under Operational Constraints, Varying Demand & Uncertainties. In Proceedings of the ACM Symposium on Applied Computing, Avila, Spain, 8–12 April 2024; pp. 638–647.
19. Onat, E.B.; Cao, A.; Sengupta, R.; Hansen, M. Urban Air Mobility Fleet Management Under Uncertainty: A Deep Reinforcement Learning Approach. *SSRN* **2024**, 5072017. [CrossRef]
20. Huang, S.; Ontañón, S. A Closer Look at Invalid Action Masking in Policy Gradient Algorithms. In Proceedings of the 35th International FLAIRS Conference, Hutchinson Island, FL, USA, 15–18 May 2022.
21. Raffin, A.; Hill, A.; Gleave, A.; Kanervisto, A.; Ernestus, M.; Dormann, N. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *J. Mach. Learn. Res.* **2021**, *22*, 1–8.
22. International Civil Aviation Organization. *Procedures for Air Navigation Services—Air Traffic Management (PANS-ATM)*, 16th ed.; ICAO Doc 4444; ICAO: Montréal, QC, Canada, 2016.
23. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. *arXiv* **2017**, arXiv:1707.06347. [CrossRef]
24. Henderson, P.; Islam, R.; Bachman, P.; Pineau, J.; Precup, D.; Meger, D. Deep Reinforcement Learning that Matters. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 3207–3214.
25. National Aeronautics and Space Administration. *UAM Vision Concept of Operations (ConOps)—UAM Maturity Level 4*; NASA: Washington, DC, USA, 2020.
26. Goyal, R.; Reiche, C.; Fernando, C.; Cohen, A. Advanced Air Mobility: Demand Analysis and Market Potential of the Airport Shuttle and Air Taxi Markets. *Sustainability* **2021**, *13*, 7421. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.