Optimizing Resource-Constrained Non-Pharmaceutical Interventions for Multi-Cluster Outbreak Control Using Hierarchical Reinforcement Learning

Xueqiao Peng[1] , **Andrew Perrault**[1]
[1]The Ohio State University
peng.969@osu.edu, perrault.17@osu.edu

Abstract

Non-pharmaceutical interventions (NPIs), such as diagnostic testing and quarantine, are crucial for controlling infectious disease outbreaks but are often constrained by limited resources, particularly in the early outbreak stages. In real-world public health settings, resources must be allocated across multiple outbreak clusters that emerge asynchronously, vary in size and risk, and compete for a shared resource budget. We define a cluster as a group of close contacts generated by a single infected index case. Thus, decisions must be made under uncertainty and heterogeneous demands while respecting operational constraints. We formulate this problem as a constrained restless multi-armed bandit and propose a hierarchical reinforcement learning framework. A global controller learns a continuous action cost multiplier that adjusts global resource demand, while a generalized local policy estimates the marginal value of allocating resources to individuals within each cluster. We evaluate the proposed framework in a realistic agent-based simulator of SARS-CoV-2 with dynamically arriving clusters. Across a wide range of system scales and testing budgets, our method consistently outperforms RMAB-inspired and heuristic baselines, improving outbreak control effectiveness by 20–30%. Experiments on up to 40 concurrently active clusters further demonstrate that the hierarchical framework is highly scalable and enables faster decision-making than the RMAB-inspired method.

1 Introduction

The rapid emergence of novel infectious diseases has brought severe challenges to public health, particularly in the early stages of an outbreak when vaccines or targeted treatments are not available. In this stage, outbreak control mainly relies on non-pharmaceutical interventions (NPIs) such as diagnostic testing, contact tracing, and isolation. However, in practice, policymakers must make these decisions under considerable uncertainty. This uncertainty involves not only the potential state of the outbreak but also the policy objectives. For policy objectives, the relative importance of preventing transmission, avoiding unnecessary isolation, and reducing operational burden is often not clear beforehand and may change as the outbreak develops.

These difficulties are even more severe given the limited public health resources. For instance, when testing capacity is insufficient, policymakers must allocate testing based on conflicting needs without a clear standard. In reality, each confirmed case generates a local cluster of exposed individuals, and multiple clusters are often active simultaneously, competing for a shared testing budget. Heterogeneity in cluster size, timing, and symptom prevalence can lead to highly uneven and asynchronous testing demand both spatially and temporally [Rahmandad *et al.*, 2021; Verelst *et al.*, 2016]. Therefore, an effective early response to an outbreak requires a decision-making mechanism that can adapt to uncertainty while strictly adhering to resource constraints.

This setting naturally relates to Restless Multi-Armed Bandit (RMAB) [Whittle, 1988], but directly applying existing deep RMAB methods faces challenges. Most Deep RMAB solvers rely on network architectures (e.g., MLPs or RNNs) which require fixed-dimensional state inputs and a predefined static set of arms [Killian *et al.*, 2021]. This structural limitation makes them unable to handle the asynchronous nature of infectious diseases. Clusters dynamically emerge, resulting in the dimension of the global state space changing over time. Moreover, classical and neural index-based methods often incur significant time-step computational overhead since the Whittle index must be computed or approximated iteratively at each time step [Weber and Weiss, 1990; Nakhleh *et al.*, 2021], increasing computational cost per arm. In large-scale scenarios, this cost becomes prohibitive, making it difficult to perform rapid, low-latency decision-making to respond to an outbreak in real time.

In this work, we propose a hierarchical reinforcement learning framework to allocate the testing resources across multiple dynamically evolving clusters under a global budget. Instead of directly allocating tests, our method separates global coordination from local decision-making. A global

An extended version with supplementary material is available at: https://arxiv.org/abs/2603.19397. Code is available at https://github.com/XueqiaoPeng/Covid_HierRL.git.

controller adjusts the test usage across all the clusters through a continuous cost multiplier, which modulates the test cost perceived by the local policy. The local policy then evaluates the testing decisions at the individual level based on epidemiological features and cost signals provided by the global controller. The local clusters employ pre-trained Deep-Q Networks (DQNs) [Mnih *et al.*, 2015] enhanced with a Transformer [Vaswani *et al.*, 2017] architecture. We trained a generalized DQN that can produce different policies under different test cost signals. This design ensures local decisions remain effective under various objective trade-offs, allowing policymakers to adjust the policy according to their objectives. Final testing decisions are made by ranking candidate actions according to their estimated benefits and selecting the most valuable actions that fit within the available budget. This ensures that testing resources are consistently directed toward the most valuable decisions while remaining feasible under strict operational constraints.

We evaluate our framework using a realistic agent-based simulator of SARS-CoV-2 transmission. Across a wide range of testing budgets and outbreak scales, our method consistently outperforms the heuristic and RMAB-inspired baselines. Our main contributions are:

- We formulate the resource allocation problem as a constrained RMAB process with asynchronous cluster arrivals, bridging the gap between theoretical bandit models and real-world public health needs. In particular, it addresses the challenge of coordinating resources across temporally asynchronous outbreaks.
- We propose a hierarchical reinforcement learning framework that decouples local intervention decisions from global resource allocation. A key contribution is the design of a generalized local DQN that can adapt to different resource constraints without retraining. By combining this with a global Proximal Policy Optimization (PPO) controller, our approach is able to dynamically adjust resource usage while maintaining computational efficiency.
- Our proposed framework improves outbreak control effectiveness by 5–12% relative to the RMAB-inspired methods and 20–30% relative to heuristic strategies, while scaling to scenarios with up to 40 concurrently active clusters and achieving approximately $5\times$ speedup in decision-making.

2 Related Work

Outbreak Control and NPIs. For a long time, research methods for outbreak control have included optimal control, compartment models, and large-scale simulations [Nowzari *et al.*, 2016; Brauer *et al.*, 2017; Probert *et al.*, 2016]. While these classical methods offer important theoretical insights, they often rely on simplifying assumptions, such as population homogeneity or known dynamics. This limits their ability to support adaptive, refined decision-making under uncertainty. More recently, data-driven and reinforcement learning methods have been explored for epidemic control and mitigation policies [Alamo *et al.*, 2021; Kompella *et al.*, 2020]. Related bandit-based approaches have also studied public-health intervention and resource allocation problems [Mate *et al.*, 2020; Killian *et al.*, 2021]. Building on this line of work, Peng et al. [Peng *et al.*, 2023] studied individual-level testing and quarantine decisions under partial observability. However, most existing studies only consider a single population or a fixed decision-making context and do not explore how to coordinate limited resources to deal with multiple outbreak clusters that develop asynchronously over time.

Restless Multi-Armed Bandits (RMAB). Traditional RMAB solutions are typically based on index policies, the most well-known being the Whittle index [Whittle, 1988]. These methods require restrictive assumptions such as indexability and known transition dynamics [Glazebrook *et al.*, 2006; Liu and Zhao, 2010; Avrachenkov and Borkar, 2022]. RMAB formulations have been successfully applied to public health problems such as tuberculosis screening and maternal health outreach [Mate *et al.*, 2020; Biswas *et al.*, 2021]. However, applying RMAB methods to outbreak response introduces additional challenges. Real-world outbreaks consist of clusters that appear asynchronously and vary widely in size, leading to state representations that have variable and time-varying dimensions, which are incompatible with standard deep RMAB architectures. Moreover, operational constraints such as daily testing capacity impose a tight per-timestep budget, whereas many existing approaches enforce constraints only in expectation.

Recent multi-agent reinforcement learning methods have addressed scalability in large-scale populations [Yang *et al.*, 2018; Mao and Perrault, 2026]. However, many decentralized MARL approaches optimize local policies without guaranteeing strict per-timestep global feasibility, while centralized approaches often suffer from combinatorial action spaces as the number of clusters grows. In contrast, our setting requires scalable coordination under hard per-timestep testing budget constraints together with dynamically arriving clusters and variable-sized state representations.

Hierarchical RL and Price-Based Coordination. Motivated by these challenges, our work draws on ideas from hierarchical reinforcement learning and price-based coordination, where the global signals regulate shared resources, while local policies focus on fine-grained decision-making [Barto and Mahadevan, 2003; Nachum *et al.*, 2018; Vezhnevets *et al.*, 2017]. Related dual-based control mechanisms have been explored in networked and constrained systems, where Lagrange multipliers or shadow prices are used to manage global constraints [Bertsekas, 1997; Neely, 2010]. In contrast to prior approaches, we have explicitly designed a framework that can adapt to variable-sized, asynchronously arriving outbreak clusters and enforces strict budget constraints through a deterministic execution layer. By separating global resource regulation from local risk assessment, our approach achieves scalable coordination without incurring the combinatorial complexity of fully centralized control.

3 Problem Description

At the individual cluster level, intervention decisions such as testing and quarantine must balance epidemiological effectiveness against economic and social costs. In our setting, a

cluster refers to a group of close contacts being exposed to a single index case simultaneously. The number of close contacts, demographic composition, and epidemiological characteristics of outbreaks may vary significantly. Following exposure, some contacts will become infected and develop symptoms, while others remain uninfected. Crucially, individuals' true infection status is not directly observable; intervention decisions must be made sequentially under uncertainty.

Prior work by Peng [2023] studies this *single-cluster* optimization problem using reinforcement learning. Each cluster is modeled as an independent sequential decision process, with the objective defined as

$$\frac{-(S_1 + \alpha_2 S_2 + \alpha_3^{\text{true}} S_3)}{N}, \tag{1}$$

where S_1 denotes infectious days before quarantine, S_2 denotes unnecessary quarantine days for uninfected contacts weighted by α_2, and S_3 denotes testing costs weighted by the true per-test cost α_3^{true}. The objective is normalized by the cluster size N. Under the reward structure shown in Equation 1, the quarantine policy has an optimal threshold form that is independent of the testing allocation. Therefore, we fix the quarantine policy and mainly focus on the resource allocation problem.

In this paper, we consider a more challenging scenario in which *multiple* clusters evolve simultaneously and compete for a shared global testing budget. Clusters may activate and deactivate asynchronously over time, reflecting realistic outbreak dynamics across heterogeneous contact groups.

Let $\mathcal{A}_t$ denote the set of active clusters at time t. At each timestep, the global testing budget B limits the total number of tests that can be performed across all active clusters. The system-level objective is to maximize the expected cumulative reward across clusters while respecting this per-timestep budget constraint:

$$\begin{aligned} \max_{\pi} \quad & \mathbb{E}\left[\sum_t \gamma^t \sum_{n \in \mathcal{A}_t} R_n(s_{n,t}, a_{n,t})\right] \\ \text{s.t.} \quad & \sum_{n \in \mathcal{A}_t} C_n(a_{n,t}) \le B, \quad \forall t. \end{aligned} \tag{2}$$

Here, R_n denotes the cluster-level reward defined in Eq. (1), $C_n(a_{n,t})$ is the realized testing cost incurred by cluster n at time t, B is the global testing budget, and $\gamma \in (0, 1)$ is the discount factor.

This cross-cluster coupling introduces a system-level coordination challenge that cannot be solved by optimizing individual clusters alone. Directly learning the joint discrete resource allocation across clusters leads to an excessively large action space and poor scalability, while purely local strategies cannot take into account the global resource scarcity. Therefore, our goal is to design a scalable decision framework that can coordinate tests between heterogeneous, asynchronously activated clusters and strictly enforce the global test budget at each time step.

4 Methods

4.1 RMAB Formulation and Lagrangian Relaxation

The multi-cluster problem in Eq. (2) can be formulated as a Restless Multi-Armed Bandit (RMAB) [Whittle, 1988], where each active cluster is an arm evolving under partial observability and competing for a shared limited budget. Each cluster-level decision problem can be modeled as a Partially Observable Markov Decision Process (POMDP) [Killian *et al.*, 2021].

To handle global constraints on test resources, we adopt a Lagrangian relaxation approach, introducing a penalty parameter λ to relax the budget constraint. For a fixed λ, the relaxed objective can be written as:

$$\mathcal{L}(\lambda) = \mathbb{E}\left[\sum_t \gamma^t \sum_{n \in \mathcal{A}_t} \big(R_n(s_{n,t}, a_{n,t}) - \lambda C_n(a_{n,t})\big)\right] + \frac{\lambda B}{1-\gamma}. \tag{3}$$

This relaxation decomposes the global constrained problem into independent subproblems indexed by λ, each trading off reward against a penalized testing cost. Equivalently,

$$J(\lambda) = \frac{\lambda B}{1-\gamma} + \sum_n V_n(s_n, \lambda), \tag{4}$$

where $V_n(s_n, \lambda)$ denotes the value of arm n under the testing cost.

Rather than explicitly solving this dual optimization via linear programming, we use this formulation as a conceptual guide for our proposed framework.

Crucially, in our setting, the penalty parameter λ can be naturally interpreted as the cost coefficient α_3^{active} for each test, and is parameterized as:

$$\alpha_3^{\text{active}} = m_t \cdot \alpha_3^{\text{true}}, \tag{5}$$

where $m_t \ge 1$ is learned by a global controller.

This formulation implies that when the budget is tight, the controller increases m_t to raise the effective test cost, thus suppressing testing actions in local policies.

4.2 Hierarchical Reinforcement Learning Framework

Directly learning globally constrained testing allocations across dynamically arriving clusters requires solving a large combinatorial action space while enforcing strict per-timestep feasibility. To improve scalability, we propose a hierarchical reinforcement learning (HRL) framework that separates global resource coordination from local intervention decisions, as shown in Figure 1. Instead of directly allocating tests, the global controller adjusts a shared cost multiplier that regulates overall testing demand, while local policies determine testing priorities and quarantine decisions within each cluster.

At each timestep, the generalized local policy evaluates the marginal value of testing each individual in all active clusters under the current perceived test cost. These evaluations produce candidate testing actions. The global controller then observes system-level statistics and outputs a scalar multiplier

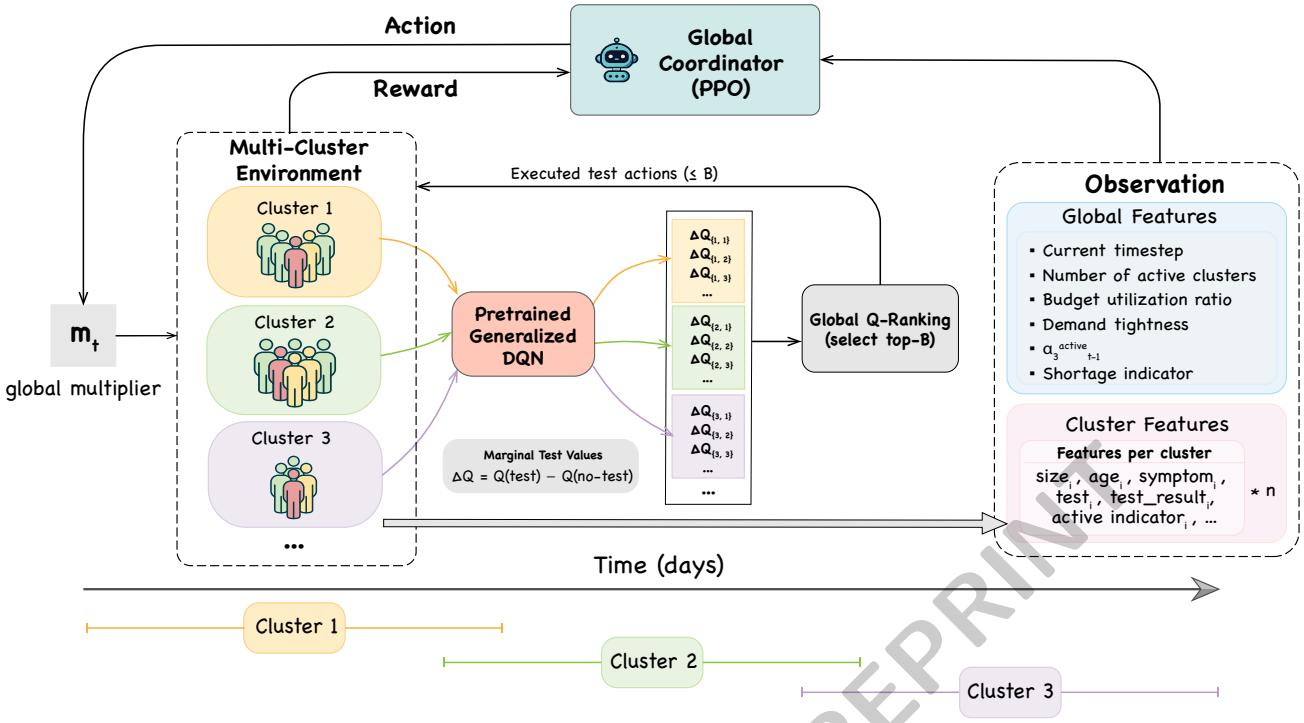

Figure 1: Overview of the proposed hierarchical RL framework for multi-cluster outbreak control. A global PPO controller adjusts a shared testing cost multiplier, which modulates the perceived test cost used by a pretrained generalized DQN applied to each active cluster. Based on these cost-conditioned local value estimates, a global Q-ranking policy selects testing actions across clusters while strictly enforcing the hard global budget constraint.

that scales the perceived testing cost shared across clusters, allowing it to regulate resource usage without selecting individuals directly.

When induced testing demand exceeds the available budget, a deterministic global Q-ranking layer pools candidate tests across clusters, ranks them by marginal value, and executes only the highest-value tests up to the budget limit. This guarantees strict budget feasibility while preserving local testing priorities. Together, these components enable scalable coordination across heterogeneous, asynchronously arriving clusters.

Local Policy: Generalized Transformer-Based DQN

At the cluster level, we employ a generalized Deep Q-Network (DQN) [Mnih *et al.*, 2015] to estimate the marginal value of testing individual contacts. To handle the dynamic number of individuals within each cluster, we employ a Transformer [Vaswani *et al.*, 2017] in DQN. Quarantine decisions follow the threshold policy proposed in [Peng *et al.*, 2023], and the learnable decision is whether to administer a diagnostic test (`test` vs. `no-test`) to each individual.

A key challenge is that optimal testing depends on the actual test cost, which affects testing usage within each cluster. Rather than retrain separate policies for different cost regimes, we use a *generalized* DQN that conditions on the perceived test cost.

Formally, the observation for cluster n at time t is:

$$o_{n,i,t} = [\, s_{n,i,t},\ \alpha_{3,t}^{\text{active}}],$$

where $s_{n,i,t}$ contains belief-based features, including estimated infection probabilities, recent symptoms, and testing history. Additional implementation details are provided in the supplementary material.

The DQN outputs action-values:

$$Q_\phi(o_{n,i,t}, a), \quad a \in \{\texttt{no-test}, \texttt{test}\}.$$

To ensure consistent behavior of the learned policy, we introduce a gradient-based regularization term. From a public health decision-support perspective, this enforces a basic economic and clinical principle: when the actual cost of testing increases, the tendency to recommend testing should not increase. Without such a constraint, deep reinforcement learning agents may exhibit counterintuitive behaviors due to function approximation errors, such as performing more tests under higher costs. By explicitly enforcing monotonicity with respect to the testing cost, we improve the logical consistency and interpretability of the learned policy in healthcare decision-support settings [Holzinger *et al.*, 2019].

The total loss at each training step is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{TD}} + \lambda_{\text{gp}} \cdot \mathcal{L}_{\text{grad}}, \tag{6}$$

where $\mathcal{L}_{\text{TD}}$ is the standard temporal-difference loss, and $\mathcal{L}_{\text{grad}}$ is the gradient penalty term:

$$\mathcal{L}_{\text{grad}} = \mathbb{E}_{o_t,\, a_t}\left[\left(\max\left(0,\ \frac{\partial Q(o_t, a_t)}{\partial \alpha_3} - g_{\text{target}}\right)\right)^2\right] \tag{7}$$

Here, $g_{\text{target}} < 0$ is a target gradient that encourages decreasing Q-values with higher α_3. The coefficient λ_{gp} controls the strength of this regularization.

After training, the policy adapts continuously to different α_3 inputs: higher values suppress testing, while lower values encourage more aggressive testing behavior. We emphasize that the environment reward is always computed using the true per-test cost α_3^{true}, while the local DQN takes $\alpha_{3,t}^{\text{active}}$ as input. The $\alpha_{3,t}^{\text{active}}$ is provided by the global controller to control the test usage within each cluster.

Because the testing penalty directly affects only the `test` action, while the `no-test` action is largely invariant to α_3, this regularization also indirectly stabilizes the marginal ranking score ΔQ used in global Q-ranking.

Global Policy: PPO-Based Controller

To coordinate testing across dynamically activated clusters, we introduce a global controller that adjusts the test usage across the system using time-varying cost signals. The controller is implemented using Proximal Policy Optimization (PPO) [Schulman *et al.*, 2017] and operates at the system level, rather than selecting test actions for individual contacts.

The PPO strategy outputs a scalar multiplier m_t, which parameterizes the active test cost coefficient defined in the RMAB formulation. This multiplier allows the controller to adjust the overall detection demand from each local cluster, while remaining independent of individual-level decisions. The environment evaluates rewards using the fixed true per-test cost α_3^{true}.

Importantly, the controller does not directly enforce feasibility. When total testing demand is below the available budget, the final allocation is identical to that under the true cost, regardless of the specific value of m_t. When demand would exceed capacity, increasing m_t will increase perceived test costs and suppress demand, while the subsequent ranking-based allocation mechanism ensures strict adherence to hard budget constraints.

At each time step t, the PPO policy operates based on a compressed observation, which includes system-level information and cluster features. This enables scalable decision-making as clusters emerge. Additional details on the controller observation space and feature construction are provided in the supplementary material.

Global Q-Ranking for Test Execution

At each decision step, the generalized local DQN evaluates testing decisions for all individuals in all active clusters under the current perceived test cost. For each individual, the DQN outputs action-value estimates for two actions: `test` and `no-test`. The difference between these values,

$$\Delta Q_{n,i} = Q_\phi(o_{n,i,t}, \texttt{test}) - Q_\phi(o_{n,i,t}, \texttt{no-test}),$$

represents the marginal benefit of administering a test, with larger values indicating higher testing priority.

Because all local Q-values are conditioned on the same perceived test cost, these marginal scores are directly comparable across clusters. To enforce the hard global testing budget, we adopt a global Q-ranking policy: candidate testing actions from all active clusters are pooled and ranked according to their marginal value ΔQ in descending order. Testing is then allocated to the top-ranked candidates with positive marginal value, up to the available budget B at the current timestep. If fewer than B candidates have positive marginal value, the remaining budget is left unused, while all other individuals are assigned the `no-test` action. This procedure guarantees feasibility by construction.

Global Q-ranking serves as a deterministic execution layer that separates marginal value estimation from budget enforcement. This decomposition avoids the instability of directly learning discrete, globally constrained allocations while remaining interpretable, as tests are allocated to individuals with the highest estimated marginal benefit under the current cost regime.

5 Experiments

5.1 Experimental Setup

We evaluate the proposed framework using a realistic agent-based simulator of SARS-CoV-2 transmission that models infection, testing, and isolation at the individual level, allowing intervention decisions to directly affect disease dynamics and future observability. Epidemiological parameters are drawn from public health literature, and the simulator explicitly captures high-transmission scenarios through a small fraction of highly infectious index cases, reflecting the heavy-tailed risk observed in real outbreaks.

The environment is organized as a multi-cluster system induced by contact tracing, where each confirmed case generates a cluster of exposed individuals evolving independently over time. Multiple clusters may be active concurrently and compete for a shared global testing budget, while all transmission and intervention dynamics remain individual-level.

At each timestep, a hard global testing budget constrains the total number of tests that can be executed across clusters. During training, the budget is randomized to improve robustness under different resource regimes. We evaluate performance under both synchronous and asynchronous cluster activation settings using the cumulative multi-objective reward defined in Equation 1, normalized by cluster size. Results are averaged over five random seeds. Additional details on simulator dynamics, cluster generation, training environments, and budget randomization are provided in the supplementary material.

5.2 Comparison Policies

We compare our hierarchical reinforcement learning approach against several baseline policies for allocating limited testing resources under a hard global budget constraint. All methods are evaluated in the same environment and strictly enforce feasibility at every timestep.

Across all methods, testing and quarantine decisions are decoupled. Unless otherwise stated, all baselines share the same fixed quarantine mechanism, while Symp-AvgRand instead applies symptom-based quarantine as a representative public health heuristic. Unless otherwise specified, all experiments use a fixed quarantine cost coefficient $\alpha_2 = 0.1$ and a true per-test cost $\alpha_3^{\text{true}} = 0.05$.

Symp-AvgRand. The global testing budget is evenly divided across active clusters, and individuals are selected uniformly at random for testing. Quarantine is determined using observed symptoms and positive test results.

Thres-AvgRand. Testing resources are evenly allocated across active clusters, while individuals within each cluster are selected uniformly at random for testing.

Thres-SizeRand. Testing resources are allocated proportionally to cluster sizes, while individuals within each cluster are selected uniformly at random for testing.

Fixed-M-QR. A fixed global cost multiplier (M=1) is applied to the pretrained generalized DQN, inducing a constant perceived testing cost across all timesteps. This baseline evaluates whether a static scarcity signal is sufficient for coordinating testing decisions. Results with alternative fixed M values are included in the supplementary material.

Bin-M-QR. The global cost multiplier is adjusted online via binary search to satisfy the testing budget at each timestep. This method provides a reactive RMAB-inspired coordination mechanism analogous to Whittle-index-style resource balancing, but without requiring indexability assumptions or explicit index computation.

Hierarchical PPO (Ours). Our approach learns a time-varying global cost multiplier using a PPO-based controller, enabling adaptive coordination of testing decisions based on the evolving global system state.

All learning-based methods share the same pretrained generalized local DQN and differ only in how the global cost multiplier is selected. Additional implementation details are provided in the supplementary material.

5.3 Generalized vs. Fixed DQN

Table 1 shows the results comparing the generalized DQN with fixed-cost DQN policies. The generalized DQN is trained with α_3 randomly sampled from $[0.0, 0.1]$, while each fixed-cost DQN is trained under a single fixed α_3 value. Both generalized and fixed-cost DQN models use a fixed quarantine cost $\alpha_2 = 0.1$.

Across all evaluated settings, the generalized policy achieves performance comparable to that of specialized fixed-cost policies. This confirms that a single conditional model can effectively capture the optimal policy across varying cost regimes without sacrificing precision. This result has significant practical implications. Rather than training and maintaining a suite of separate models for every possible cost scenario, a single generalized model can be deployed to handle a wide range of resource conditions.

This substantially reduces computational overhead, as policy adaptation to different resource conditions requires only changing the input cost parameter rather than retraining the model. Moreover, the generalized formulation enables efficient what-if analysis. Policymakers can rapidly evaluate how testing behavior and outcomes would change under alternative cost or budget assumptions using a single trained model. This flexibility is particularly valuable in real-world public health settings, where resource constraints can shift over time and rapid scenario evaluation is often required.

	α_3=0.01	α_3=0.05	α_3=0.08	α_3=0.10
Fixed-DQN	-0.23 ± 0.02	-0.33 ± 0.02	-0.42 ± 0.02	-0.50 ± 0.04
Gen-DQN	-0.25 ± 0.02	-0.34 ± 0.02	-0.41 ± 0.02	-0.47 ± 0.03

Table 1: Generalized vs. Fixed DQN performance under varying test penalty cost (α_3). Results indicate that the generalized policy achieves comparable performance, demonstrating adaptability to different test penalties.

5.4 Analysis

Table 2 summarizes performance under asynchronous cluster activation. Hier-PPO achieves the highest average return across all settings. Compared with Bin-M-QR, Hier-PPO improves overall returns by approximately 5%–12% under low-to-moderate budgets. Moreover, all globally coordinated Q-ranking methods (Fixed-M-QR, Bin-M-QR, and Hier-PPO) substantially outperform heuristic baselines, yielding 20%–30% improvements over threshold-based random allocation in representative constrained settings (e.g., around -1.11 vs. -0.65 for $\#C$=20, $\#B$=40). When budgets are loose relative to demand, performance gaps among globally coordinated methods narrow, consistent with diminishing returns from scarcity control.

Table 3 decomposes performance into S_1 (unquarantined infections), S_2 (unnecessary quarantine), and S_3 (testing cost). Under tight budgets, Symp-AvgRand has high S_1, while Thres-AvgRand and Thres-SizeRand reduce infections at the cost of higher unnecessary quarantine and testing. In contrast, Q-ranking methods achieve lower S_1 while keeping S_3 near the effective budget limit. Hier-PPO further reduces S_1 and S_2, suggesting more effective prioritization of high-impact tests.

Under loose budgets, threshold-based random allocation tests nearly everyone, achieving low S_1 but incurring much higher S_3. Q-ranking methods instead test only when the estimated marginal benefit is positive, yielding a more efficient cost–benefit trade-off.

Table 3 further highlights the distinction between Bin-M-QR and Hier-PPO. Both methods intervene only when the proposed testing demand exceeds the per-timestep budget. However, Bin-M-QR relies on binary search to adjust the cost multiplier reactively at each timestep, whereas Hier-PPO learns a smooth, state-dependent adjustment policy. This learned policy better aligns the scarcity signal with the evolving system state when constraints bind, yielding consistently lower S_1 and S_2 at comparable testing costs.

Computational efficiency. In addition to improved performance, Hier-PPO is computationally more efficient at decision time, reducing runtime by approximately 4–8$\times$ compared with Bin-M-QR. Hier-PPO directly outputs the test cost multiplier through a single forward pass of the global PPO, whereas Bin-M-QR determines the multiplier via iterative binary search and therefore requires multiple demand evaluations per timestep. Additional runtime analysis across different numbers of active clusters is provided in the supplementary material.

	Symp-AvgRand	Thres-AvgRand	Thres-SizeRand	Fixed-M-QR(1)	Bin-M-QR	Hier-PPO
#C=10, #B=20	-2.12 ± 0.03	-1.02 ± 0.01	-0.93 ± 0.02	-0.64 ± 0.01	-0.61 ± 0.01	$\mathbf{-0.56\pm0.01}$
#C=10, #B=200	-2.64 ± 0.03	-2.04 ± 0.01	-2.12 ± 0.01	-0.67 ± 0.01	-0.66 ± 0.02	$\mathbf{-0.63\pm0.01}$
#C=20, #B=40	-2.31 ± 0.04	-1.11 ± 0.01	-1.01 ± 0.02	-0.65 ± 0.01	-0.64 ± 0.01	$\mathbf{-0.57\pm0.01}$
#C=20, #B=400	-2.79 ± 0.02	-2.15 ± 0.01	-2.21 ± 0.01	-0.68 ± 0.01	-0.68 ± 0.01	$\mathbf{-0.63\pm0.01}$
#C=40, #B=80	-2.34 ± 0.03	-1.13 ± 0.01	-1.02 ± 0.01	-0.66 ± 0.01	-0.63 ± 0.02	$\mathbf{-0.59\pm0.01}$
#C=40, #B=800	-2.82 ± 0.02	-2.18 ± 0.06	-1.57 ± 0.02	-0.78 ± 0.01	-0.68 ± 0.02	$\mathbf{-0.64\pm0.01}$

Table 2: Performance comparison of various test allocation methods under asynchronous cluster activation environments. Results highlight the advantage of the proposed hierarchical allocation PPO (Hier-PPO) method (#C: number of clusters, #B: daily testing budget).

Method	$(\#C, \#B)$	S1	S2	S3
Symp-AvgRand	$(20, 40)$	1.64 ± 0.04	0.44 ± 0.02	12.54 ± 0.09
Thres-AvgRand		0.33 ± 0.01	1.59 ± 0.03	12.54 ± 0.09
Thres-SizeRand		0.39 ± 0.01	1.58 ± 0.03	9.33 ± 0.06
Fixed-M-QR		0.35 ± 0.01	1.02 ± 0.02	3.45 ± 0.05
Bin-M-QR		0.36 ± 0.01	0.99 ± 0.02	3.35 ± 0.04
Hier-PPO		0.30 ± 0.04	0.95 ± 0.02	3.41 ± 0.06
Symp-AvgRand	$(20, 400)$	0.94 ± 0.02	0.58 ± 0.01	35.95 ± 0.01
Thres-AvgRand		0.17 ± 0.01	1.81 ± 0.04	35.95 ± 0.05
Thres-SizeRand		0.15 ± 0.01	1.76 ± 0.04	37.71 ± 0.05
Fixed-M-QR		0.37 ± 0.01	1.09 ± 0.03	4.07 ± 0.04
Bin-M-QR		0.36 ± 0.01	1.08 ± 0.03	4.08 ± 0.04
Hier-PPO		0.30 ± 0.01	0.99 ± 0.02	4.10 ± 0.03

Table 3: Comparison of different methods under asynchronous settings for $\#C$=20 with two global budgets. Results are reported for metrics S_1, S_2, and S_3.

6 Discussion

This work presents a hierarchical reinforcement learning framework for allocating limited testing resources across dynamically evolving outbreak clusters under hard per-timestep budget constraints. By separating global resource coordination from local intervention decisions, the framework provides a scalable approach for multi-cluster outbreak control in realistic public health settings.

A central feature of the proposed framework is the use of cost-based coordination rather than direct centralized allocation. Instead of optimizing a joint policy over a large combinatorial action space, the framework regulates aggregate testing demand through a shared cost signal, while local policies estimate the marginal value of testing individual contacts. This decomposition is particularly useful in outbreak settings, where clusters arrive asynchronously and resource pressure changes over time. By combining cost-based coordination with ranking-based execution, the framework can adapt testing intensity to system-level scarcity while still prioritizing individuals with the highest estimated marginal benefit. More broadly, this design suggests that price-based coordination with learned local value functions can provide a scalable alternative to fully centralized control in large-scale sequential resource allocation problems.

This framework uses a threshold-based quarantine policy rather than jointly learning testing and quarantine decisions. Under the POMDP formulation and reward structure considered in this paper, the threshold policy is provably optimal [Peng *et al.*, 2023]. Fixing the quarantine policy provides theoretical guarantees while also improving training stability in large, multi-cluster environments. The quarantine cost parameter α_2 also plays an important role because it determines the value of the information provided by testing. When α_2 is very small, unnecessary quarantine becomes inexpensive, making it optimal to quarantine nearly all contacts regardless of testing outcomes. In this regime, testing provides limited additional decision-making value because intervention decisions become largely independent of test results. Conversely, when α_2 is large, quarantine becomes costly and is applied less frequently, reducing the practical impact of testing because positive test results are less likely to trigger quarantine actions. For this reason, the main experiments focus on the setting with $\alpha_2 = 0.1$. We also explored extensions in which both quarantine and testing costs are generalized; additional details are provided in the supplementary material.

Several limitations point to important directions for future work. First, although the global cost signal provides an interpretable summary of system-level scarcity, local testing decisions are produced by neural networks and are not directly interpretable at the individual level. This lack of transparency is a common challenge for reinforcement learning methods applied to public health decision support and may limit trust and adoption in practice [Amann *et al.*, 2020]. Improving interpretability and alignment with public health reasoning remains an important direction.

Second, the current framework does not explicitly model behavioral factors such as imperfect compliance with testing and quarantine guidance. In real outbreaks, adherence changes over time due to fatigue, risk perception, and trust in public health messaging, all of which can substantially affect intervention effectiveness [Funk *et al.*, 2010; Verelst *et al.*, 2016]. Modeling these behavioral dynamics remains an open challenge. Recent advances in large language models may provide promising opportunities for incorporating more realistic patterns of human decision-making and compliance inferred from unstructured behavioral data [Park *et al.*, 2023; Argyle *et al.*, 2023].

Overall, this work demonstrates the potential of hierarchical reinforcement learning as a decision-support tool for outbreak response under resource constraints. Rather than replacing human judgment, the proposed framework is designed to support public health practitioners in reasoning about trade-offs, prioritizing testing, and adapting interventions. We view this approach as a step toward AI systems that enable more informed, timely, and accountable decision-making in real-world public health settings.

References

[Alamo *et al.*, 2021] Teodoro Alamo, Daniel G Reina, Pablo Millán Gata, Victor M Preciado, and Giulia Giordano. Data-driven methods for present and future pandemics: Monitoring, modelling and managing. *Annual Reviews in Control*, 52:448–464, 2021.

[Amann *et al.*, 2020] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I Madai, and Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, 20(1):310, 2020.

[Argyle *et al.*, 2023] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

[Avrachenkov and Borkar, 2022] Konstantin E Avrachenkov and Vivek S Borkar. Whittle index based q-learning for restless bandits with average reward. *Automatica*, 139:110186, 2022.

[Barto and Mahadevan, 2003] Andrew G Barto and Sridhar Mahadevan. Recent advances in hierarchical reinforcement learning. *Discrete event dynamic systems*, 13(4):341–379, 2003.

[Bertsekas, 1997] Dimitri P Bertsekas. *Nonlinear Programming*. Athena Scientific, 1997.

[Biswas *et al.*, 2021] Arpita Biswas, Gaurav Aggarwal, Pradeep Varakantham, and Milind Tambe. Learn to intervene: An adaptive learning policy for restless bandits in application to preventive healthcare. *arXiv preprint arXiv:2105.07965*, 2021.

[Brauer *et al.*, 2017] Fred Brauer, Carlos Castillo-Chavez, and Zhilan Feng. *Mathematical epidemiology*. Springer, 2017.

[Funk *et al.*, 2010] Sebastian Funk, Marcel Salathé, and Vincent AA Jansen. Modelling the influence of human behaviour on the spread of infectious diseases: a review. *Journal of the Royal Society Interface*, 7(50):1247–1256, 2010.

[Glazebrook *et al.*, 2006] Kevin D Glazebrook, Diego Ruiz-Hernandez, and Christopher Kirkbride. Some indexable families of restless bandit problems. *Advances in Applied Probability*, 38(3):643–672, 2006.

[Holzinger *et al.*, 2019] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 9(4):e1312, 2019.

[Killian *et al.*, 2021] Jackson A Killian, Andrew Perrault, and Milind Tambe. Beyond" to act or not to act": Fast lagrangian approaches to general multi-action restless bandits. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pages 710–718, 2021.

[Kompella *et al.*, 2020] Varun Kompella, Roberto Capobianco, Stacy Jong, Jonathan Browne, Spencer Fox, Lauren Meyers, Peter Wurman, and Peter Stone. Reinforcement learning for optimization of covid-19 mitigation policies. *arXiv preprint arXiv:2010.10560*, 2020.

[Liu and Zhao, 2010] Keqin Liu and Qing Zhao. Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. *IEEE Transactions on Information Theory*, 56(11):5547–5567, 2010.

[Mao and Perrault, 2026] Yi Mao and Andrew Perrault. Optimizing urban service allocation with time-constrained restless bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 39025–39032, 2026.

[Mate *et al.*, 2020] Aditya Mate, Jackson Killian, Haifeng Xu, Andrew Perrault, and Milind Tambe. Collapsing bandits and their application to public health intervention. *Advances in Neural Information Processing Systems*, 33:15639–15650, 2020.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[Nachum *et al.*, 2018] Ofir Nachum, Shixiang Shane Gu, Honglak Lee, and Sergey Levine. Data-efficient hierarchical reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

[Nakhleh *et al.*, 2021] Khaled Nakhleh, Santosh Ganji, Ping-Chun Hsieh, I-Hong Hou, and Srinivas Shakkottai. Neurwin: Neural whittle index network for restless bandits via deep rl. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 828–839. Curran Associates, Inc., 2021.

[Neely, 2010] Michael J Neely. Stochastic network optimization with application to resource allocation and queueing. *Synthesis Lectures on Communication Networks*, 3:1–211, 2010.

[Nowzari *et al.*, 2016] Cameron Nowzari, Victor M Preciado, and George J Pappas. Analysis and control of epidemics: A survey of spreading processes on complex networks. *IEEE Control Systems Magazine*, 36(1):26–46, 2016.

[Park *et al.*, 2023] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

[Peng *et al.*, 2023] Xueqiao Peng, Jiaqi Xu, Xi Chen, Dinh Song An Nguyen, and Andrew Perrault. Using reinforcement learning for multi-objective cluster-level optimization of non-pharmaceutical interventions for infec-

tious disease. In *Machine Learning for Health (ML4H)*, pages 445–460. PMLR, 2023.

[Probert *et al.*, 2016] William JM Probert, Katriona Shea, Christopher J Fonnesbeck, Michael C Runge, Tim E Carpenter, Salome Dürr, M Graeme Garner, Neil Harvey, Mark A Stevenson, and Colleen T Webb. Decision-making for foot-and-mouth disease control: objectives matter. *Epidemics*, 15:10–19, 2016.

[Rahmandad *et al.*, 2021] Hazhir Rahmandad, Tse Yang Lim, and John Sterman. Behavioral dynamics of covid-19: estimating underreporting, multiple waves, and adherence fatigue across 92 nations. *System dynamics review*, 37(1):5–31, 2021.

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

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Verelst *et al.*, 2016] Frederik Verelst, Lander Willem, and Philippe Beutels. Behavioural change models for infectious disease transmission: a systematic review (2010–2015). *Journal of The Royal Society Interface*, 13(125):20160820, 2016.

[Vezhnevets *et al.*, 2017] Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. Feudal networks for hierarchical reinforcement learning. In *International conference on machine learning*, pages 3540–3549. PMLR, 2017.

[Weber and Weiss, 1990] Richard R Weber and Gideon Weiss. On an index policy for restless bandits. *Journal of applied probability*, 27(3):637–648, 1990.

[Whittle, 1988] Peter Whittle. Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, 25(A):287–298, 1988.

[Yang *et al.*, 2018] Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. In *International conference on machine learning*, pages 5571–5580. PMLR, 2018.