HARL: Hierarchical Reinforcement Learning for Real-Time Policy Optimization in Complex Logistics Networks

Hadi Aghazadeh
Department of Geomatics Engineering, University of Calgary
Calgary, Alberta, Canada
hadi.aghazadeh@ucalgary.ca

Xin Wang*
Department of Geomatics Engineering, University of Calgary
Calgary, Alberta, Canada
xcwang@ucalgary.ca

Abstract

Efficiently managing logistics operations, particularly Vehicle Routing Problems (VRPs), is critical in modern supply chains. These operations are often characterized by complex challenges including heterogeneous vehicle fleets, diverse demand types, and stochastic environmental factors like travel times, all requiring real-time adaptive decision-making. Existing approaches often struggle to simultaneously address these multifaceted issues. This paper introduces HARL (Hierarchical Emergency Logistics Planning with Reinforcement Learning), a novel framework designed for real-time policy optimization in such complex logistics scenarios. HARL features an attention-based policy optimizer with a unique hierarchical decoder architecture and dilated temporal convolutions to manage intricate action spaces and temporal dependencies. Trained using the REINFORCE algorithm, the model dynamically adapts to changing conditions. We demonstrate HARL's effectiveness through experiments on synthetic VRP instances and a real-world case study derived from disaster response logistics. Results show that HARL significantly outperforms state-of-the-art reinforcement learning baselines and traditional heuristics in both solution quality and computational efficiency, offering a robust and generalizable approach for complex VRP research and AI-driven supply chain optimization.

CCS Concepts

• **Computing methodologies → Reinforcement learning**; • **Applied computing** → *Transportation*; • **Theory of computation** → *Routing and network design problems.*

Keywords

Vehicle Routing Problem, Deep Reinforcement Learning, Hierarchical Policy, Logistics Optimization, Stochastic Environments, Real-Time Decision Making

ACM Reference Format:
Hadi Aghazadeh and Xin Wang. 2025. HARL: Hierarchical Reinforcement Learning for Real-Time Policy Optimization in Complex Logistics Networks. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

*Corresponding author.

International Conference on Knowledge Discovery and Data Mining, Toronto, CA

ACM ISBN 978-1-4503-XXXX-X/2025/08
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The efficient flow of goods and services is the backbone of modern economies, with logistics operations and Vehicle Routing Problems (VRPs) playing a pivotal role in supply chain performance [5]. However, real-world logistics are fraught with complexities that challenge traditional optimization approaches. Disaster response logistics starkly exemplifies these challenges: events like Hurricane Katrina, where the Defense Logistics Agency distributed millions of meals and billions of gallons of fuel [2], or the 2023 Canadian wildfires necessitating large-scale evacuations and supply distribution [8], underscore the critical need for highly adaptive and responsive systems. While these emergency scenarios are extreme, the core operational difficulties resonate across a wide spectrum of general logistics and supply chain management.

The intricate nature of such logistics problems is illustrated in Figure 1. Consider a central hub (H_0) tasked with distributing essential resources (e.g., water, food, medical supplies) to various help centers using a heterogeneous fleet of vehicles (e.g., helicopters, specialized trucks, standard vans), each with different capacities and speeds.

In a stable situation (Scenario 1 in Figure 1), where road conditions are known and demand distribution is fixed, an optimal routing plan can be derived using standard optimization techniques. Vehicles are assigned to routes and demand types based on their capabilities to minimize costs or delivery times. However, real-world logistics, especially in dynamic contexts like disaster response or even daily urban deliveries, rapidly shift towards more complex situations (Scenario 2). Here, unforeseen events like major road blockages drastically alter travel times, and the spatial distribution or intensity of demands can change abruptly. A plan optimized for Scenario 1 would become highly suboptimal or even infeasible in Scenario 2. This necessitates a system capable of real-time re-evaluation and adaptation to maintain operational effectiveness.

This illustrative example highlights three primary interacting challenges that make real-time VRP optimization particularly difficult:

(1) **Heterogeneous Fleets and Demands:** Logistics networks typically involve diverse vehicle types (e.g., trucks, vans, drones, specialized vehicles) each with unique capacities, operating costs, speeds, and compatibilities with different cargo or service types. Simultaneously, demands are not uniform; they vary by type (e.g., perishable goods, bulk items, medical

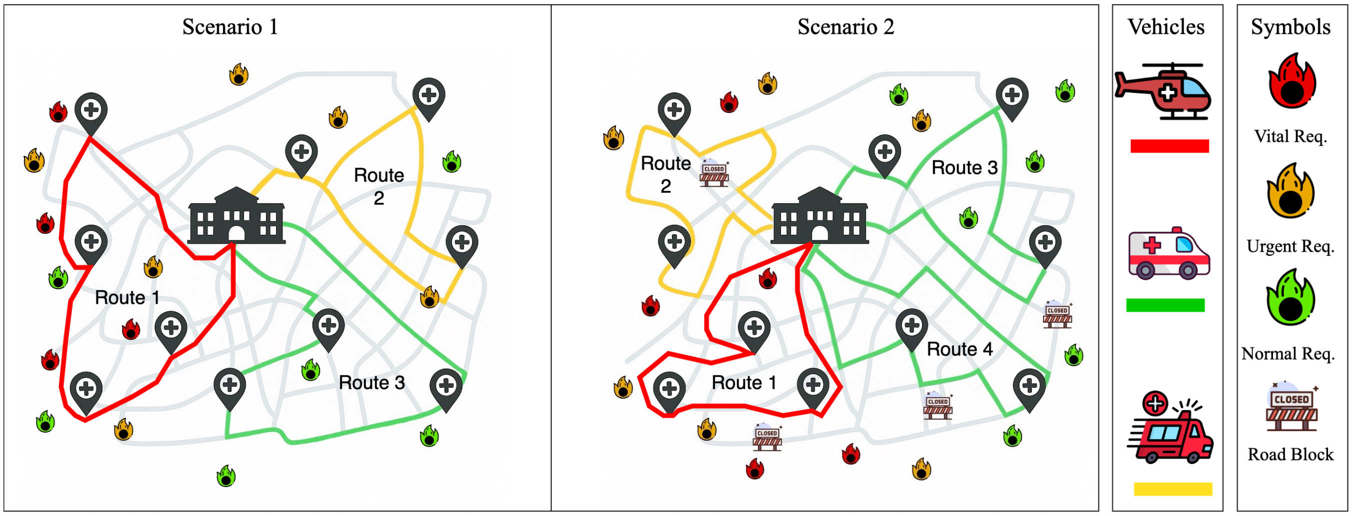


Figure 1: Illustrative scenarios highlighting complexities in logistics operations. Scenario 1 depicts a stable situation amenable to pre-planning. Scenario 2 introduces dynamic disruptions (e.g., blocked routes, shifting demand patterns).

supplies), priority, and handling requirements. Effectively matching the right vehicle to the right demand at the right time is a significant combinatorial challenge.

(2) **Stochastic and Dynamic Environments:** Travel times are rarely deterministic due to traffic congestion, weather conditions, road closures, or (in disaster contexts) damaged infrastructure [4]. Demand itself can be dynamic, with new orders arriving or existing ones changing. This environmental uncertainty requires solutions that can adapt to evolving conditions rather than relying on static plans.

(3) **Real-Time Decision-Making:** The dynamic nature of these factors necessitates rapid re-planning and decision-making. Pre-computed optimal routes can quickly become obsolete, leading to inefficiencies or failures if the system cannot respond in near real-time to new information [13].

Traditional optimization methods, such as Mixed Integer Linear Programming (MILP), often struggle with the scale and dynamism of such problems, especially when all three challenges are present [3]. Heuristic and metaheuristic approaches offer more scalability but may not guarantee optimality or adapt sufficiently quickly. Deep Reinforcement Learning (DRL) has shown considerable promise for tackling complex, sequential decision-making problems under uncertainty [6, 7]. However, many existing DRL applications to VRPs tend to simplify one or more of the aforementioned heterogeneities or assume deterministic settings, and often do not explicitly address the deeply hierarchical nature of the action space (e.g., selecting a location, then a specific vehicle from a diverse fleet, then a specific demand type to service).

To address these limitations, we introduce HARL (Hierarchical Adaptive Routing for Logistics). While motivated and validated using a challenging disaster logistics scenario, HARL is designed as a general framework for real-time policy optimization in complex VRPs. Its main contributions are:

- A unified DRL framework that simultaneously considers heterogeneous demand types, a diverse fleet of vehicles with varying capabilities for each demand type, and stochastic travel times.
- An end-to-end, attention-based policy optimizer using a policy-gradient DRL model capable of generating adaptive real-time policies.
- A novel hierarchical decoder architecture specifically designed to manage the complex, multi-level action space inherent in assigning heterogeneous vehicles to heterogeneous demands at various locations under stochastic conditions.
- The integration of dilated temporal convolutions within the decoder to capture long-range spatio-temporal dependencies and sequential patterns in routing decisions.

This work presents a significant step towards more robust and intelligent AI systems for complex, real-world logistics and supply chain optimization.

2 Related Works

The VRP and its variants have been extensively studied, with traditional approaches often relying on Mixed Integer Linear Programming (MILP) [3, 10] and metaheuristics like Tabu Search or Genetic Algorithms [11]. While powerful, these methods can be computationally intensive for large-scale, dynamic problems and often assume deterministic environments.

Incorporating stochasticity, particularly in travel times or demand, has been addressed [4, 13], but often with simplified assumptions about resource heterogeneity. Some works consider heterogeneous fleets or multi-modal transport [1], yet few integrate this with comprehensive stochastic modeling and real-time decision-making requirements.

Deep Reinforcement Learning (DRL) has emerged as a promising direction for solving VRPs, with notable attention-based models

[6] and policy optimization techniques [7]. These learning-based approaches can generate solutions quickly once trained. However, many existing DRL models for VRPs do not fully address the combined challenge of managing heterogeneous resources (both vehicles and demands with specific compatibilities), stochastic travel times, and a deeply hierarchical action space in a real-time context. Our work aims to tackle these aspects simultaneously through a novel hierarchical policy architecture. Few models, to our knowledge, holistically address all three aspects—heterogeneous demands/fleets and stochastic travel times—within a unified real-time DRL framework for complex logistics problems.

3 The HARL Framework

To apply DRL, we formulate the logistics optimization problem as a Markov Decision Process (MDP), defined by a tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$.

3.1 Problem Formulation

State Space $\mathcal{S}$: The state $s_t \in \mathcal{S}$ at time step t encapsulates all relevant information for decision-making:

- Vehicle locations l_t: Current location of each vehicle (depot or a help center).
- Vehicle capacities C_t: A matrix representing the remaining capacity of each vehicle $m \in \mathcal{V}$ for each demand type $k \in \mathcal{K}$. This captures vehicle heterogeneity in terms of what they can carry and how much.
- Remaining demands D_t: A matrix indicating the unfulfilled demand of type k at each help center $n \in \mathcal{H}$.
- Visit flags V_t: Binary indicators to track which help centers have been fully serviced or which vehicle-hub pairings have occurred, used with action masking to prevent redundant actions.

Action Space $\mathcal{A}$: The action $a_t \in \mathcal{A}$ is complex and hierarchical. A full action involves selecting a help center n to visit, a specific vehicle m to dispatch, and a particular demand type k that vehicle m will service at n. An action can also be to return a vehicle to the central depot H_0.

$$a_t = \begin{cases} (n, m, k) & \text{Send vehicle } m \text{ to hub } n \text{ for demand type } k \\ (H_0, m, \emptyset) & \text{Vehicle } m \text{ returns to depot} \end{cases} \tag{1}$$

This structure naturally lends itself to a hierarchical decision process, which is a core feature of our proposed decoder.

Transition Dynamics $P(s_{t+1}|s_t, a_t)$: These govern how the state evolves. Vehicle locations change based on the chosen n. Capacities C_t decrease, and demands D_t are reduced based on the service provided. Visit flags V_t are updated.

Travel Time Stochasticity: Travel time τ_{ij} between locations i and j is stochastic. We model it based on Euclidean distance $||p_i - p_j||_2$ modulated by a random variable $\xi \sim \mathcal{N}(\mu_t, \sigma_t^2)$, where μ_t and σ_t^2 can themselves be learned or set to reflect environmental conditions.

$$\tau_{ij} = ||p_i - p_j||_2 \cdot \xi \tag{2}$$

Reward Function $R(s_t, a_t, s_{t+1})$: The reward r_t guides the learning process. It's designed to minimize operational costs and prioritize urgent demands:

$$r_t = -\underbrace{(\tau_{action} \cdot c_m)}_{\text{Travel cost}} - \underbrace{(\omega_k \cdot \phi_{m,k} \cdot \delta_{m,k})}_{\text{Delivery cost/penalty}} \tag{3}$$

where τ_{action} is the travel time for the action, c_m is the travel cost coefficient for vehicle m, ω_k is the urgency weight for demand type k, $\phi_{m,k}$ is the cost rate for vehicle m delivering demand type k, and $\delta_{m,k}$ is the amount delivered. The negative sign indicates we are minimizing costs (maximizing reward).

The formulation satisfies the Markov property as the current state s_t contains all necessary information for future decisions, and transitions/rewards depend only on s_t and a_t.

3.2 Hierarchical Policy Optimization Model

HARL utilizes an encoder-decoder architecture within an actor-critic REINFORCE framework

3.2.1 Encoder Architecture. The encoder processes static and dynamic state information. Static inputs, such as help center locations p_i and their initial demand profiles d_i, are first projected into high-dimensional embeddings using an MLP: $h_i^0 = \text{MLP}([p_i; d_i])$. These initial node embeddings $H^0 = \{h_0^0, h_1^0, \ldots, h_N^0\}$ are then fed into a multi-head self-attention (MHSA) layer [12]. The MHSA layer allows the model to learn rich contextual representations $H = \text{MHSA}(H^0)$ by capturing dependencies and relationships among all help centers. This provides a global understanding of the problem instance. Dynamic information (current vehicle locations, capacities) is separately embedded and combined with these static embeddings at the decoder stage.

3.2.2 Hierarchical Decoder. A core novelty of HARL is its hierarchical decoder, which breaks down the complex action selection into two sequential levels, making the learning process more tractable and the policy more interpretable.

(1) **Level 1 - Hub Selection:** The first level decides which help center n (or the depot H_0) to target next. Given the global context H from the encoder and the current dynamic state embedding h_{dyn} (representing current vehicle locations, time, etc.), an attention mechanism computes attention scores α_n over all possible destination nodes:

$$\alpha_n = \text{Softmax}\left(\frac{(W_q h_{dyn})^T (W_k h_n)}{\sqrt{d_{key}}}\right) \tag{4}$$

where $h_n \in H$ is the embedding of hub n, and W_q, W_k are learned weight matrices. This produces a probability distribution over hubs.

(2) **Level 2 - Vehicle and Resource Allocation:** Once a hub n^* is selected (e.g., by sampling from α_n), the second level determines which available vehicle m should transport which type of required demand k to n^*. This decision is conditioned on h_{n^*}, h_{dyn} (including current vehicle capacities $C_t^{m,k}$ and demand $D_t^{n^*,k}$). An MLP followed by a softmax layer computes a joint probability distribution $\beta_{m,k}^{n^*}$ over valid (vehicle,

demand type) pairs for the selected hub n^*:

$$\beta^{n^*}_{m,k} = \text{Softmax}(\text{MLP}([h_{n^*}; h_{dyn}; C^{m,k}_t; D^{n^*,k}_t])) \quad (5)$$

The overall action probability for $a_t = (n^*, m^*, k^*)$ is then $\pi(a_t|s_t) = \alpha_{n^*} \cdot \beta^{n^*}_{m^*,k^*}$. This factorization manages the combinatorial complexity of the action space.

3.2.3 Dilated Temporal Convolutions for Sequential Context. To effectively model long-range dependencies and sequential patterns in the sequence of decisions (e.g., the order in which hubs are visited), we incorporate dilated temporal convolutions [9] within the decoder, particularly influencing the hub selection stage. Standard RNNs can struggle with long sequences, while full Transformers can be computationally expensive for very long horizons. Dilated convolutions offer an efficient alternative by applying filters over an input sequence with exponentially increasing dilation rates across layers. This allows the network to have a large receptive field with relatively few layers and parameters, capturing temporal context from distant past decisions without explicit recurrent connections. These convolutions operate on sequences of hub embeddings generated during the rollout or features derived from the history of actions, enhancing the model's ability to make contextually aware, far-sighted decisions.

3.2.4 Actor-Critic Policy Optimization. We employ an actor-critic architecture based on the REINFORCE algorithm to train the policy network $\pi_\theta(a_t|s_t)$.

- **Actor (π_θ):** The actor is the policy network described above, including the encoder and hierarchical decoder. It outputs a probability distribution over actions.
- **Critic ($V_\phi(s_t)$):** The critic is a separate neural network that estimates the state-value function $V(s_t)$, which is the expected cumulative discounted reward from state s_t. It helps reduce the variance of the policy gradient estimates. The critic typically processes the same encoded state representations as the actor.

The policy parameters θ are updated using the policy gradient theorem:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) A(s_t, a_t)\right] \quad (6)$$

where $A(s_t, a_t) = R_t - V_\phi(s_t)$ is the advantage function. $R_t = \sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t'}$ is the discounted cumulative return. The critic parameters ϕ are updated by minimizing a loss function, typically the mean squared error between $V_\phi(s_t)$ and R_t.

3.2.5 Action Masking and Constraint Handling. To ensure that the policy only selects feasible actions, an action masking mechanism is applied. For example, a vehicle cannot be dispatched if it lacks the capacity for a specific demand, or a hub that has all its demands satisfied might be masked out. This is typically done by adding large negative values to the logits of invalid actions before the final softmax activation, effectively assigning them zero probability. Stochastic travel times are handled by sampling from their distributions during simulation for training and evaluation. Dynamic capacity updates are part of the environment's transition logic.

Table 1: Inference Reward (R) and Time (T, secs) for HARL vs. Best Baselines (selected scenarios).

Model	(20,3,3)		(30,5,5)		(50,7,5)	
	R	T	R	T	R	T
A3C (RL)	-0.619	0.46	-0.732	0.93	-0.632	0.80
POMO (RL)	-0.391	1.32	-0.720	2.12	-0.583	1.88
LKH-3 (Heur.)	-0.278	783	-0.628	943	-0.733	1546
HARL (Ours)	**-0.163**	**0.14**	**-0.263**	**0.19**	**-0.326**	**0.32**

4 Experiments

We evaluated HARL on synthetic VRP instances with varying numbers of help centers (20 to 50), vehicle types (3 to 7), and demand types (3 to 7), denoted as (hubs, vehicles, demands). Stochastic travel times were used throughout. We focus on two key experiments and a real-world case study. All models were trained for 10,000 epochs.

4.1 Comparison with Baselines

HARL was compared against several policy-based RL baselines including A3C, POMO [7]), and traditional heuristics (LKH-3 (simplified). For LKH-3 we used a rollout mechanism for stochastic travel times compatibility.

Table 1 summarizes inference-time reward (higher is better, values are typically negative costs) and response time (seconds). HARL consistently achieves superior rewards with significantly faster response times, crucial for real-time applications. For instance, in the (50,5,5) scenario, HARL achieved a reward of -0.301 in 0.351s, while the best RL baseline (POMO) scored -0.429 in 3.21s, and the best heuristic (LKH-3) scored -0.418 but took 1342s. This highlights HARL's strong balance of solution quality and speed.

4.2 Ablation Study

We analyzed the impact of the hierarchical decoder (H) and dilated temporal convolutions (D). Configurations: (H+, D+) is the full model; (H+, D-) removes dilated convolutions (uses linear layers); (H-, D+) uses a flat decoder; (H-, D-) removes both.

Table 2 shows inference-time reward and Consecutive Node Actions (CNA), where higher CNA indicates better spatial coherence in routing. The full model (H+, D+) consistently yields the best rewards and higher CNA values across scenarios. For example, in (30,7,7), (H+, D+) achieved -0.289 reward and 92 CNA, while (H-, D-) scored -1.31 reward and 36 CNA. This confirms that both the hierarchical structure and dilated convolutions are crucial for HARL's performance and ability to generate spatially coherent plans.

4.3 Case Study: Wildfire Relief Logistics

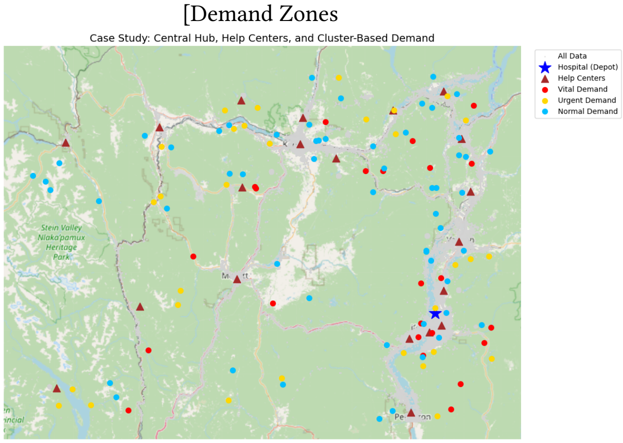
To demonstrate real-world applicability, we adapted HARL to a scenario inspired by the 2021 Kelowna Wildfire, Canada, using anonymized location-based movement data to model demand zones (20 help centers, 3 vehicle types, 3 demand types). Figure 2 illustrates the setup and the optimized policy by HARL.

HARL achieved an average reward of -0.37, outperforming other RL methods (e.g., POMO: -0.54, A3C: -0.86). This underscores its

Table 2: Ablation: Reward (R) and CNA for different model configurations (selected scenarios).

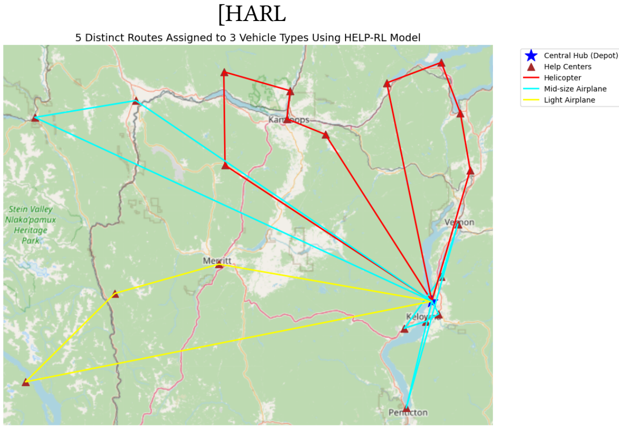
2*Config.	(20,5,5)		(30,7,7)		(50,5,7)	
	R	CNA	R	CNA	R	CNA
H+, D+ (Full)	-0.211	44	-0.289	92	-0.429	113
H+, D-	-0.423	37	-0.351	81	-0.832	92
H-, D+	-0.366	32	-0.311	61	-0.855	104
H-, D-	-1.42	17	-1.310	36	-2.790	42

[Demand Zones



Hubs]

[HARL



Policy]

Figure 2: Kelowna Wildfire case study: (a) Derived help centers and demand types. (b) Optimized routing policy by HARL.

effectiveness in handling complex, spatio-temporally evolving demands in a realistic setting, further supporting its potential as a general tool for dynamic logistics optimization.

5 Conclusion

This paper introduced HARL, a novel hierarchical reinforcement learning framework for real-time policy optimization in complex logistics networks characterized by heterogeneous resources, diverse demands, and stochastic travel times. Key innovations include a hierarchical decoder for structured action selection and the use of dilated temporal convolutions for capturing sequential dependencies in traversed routing policy.

Comprehensive experiments, including comparisons with state-of-the-art RL and heuristic methods, and an ablation study, demonstrated HARL's superior performance in terms of both solution quality and computational speed. A real-world case study based on wildfire relief logistics further validated its applicability. HARL offers a robust and generalizable approach for AI-driven optimization in dynamic supply chain and VRP research, capable of providing high-quality, real-time decisions under uncertainty. Future work will explore extensions to multi-objective optimization and richer real-time information fusion.

References

[1] Abbas Afshar and Ali Haghani. 2012. Modeling integrated supply chain logistics in real-time large-scale disaster relief operations. *Socio-Economic Planning Sciences* 46, 4 (2012), 327–338. doi:10.1016/j.seps.2011.12.003 Special Issue: Disaster Planning and Logistics: Part 2.

[2] Defense Logistics Agency. 2005. Special Report: Hurricane Katrina – The Defense Department Looks Back. https://www.defense.gov/News/News-Stories/Article/Article/615297/special-report-hurricane-katrina-the-defense-department-looks-back/

[3] Djamel Berkoune, Jacques Renaud, Monia Rekik, and Angel Ruiz. 2012. Transportation in disaster response operations. *Socio-Economic Planning Sciences* 46, 1 (2012), 23–32. doi:10.1016/j.seps.2011.05.002 Special Issue: Disaster Planning and Logistics: Part 1.

[4] M.E. Bruni, P. Beraldi, and S. Khodaparasti. 2018. A fast heuristic for routing in post-disaster humanitarian relief logistics. *Transportation Research Procedia* 30 (2018), 304–313. doi:10.1016/j.trpro.2018.09.033 EURO Mini Conference on "Advances in Freight Transportation and Logistics".

[5] ULN Global. 2021. Logistics in Disaster Relief and Humanitarian Aid. https://ulnglobal.com/media/article/logistics-in-disaster-relief-and-humanitarian-aid

[6] Wouter Kool, Herke van Hoof, and Max Welling. 2018. Attention, Learn to Solve Routing Problems!. In *International Conference on Learning Representations*. https://api.semanticscholar.org/CorpusID:59608816

[7] Yeong-Dae Kwon, Jinho Choo, Byoungjip Kim, Iljoo Yoon, Youngjune Gwon, and Seungjai Min. 2020. POMO: Policy Optimization with Multiple Optima for Reinforcement Learning. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 21188–21198. https://proceedings.neurips.cc/paper_files/paper/2020/file/f231f2107df69eab0a3862d50018a9b2-Paper.pdf

[8] Convoy of Hope. 2023. Convoy Responds to 2023 Canada Wildfires. https://convoyofhope.org/disaster-services/2023-canada-wildfires/ Accessed: April 24, 2025.

[9] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

[10] Luca Talarico, Frank Meisel, and Kenneth Sörensen. 2015. Ambulance routing for disaster response with patient groups. *Computers and Operations Research* 56 (2015), 120–133. doi:10.1016/j.cor.2014.11.006

[11] Hamid Tikani and Mostafa Setak. 2019. Ambulance routing in disaster response scenario considering different types of ambulances and semi soft time windows. *Journal of Industrial and Systems Engineering* 12, 1 (2019), 95–128.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *CoRR* abs/1706.03762 (2017). arXiv:1706.03762 http://arxiv.org/abs/1706.03762

[13] Sascha Wohlgemuth, Richard Oloruntoba, and Uwe Clausen. 2012. Dynamic vehicle routing with anticipation in disaster relief. *Socio-Economic Planning Sciences* 46, 4 (2012), 261–271. doi:10.1016/j.seps.2012.06.001 Special Issue: Disaster Planning and Logistics: Part 2.