PG-DyRA: Potential-Game Guided Multi-Agent Collaboration for Dynamic Urban Resource Allocation

Yaxin Mei
Beijing Normal University
Beijing, China
meiyax@163.com

Huiling Qin*
Beijing Normal University
Zhuhai, China
orekinana@gmail.com

Chenhao Wang
Beijing Normal University
Zhuhai, China
chenhwang@bnu.edu.cn

Weijia Jia
Beijing Normal University
Zhuhai, China
jiawj@bnu.edu.cn

Yu Zheng
Southwest Jiaotong University
Chengdu, China
Xidian University
Xi'an, China
msyuzheng@outlook.com

Tian Wang
Beijing Normal University
Zhuhai, China
cs_tianwang@163.com

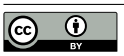
Abstract

Efficient allocation of urban public resources such as transportation, logistics, and emergency response is crucial for modern cities to maintain operational efficiency and service equity. This task requires coordinating multiple mobile units to serve spatiotemporally heterogeneous demands under coupled physical constraints such as finite service capacity, energy budgets, and mobility ranges. Beyond physical limitations, practical deployments also impose information constraints: sensing costs, communication bandwidth, and privacy regulations restrict each agent to observing only nearby peers, yet effective coordination demands reasoning about unobserved actions across the entire system. We propose PG-DyRA, a decentralized coordination framework that operates under partial observability without inter-agent communication. Building on a potential game perspective, we design a marginal contribution utility under which the resulting game admits pure strategy Nash equilibria that coincide with the global optimum, providing a theoretically grounded coordination signal for independent decision-making. To operationalize this signal under this observability constraint, we develop a dual-branch intention inference network that estimates other agents' service capacity allocation from local observations, with one branch modeling visible agents via trajectory-based state bank alignment and the other inferring invisible agents from demand residuals. Experiments on four urban datasets demonstrate that PG-DyRA achieves 2.23-8.04 percentage points higher task coverage rates while reducing energy consumption by 2.99-8.35 percentage points compared to the state-of-the-art methods.

CCS Concepts

• **Computing methodologies** → **Multi-agent planning**; *Cooperation and coordination*; • **Information systems** → Decision support systems.

*Corresponding author.



KDD '26, Jeju Island, Republic of Korea

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3817998

Keywords

Urban Resource Allocation; Potential Game; Physical Constraints; Decentralized Coordination; Urban Computing

ACM Reference Format:
Yaxin Mei, Huiling Qin, Chenhao Wang, Weijia Jia, Yu Zheng, and Tian Wang. 2026. PG-DyRA: Potential-Game Guided Multi-Agent Collaboration for Dynamic Urban Resource Allocation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3770855.3817998

1 Introduction

In large-scale modern urban scenarios, the effective allocation of public resources, e.g. transportation, logistics services, and emergency response, constitutes a critical decision that shapes operational efficiency and service equity [25, 27]. Efficient allocation not only sustains daily civic life and economic activity but also underpins sustainable urban development. However, resource supply and public demand exhibit pronounced spatio-temporal heterogeneity [20]. Coordinating multiple mobile units under this heterogeneity is complicated by coupled physical constraints governing each unit: finite service capacity, movement bounded by speed limits, and energy budgets that risk stranding units far from charging depots [20]. When regional demand exceeds individual service capacity, multiple units must collaborate, yet aggressive repositioning toward demand hotspots accelerates energy depletion. While centralized optimization methods can compute globally optimal assignments under complete information [6], they suffer from exponential computational complexity at urban scale and impose unacceptable data centralization burdens.

Beyond physical limitations, practical deployments impose stringent information constraints. Sensing costs, communication bandwidth, and privacy regulations restrict each agent to observing only nearby peers, creating partial observability where agents cannot access others' intentions [16, 23]. This induces strategic uncertainty: invisible agents may simultaneously target identical regions, causing redundant coverage and wasted service capacity. Multi-Agent Reinforcement Learning (MARL) offers a decentralized alternative, yet existing methods either fail to achieve effective coordination due to unresolved action uncertainty, or compensate through explicit

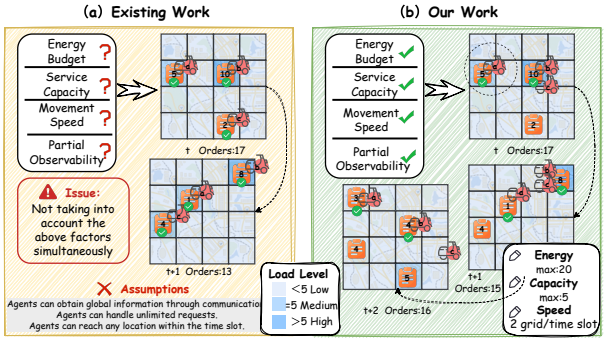


Figure 1: Existing work vs. our work on dynamic urban resource allocation. Existing work typically assumes idealized conditions, while our work is more in line with reality.

inter-agent communication during execution that incurs bandwidth and latency overhead [9, 24]. More critically, these approaches lack theoretical guarantees that decentralized decisions converge to globally optimal allocations. The coupling of physical and information constraints thus remains inadequately addressed: agents must coordinate service capacity allocation without knowing others' intentions, while respecting operational limits that are not fully observable.

To bridge the aforementioned gaps, we introduce PG-DyRA (Potential-Game guided Dynamic urban Resource Allocation), a decentralized coordination framework that operates under physical constraints and partial observability without inter-agent communication, as illustrated in Figure 1. From a potential game perspective, we design a marginal contribution utility under which each agent's local incentive equals its contribution to a global potential function aligned with the system objective. This guarantees the theoretical existence of pure strategy Nash equilibria where globally optimal allocations are attainable through independent utility maximization without centralized coordination or explicit communication.

However, computing marginal contributions requires knowing other agents' service capacity allocation, which is unavailable under partial observability. We address this through a dual-branch intention inference network that predicts system-wide capacity distribution from local observations. One branch models visible agents through trajectory-based state bank alignment, while the other infers invisible agents' intentions from spatio-temporal demand residuals via gated filtering. Since the theoretical guarantee relies on exact capacity knowledge, this inference bridges the gap between the ideal potential game and practical partial observability, enabling agents to approximate marginal contributions and make distributed decisions that respect physical constraints. Training proceeds in two phases, where the pre-training phase uses Nash equilibrium solutions to warm-initialize the inference network, followed by alternating optimization of inference and policy networks to address distribution shift as the policy evolves. The resulting framework enables fully decentralized execution where each agent makes independent decisions that collectively achieve near-optimal system performance. Our main contributions are summarized as follows:

- We formulate dynamic urban resource allocation as a potential game with a marginal contribution utility, under which the global optimum is theoretically realized as a pure strategy Nash equilibrium, turning system-wide optimization into per-agent local incentives.
- We design a dual-branch intention inference network that reconstructs agents' capacity allocation from partial observations, modeling visible agents via trajectory-based state-bank alignment and invisible agents through spatio-temporal demand residuals, bridging the gap between theoretical guarantees and real-world deployability.
- Extensive experiments on various scenarios demonstrate that PG-DyRA outperforms the state-of-the-art baseline methods. It achieves 2.23-8.04 percentage points higher task coverage rates while reducing energy consumption by 2.99-8.35 percentage points.

2 Preliminary

2.1 Definitions

In this paper, we mainly have the following definition:

Definition 1. *(Urban Area). We model the urban area as an $H \times W$ grid world containing $R = H \times W$ service regions $\mathcal{R} = \{1, 2, \ldots, R\}$. Each region $r \in \mathcal{R}$ corresponds to grid coordinates $p(r) = [h_r, w_r] \in \mathbb{Z}^2$, and the Manhattan distance between regions is $d(r_1, r_2) = |h_{r_1} - h_{r_2}| + |w_{r_1} - w_{r_2}|$. The system comprises N mobile resource units (agents) $\mathcal{N} = \{1, 2, \ldots, N\}$ operating over a planning horizon of T time steps.*

Definition 2. *(Tasks). Service demands arrive dynamically as tasks. Each task τ is characterized by a tuple (p_τ, q_τ, t_τ), where p_τ denotes the task position, q_τ represents the task quantity, and t_τ specifies the task arrival timestep. At timestep t, the aggregated uncompleted demand in region r is D_r^t.*

Definition 3. *(Agent State). Agent i's state at timestep t is $\mathbf{s}_i^t = (p_i^t, c_i^t, e_i^t)$, where p_i^t is the agent's position, c_i^t is the agent's service capacity bounded by the maximum capacity $C_{\max}$, and $e_i^t \in [0, E_{\max}]$ is the residual mobile energy. In addition, each agent i has a movement speed of v, which means $d(p_i^t, p_i^{t+1}) \le v$.*

Definition 4. *(Depot). The depot $r_0 \in \mathcal{R}$ is a designated region where all agents are initially based and must return by the final timestep $T + 1$. Agents can recharge their energy at the depot.*

Definition 5. *(Joint Action). At each timestep t, each agent i selects an action $a_i^t \in \mathcal{R}$, where the action specifies a target region. The action $a_i^t = r_0$ corresponds to returning to the depot, and selecting the agent's current region keeps the agent in place (i.e., a Wait action). The joint action is $\mathbf{a}^t = (a_1^t, \ldots, a_N^t)$.*

Definition 6. *(Partial Observability). Agent i observes nearby agents $\mathcal{N}_i^{vis}(t) = \{j \in \mathcal{N} \setminus \{i\} \mid d(p_i^t, p_j^t) \le d_0\}$, where d_0 is the vision range. Invisible agents are $\mathcal{N}_i^{inv}(t) = \mathcal{N} \setminus (\mathcal{N}_i^{vis}(t) \cup \{i\})$. Agent i's partial observation $\mathbf{o}_i^t$ includes its own state $\mathbf{s}_i^t$, global demand distribution $\{D_r^t\}_{r \in \mathcal{R}}$, and visible agents' states $\{\mathbf{s}_j^t\}_{j \in \mathcal{N}_i^{vis}(t)}$.*

2.2 Problem Statement

Given a targeted urban area $\mathcal{R}$ with N mobile resources (agents) and tasks over T time steps, the formulation of this problem is finding

a joint policy $\boldsymbol{\pi}$ that maximizes the system benefit after allocating the resources:

$$\max_{\boldsymbol{\pi}} \mathbb{E}_{\mathbf{a}^{1:T}\sim\boldsymbol{\pi}}\left[\sum_{t=1}^{T}\Phi(\mathbf{a}^t)\right], \tag{1}$$

where the system reward at timestep t is:

$$\Phi(\mathbf{a}^t) = \sum_{r\in\mathcal{R}} \min\left(D_r^t, \sum_{i:a_i^t=r} c_i^t\right), \tag{2}$$

$$\begin{aligned} \text{s.t.}\quad & C1: d(p_i^t, a_i^t) \le v, \quad \forall i \in \mathcal{N}, t \in \{1,\dots,T\}, \\ & C2: e_i^t \ge e(p_i^t, a_i^t), \quad \forall i \in \mathcal{N}, t \in \{1,\dots,T\}, \\ & C3: a_i^t \sim \pi_i(\cdot \mid \mathbf{o}_i^t), \quad \forall i \in \mathcal{N}, t \in \{1,\dots,T\}, \\ & C4: p_i^0 = p_i^{T+1} = r_0, \quad \forall i \in \mathcal{N}. \end{aligned} \tag{3}$$

Constraint $C1$ enforces motion range limitations. When agent i selects a region action, the target region must be within Manhattan distance v from its current position. Constraint $C2$ ensures energy feasibility. Agent i must have sufficient residual energy to reach the target region, where $e(p_i^t, a_i^t)$ represents the energy consumption from p_i^t to a_i^t. Constraint $C3$ enforces decentralized execution. Each agent independently samples actions from its local policy conditioned on partial observations, without centralized control or inter-agent communication. Constraints $C4$ specifies boundary conditions. All agents start from the depot r_0 at $t = 0$ and must return by the final timestep $t = T + 1$.

Theorem 1. *The dynamic resource allocation problem is NP-hard.*

Due to space limitations, we provide the proof in Appendix A. The NP-hardness motivates our learning-based approach in the next section, in which we reformulate the problem as a potential game to enable scalable decentralized solutions.

3 Game-Theoretic Framework for Dynamic Resource Allocation

Since dynamic resource allocation is NP-hard (THEOREM 1), traditional optimization methods, such as integer programming [32] and dynamic programming [5], are impractical for large-scale deployment. To circumvent this intractability, we adopt a potential game perspective and design a marginal contribution utility that aligns each agent's local incentive with the global system objective. This alignment provides a theoretical foundation for scalable, fully decentralized coordination with guaranteed convergence to Nash equilibria.

We model the resource allocation at timestep t as a game and design individual utilities that align with the system objective.

Definition 7. *(Game Model). The resource allocation game at timestep t is defined as:*

$$\mathcal{G}^t = (\mathcal{N}, \{\mathcal{A}_i^t\}_{i\in\mathcal{N}}, \{u_i^t\}_{i\in\mathcal{N}}), \tag{4}$$

where $\mathcal{N}$ is the player set, $\mathcal{A}_i^t$ is agent i's strategy space, and $u_i^t : \mathcal{A}_i^t \times \mathcal{A}_{-i}^t \to \mathbb{R}$ is agent i's utility function.

Definition 8. *(Strategy Space). Agent i's strategy space at time step t consists of feasible actions satisfying physical constraints:*

$$\begin{aligned} \mathcal{A}_i^t = \{a \in \mathcal{R} \mid\ & d(p_i^t, a) \le v, \lceil d(a, r_0)/v \rceil \le T - t, \\ & e_i^t \ge e(p_i^t, a) + e(a, r_0)\}. \end{aligned} \tag{5}$$

The three constraints jointly enforce single-step reachability of the target, sufficient remaining horizon to return to the depot, and sufficient energy for the round trip.

Definition 9. *(Marginal Contribution Utility). Agent i's utility for selecting action a_i^t given other agents' joint action $\mathbf{a}_{-i}^t$ is:*

$$u_i(a_i^t, \mathbf{a}_{-i}^t) = \min(D_r^t, C_r^{-i} + c_i^t) - \min(D_r^t, C_r^{-i}), \tag{6}$$

where $r = a_i^t$ is the target region selected by agent i, and $C_r^{-i} = \sum_{j\neq i: a_j^t = r} c_j^t$ denotes the total capacity allocated to region r by other agents.

This utility function measures agent i's marginal contribution, namely the additional task demand satisfied by selecting a_i^t given others' allocations. Specifically, if $C_r^{-i} \ge D_r^t$, then $u_i^t = 0$ (we write $u_i^t = u_i(a_i^t, \mathbf{a}_{-i}^t)$ for readability, placing the timestep on the symbol when arguments are omitted), and if $C_r^{-i} < D_r^t$, then $u_i^t = \min(c_i^t, D_r^t - C_r^{-i})$.

Theorem 2. *For any agent $i \in \mathcal{N}$, any joint action $\mathbf{a}_{-i}^t$ of other agents, and any two feasible actions $a_i^t, a'^t_i \in \mathcal{A}_i^t$, the change in agent i's marginal contribution utility equals the corresponding change in the system objective:*

$$u_i(a_i^t, \mathbf{a}_{-i}^t) - u_i(a'^t_i, \mathbf{a}_{-i}^t) = \Phi(a_i^t, \mathbf{a}_{-i}^t) - \Phi(a'^t_i, \mathbf{a}_{-i}^t). \tag{7}$$

Due to space limitations, we provide the proof in Appendix A.

THEOREM 2 establishes that any unilateral change in an agent's utility produces an identical change in Φ^t, so maximizing the individual utility u_i^t is equivalent, at every unilateral move, to maximizing the system objective. This lays the foundation for constructing a potential game.

Definition 10. *(Potential Game). A game $\mathcal{G}^t$ qualifies as a potential game if there exists a potential function $\Psi^t : \mathcal{A}^t \to \mathbb{R}$ satisfying the following condition: for any agent $i \in \mathcal{N}$, any joint strategy of other agents $\mathbf{a}_{-i}^t \in \mathcal{A}_{-i}^t$, and any two feasible strategies $a_i^t, a'^t_i \in \mathcal{A}_i^t$:*

$$u_i(a_i^t, \mathbf{a}_{-i}^t) - u_i(a'^t_i, \mathbf{a}_{-i}^t) = \Psi(a_i^t, \mathbf{a}_{-i}^t) - \Psi(a'^t_i, \mathbf{a}_{-i}^t). \tag{8}$$

This property ensures that the game admits a global coordination structure. All agents pursuing individual utility improvements collectively ascend the same potential function, making decentralized best-response updates a viable path toward system-wide optimality.

Theorem 3. *The resource allocation game $\mathcal{G}^t$ is a potential game with potential function $\Psi(\mathbf{a}^t) = \Phi(\mathbf{a}^t)$.*

Corollary 1. *The game $\mathcal{G}^t$ possesses at least one pure strategy Nash equilibrium, and the global optimum of the system objective $\Phi(\mathbf{a}^t)$ is attained at a pure strategy Nash equilibrium.*

THEOREM 3 follows directly from THEOREM 2 and DEFINITION 10. Due to space limitations, we provide the proof of COROLLARY 1 in Appendix A.

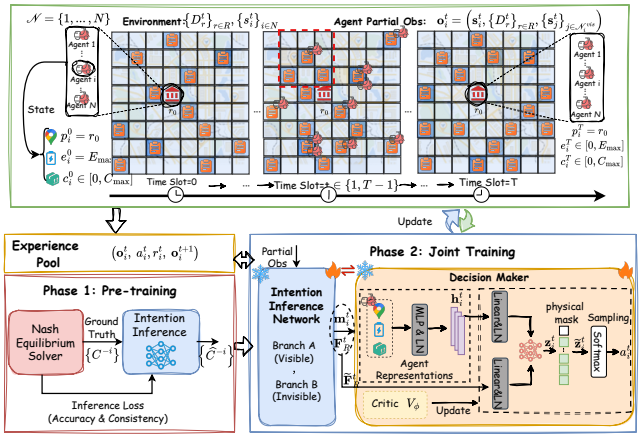


Figure 2: Overall framework of PG-DyRA.

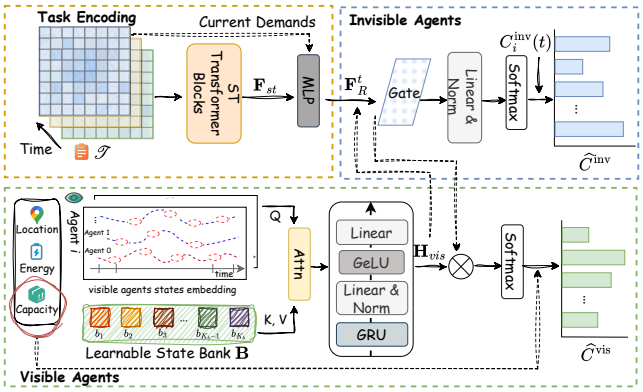


Figure 3: Details of Dual-Branch Intention Inference Network.

Remark 1. *(Scope of the Potential Property). THEOREM 2 rests on two structural conditions of our setting: (i) the system objective* $\Phi(\mathbf{a}^t)$ *decomposes additively over regions, which holds naturally for resource-allocation problems where regional service depends on local capacity; and (ii) each agent's feasible action set in DEFINITION 8 depends only on its own state, so physical constraints introduce no cross-agent coupling. Both conditions are intrinsic to our problem setting and ensure* $\mathcal{G}^t$ *qualifies as a potential game (THEOREM 3) at every timestep.*

The potential game structure provides two key properties that underpin our framework. First, convergence is guaranteed: any sequence of better response updates terminate at a pure-strategy Nash equilibrium in finite steps, since each improvement strictly increases the potential function over a finite action space [15]. Second, the global optimum is itself a pure-strategy Nash equilibrium (COROLLARY 1), so agents can in principle reach globally optimal resource allocation through independent utility maximization without centralized coordination or inter-agent communication. The remaining challenge is that computing marginal contributions requires knowing other agents' capacity allocation, which is unavailable under partial observability. This motivates the inference network designed in the next section.

4 PG-DyRA Architecture

We propose PG-DyRA, a decentralized coordination framework that operationalizes potential game principles under partial observability (Figure 2). The core idea is to infer the unknown service capacity allocation, enabling each agent to approximate its marginal contribution and make decisions that implicitly coordinate toward a pure-strategy Nash equilibrium. The framework comprises three components: (i) a dual-branch intention inference network that predicts system-wide capacity distribution from partial observations, separately modeling visible agents via trajectory-based state-bank alignment and invisible agents via spatio-temporal demand residual analysis; (ii) a decision maker that computes approximate marginal contributions from inferred capacity allocation and optimizes policies through actor-critic learning with potential-game-based rewards; and (iii) a two-phase training strategy where Nash equilibrium solutions pre-train the inference network, followed by alternating optimization to address distribution shift while preserving alignment with the potential game structure. This enables fully decentralized execution where agents independently maximize utilities that, by the potential game's construction, collectively achieve near-optimal performance.

4.1 Dual-Branch Intention Inference

Computing the marginal contribution utility requires knowing the capacity allocation $\{C_r^{-i}\}_{r\in\mathcal{R}}$ of all other agents, which is unavailable under partial observability. We design a dual-branch intention inference network that estimates $\{\hat{C}_r^{-i}\}_{r\in\mathcal{R}}$ from agent i's local observations (Figure 3). Visible agents provide trajectory signals from which intentions can be directly modeled, while invisible agents' contributions are inferred from their residual imprint on regional demand patterns. This motivates our architecture: a trajectory-based branch for visible agents, a demand-residual branch for invisible agents, and a shared task encoding that provides spatio-temporal features for both.

4.1.1 Task Encoding. Both inference branches need to reason about where agents' service capacity is likely allocated, which closely relates to the spatial distribution and temporal evolution of demands. Task encoding extracts these shared features into region-level representations that serve as the common foundation for both branches.

We organize the historical task information into a tensor, denoted as $\mathcal{T} \in \mathbb{R}^{T'\times H\times W\times C_{\text{in}}}$, where T' is the history window length and C_{in} channels respectively encode per-region statistics, such as the outstanding demand, the arrival of new tasks, and completion rates. A transformer-based spatiotemporal encoder processes $\mathcal{T}$ to model spatial correlations across regions and temporal dependencies across the history window, producing region-level features $\mathbf{F}_{\text{st}} \in \mathbb{R}^{(H\times W)\times d}$. For each grid r, we further fuse this aggregated context with the signal from the current timestep:

$$\mathbf{f}_r^t = \text{MLP}\big([\mathbf{F}_{\text{st}}[h_r W + w_r, :];\ \text{Embed}(\mathcal{T}(t, r))]\big) \in \mathbb{R}^d. \quad (9)$$

Stacking across all grids yields $\mathbf{F}_R^t \in \mathbb{R}^{R\times d}$. The visible branch uses these representations as reference keys for deriving each visible

agent's action probability distribution over regions, while the invisible branch applies a learned gating mask to filter out the observable agents' contribution, retaining the residual signal attributable to invisible agents.

4.1.2 Visible Agent Capacity Estimation. For agents within the observation range $\mathcal{N}_i^{\text{vis}}(t)$, agent i has access to their states and historical trajectories. The core challenge is to extract intention signals from irregular-length observation sequences and aggregate them into expected capacity allocation despite uncertainty over future actions.

Each visible agent j has an observation sequence $\{(t_j^k, \mathbf{s}_j^{t_k})\}_{k=1}^{L_j}$ of varying length L_j, where $\mathbf{s}_j^{t_k}$ is agent j's state at the observed timestep t_j^k. To project heterogeneous trajectory observations into a shared behavioral space, we introduce a learnable state bank $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_{K_b}] \in \mathbb{R}^{K_b \times d}$ encoding prototypical trajectory patterns. Each observation pair is embedded into a d-dimensional token, yielding $\mathbf{E}_j \in \mathbb{R}^{L_j \times d}$, and cross-attention aligns it with these prototypes:

$$\mathbf{Z}_j = \text{Attn}(\mathbf{E}_j, \mathbf{B}, \mathbf{B}). \tag{10}$$

Each timestep is thereby projected into a shared semantic space defined by the prototype patterns, so that semantically similar movements map to similar regions of this space. A GRU then processes the aligned sequence, and its final hidden state serves as the fixed-dimensional representation $\mathbf{h}_j$. Stacking across all $M = |\mathcal{N}_i^{\text{vis}}(t)|$ visible agents yields $\mathbf{H}_{\text{vis}} = [\mathbf{h}_1, \ldots, \mathbf{h}_M]^\top \in \mathbb{R}^{M \times d}$.

To convert these representations into capacity estimates, we score each agent–region pair by the compatibility between the agent's representation and the region's task feature, applying Softmax to obtain each agent's action distribution over regions:

$$\hat{\mathbf{C}}^{\text{vis}} = \text{Softmax}\left(\frac{\mathbf{H}_{\text{vis}} \mathbf{F}_R^{t\top}}{\sqrt{d}}\right)^\top \mathbf{c}^t, \tag{11}$$

where $\mathbf{c}^t = [c_1^t, \ldots, c_M^t]^\top$ contains visible agents' service capacities. The resulting $\hat{\mathbf{C}}^{\text{vis}}$ is the expected capacity that visible agents allocate across regions $\mathcal{R}$, computed as the sum of their capacities weighted by the predicted action probabilities. This expectation-based aggregation yields a smooth allocation estimate that remains stable under the action uncertainty of individual agents.

4.1.3 Invisible Agent Capacity Inference. Agents outside the observation range contribute to the same global capacity allocation as visible ones, and must therefore be incorporated into the inference. For these invisible agents, the current position is unobserved, so the tight bound on reachable regions provided by a known location and speed limit is no longer available. Even when historical observations exist, the uncertainty over current position accumulates over time and quickly renders inference unreliable. We instead infer their aggregate capacity distribution by exploiting two indirect sources: (i) the total invisible capacity $C_i^{\text{inv}}(t)$, obtained by subtracting observed capacities from the total system capacity, which provides the budget of capacity to be allocated, and (ii) the task features $\mathbf{F}_R^t$. Since task features encode the aggregate service context across regions, suppressing the component explained by observed agents yields a residual representation that, together with the total invisible-capacity budget, parameterizes the allocation distribution of invisible agents over regions.

Concretely, we use the task representations $\mathbf{F}_R^t$ to capture local demand context at each region, and the mean visible representation $\bar{\mathbf{h}}_{\text{vis}}$ to summarize the global state of observed agents. Concatenating these two signals, an FFN estimates a per-region observable-influence score:

$$\mathbf{G} = \sigma(\text{FFN}([\mathbf{F}_R^t; \mathbf{1}_R \bar{\mathbf{h}}_{\text{vis}}^\top])), \tag{12}$$

where σ is the sigmoid activation. The residual $\mathbf{F}_R^{\prime t} = \mathbf{F}_R^t \odot (\mathbf{1} - \mathbf{G})$ suppresses features attributable to observed agents while preserving signals driven by invisible ones.

The residual features are projected into per-region logits, normalized into a probability distribution over regions, and scaled by the total invisible capacity to produce the predicted allocation

$$\hat{\mathbf{C}}^{\text{inv}} = \text{Softmax}(\text{Linear}(\text{LN}(\mathbf{F}_R^{\prime t}))) \cdot C_i^{\text{inv}}(t). \tag{13}$$

Since the visible and invisible branches cover disjoint sets of agents, their predicted allocations sum directly to recover the total capacity contributed by all other agents: $\hat{\mathbf{C}}^{-i} = \hat{\mathbf{C}}^{\text{vis}} + \hat{\mathbf{C}}^{\text{inv}}$. The inference network is optimized via $\mathcal{L}_{\text{inf}} = \mathbb{E}_{i,t}[\mathcal{L}_{\text{cap}} + \lambda_{\text{rank}} \mathcal{L}_{\text{rank}}]$, where $\mathcal{L}_{\text{cap}} = \frac{1}{R} \sum_r (\hat{C}_r^{-i} - C_r^{-i})^2$ measures capacity estimation accuracy. The ranking loss $\mathcal{L}_{\text{rank}}$ additionally enforces consistency in the relative ordering of regions, which more directly aligns with downstream marginal contribution comparisons that depend on relative rather than absolute capacity values (detailed explanation in Appendix B).

4.2 Decision Maker

The Decision Maker translates each agent's local information into concrete action selections, choosing which region to serve at each timestep while respecting physical constraints. Its central task is to leverage the inferred capacity allocation to compute approximate marginal contributions, and to learn a policy that maximizes these contributions, thereby approaching the potential game's coordination guarantee in a fully decentralized manner.

With the inferred $\hat{\mathbf{C}}^{-i}$ from the dual-branch inference network, each agent can approximate its marginal contribution to every region according to the potential game formulation. Agent i constructs augmented observation $\tilde{\mathbf{o}}_i^t = [\mathbf{o}_i^t, \hat{\mathbf{C}}^{-i}]$ and computes the estimated marginal contribution vector $\mathbf{m}_i^t = (m_i^{1,t}, \ldots, m_i^{R,t})^\top$, where $m_i^{r,t} = \min(D_r^t,\ \hat{C}_r^{-i} + c_i^t) - \min(D_r^t,\ \hat{C}_r^{-i})$ approximates the true marginal contribution $u_i(r, \mathbf{a}_{-i}^t)$ by substituting the inferred $\hat{C}_r^{-i}$ for the unobservable C_r^{-i}. By THEOREM 3, when the inference is accurate, maximizing these estimated marginal contributions aligns with optimizing the global objective.

The marginal contributions gate regional task representations through a Sigmoid modulation, letting the agent emphasize regions where its contribution is most valuable

$$\boldsymbol{\alpha} = \text{Sigmoid}(\mathbf{W}_\alpha \mathbf{m}_i^t + \mathbf{b}_\alpha), \quad \tilde{\mathbf{F}}_R^t = (\boldsymbol{\alpha} \mathbf{1}_d^\top) \odot \mathbf{F}_R^t. \tag{14}$$

The Actor network produces action logits $\mathbf{z}_i^t$ from agent i's state embedding and modulated representations $\tilde{\mathbf{F}}_R^t$. Physical constraints are enforced through action masking on $\mathbf{z}_i^t$, yielding masked logits $\tilde{\mathbf{z}}_i^t$ and the final policy $\pi_\theta(a_i^t | \tilde{\mathbf{o}}_i^t) = \text{Softmax}(\tilde{\mathbf{z}}_i^t)$. During training, actions are sampled from this distribution to maintain exploration;

during evaluation, the highest probability action is selected deterministically. The critic network V_ϕ estimates state values, and training proceeds via standard Actor-Critic updates with GAE-based advantage estimation.

Individual rewards implement the potential game utility with movement costs:

$$r_i^t = u_i(a_i^t, \mathbf{a}_{-i}^t) - \beta \cdot d(p_i^t, a_i^t), \tag{15}$$

where the first term is the marginal contribution to task coverage computed by the environment using the actual joint action, and the second penalizes movement distance. The movement cost does not disrupt the potential game structure (proof in Appendix A), encouraging energy-efficient behavior.

4.3 Two-Phase Training

Training the inference network and the policy jointly from scratch faces a circular dependency: the policy relies on accurate capacity predictions to compute meaningful marginal contributions, while the inference network requires trajectories generated by a reasonable policy to learn informative patterns. Random initialization of both components leads to a negative feedback loop where poor predictions produce poor policies, which in turn generate uninformative trajectories for inference learning. We resolve this through a two-phase training strategy that breaks this cycle.

Phase 1: Pre-training. We first pre-train the inference network using Nash equilibrium solutions computed by maximizing the potential function $\Phi(\mathbf{a}^t)$ in small-scale offline instances. These optimal solutions provide high-quality supervision that anchors the inference network near optimal coordination patterns, giving the subsequent policy learning a reliable starting point for capacity estimation.

Phase 2: Joint training. Once the inference network provides reasonable capacity predictions, we train the policy and inference networks through alternating optimization. Each iteration consists of two stages:

(i) *Policy update*: with inference network parameters ψ frozen, agents collect trajectories using augmented observations $\tilde{\mathbf{o}}_i^t|_\psi$ and update Actor-Critic parameters (θ, ϕ) via policy gradient.

(ii) *Inference update*: with policy parameters frozen, ground truth actions and capacity allocations are extracted from collected trajectories to update ψ via supervised learning.

This alternating scheme addresses the distribution shift that arises as the policy evolves: the inference network continuously adapts to the current policy's trajectory distribution, while decoupled updates prevent interference between simultaneously changing components.

5 Experiments

This section presents a comprehensive empirical evaluation of the proposed system. The evaluation begins with an introduction to the real-world datasets and experimental setup, followed by systematic comparisons across different parameter configurations and deployment strategies, and concludes with case studies demonstrating practical effectiveness.

5.1 Experimental Settings

Datasets. To thoroughly evaluate the effectiveness of PG-DyRA, we conduct a series of experiments on three real-world datasets (Happy Valley [20], NYCTaxi [29] and DHRD-SE [2]) and one synthetic data SYN. These datasets cover several representative aspects of urban life and production, including infrastructure, transportation and food delivery. Statistical information is summarized in Appendix C. It includes the scenario size, the average number of demands per day, and the data collection interval time, etc. Due to the differences in size and requirements of each scenario, we make different initial settings for the agents, including the agent's service capabilities, initial energy per day, movement speed and vision range.

Baselines. We compare various baseline methods to evaluate the performance of our method. These baselines can be categorized into four groups: (i) Rule-based, include Greedy and EADS [20]; (ii) MARL Algorithms, include QMIX [19], IA2C, IPPO and MAPPO [36]; (iii) Communication-based MARL, include Tom2C [31] and HIT-MAC [33]; (iv) Structure-based Cooperative MARL, include DyPS [27] and CoopRide [28].

Metrics. We mainly use two indicators to evaluate our method: (i) Task Coverage Rate (TCR): TCR is the percentage of daily tasks that have been completed; (ii) Energy Consumption Rate (ECR): ECR describes the percentage of energy consumption of mobile resources (agents) during the process of completing tasks each day.

5.2 Overall Performance

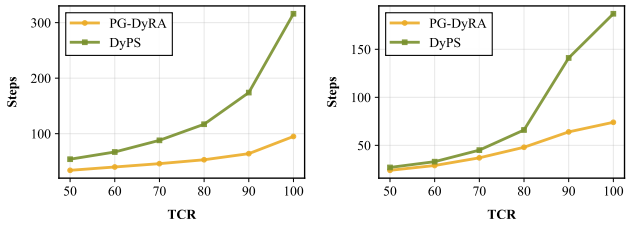
Table 1 presents the performance comparison across four datasets on TCR and ECR when TCR first reaches 50% (ECR@50). Our method consistently achieves the best performance on both metrics, attaining the highest TCR while maintaining the lowest ECR@50 across all datasets. We outperform the second-best cooperative baseline by 2.23-8.04 percentage points in TCR and 2.99-8.35 percentage points in energy efficiency. Notably, ECR@50 varies significantly across datasets, reflecting inherent difficulty differences. For example, HappyValley's sparse demand distribution and large grid (51×108) require substantially more energy to achieve initial coverage, while NYCTaxi's dense demand and compact grid (10×20) enable faster service ramp-up with lower energy overhead. This validates that our method adapts effectively to varying environmental conditions while consistently optimizing energy utilization through marginal contribution-based utilities.

Critically, ECR@50 reveals fundamental algorithmic efficiency differences. CoopRide achieves competitive TCR but requires 5.70-8.35 percentage points more energy than our method to reach 50% coverage, as its GNN-based global cooperation prioritizes coverage maximization without explicit early-stage energy optimization, and agents may undertake long-distance trips for marginal gains. Communication-based methods exhibit lower efficiency, with Tom2C consuming 7.40-11.91 percentage points more energy due to coordination overhead during the initial service phase. DyPS demonstrates balanced performance, ranking second in most datasets through dynamic parameter sharing that implicitly reduces exploration waste. Notably, several methods in HappyValley cannot reach 50% TCR (marked as '-'), highlighting that inefficient coordination prevents achieving minimum viable service levels in challenging scenarios.

Table 1: Performance Comparison on Task Coverage Rate (TCR) (%) and Energy Consumption Rate at 50% TCR (ECR@50) (%)

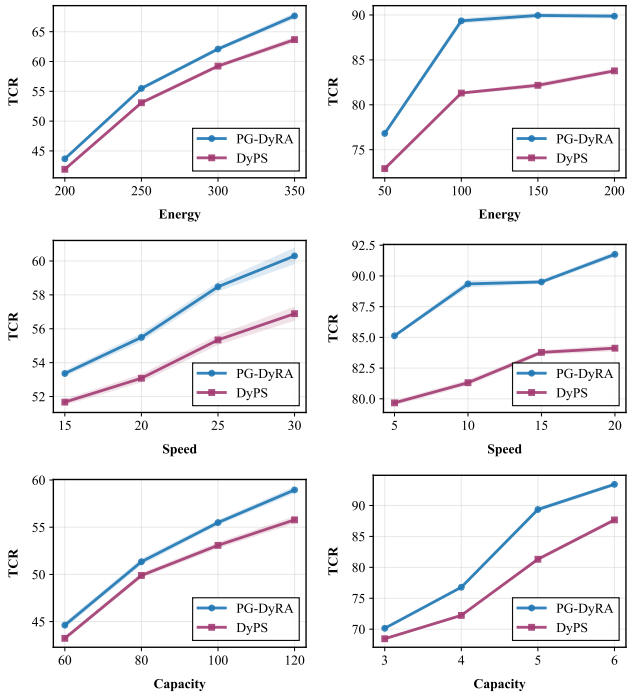
Category	Method	HappyValley		NYCTaxi		DHRD-SE		SYN	
		TCR ↑	ECR@50 ↓	TCR ↑	ECR@50 ↓	TCR ↑	ECR@50 ↓	TCR ↑	ECR@50 ↓
Rule-based	Greedy	23.04±0.00	-	53.91±0.00	82.41±0.00	31.28±0.00	-	38.65±0.00	-
	EADS	49.13±0.27	-	70.88±0.32	63.83±0.26	67.81±0.34	73.92±0.16	55.47±0.21	80.04±0.16
MARL	QMIX	45.73±0.21	-	70.49±0.28	65.12±0.24	65.32±0.31	75.34±0.14	51.52±0.21	84.01±0.13
	IA2C	46.87±0.29	-	73.54±0.23	63.56±0.18	68.72±0.26	73.39±0.13	54.28±0.10	83.44±0.13
	IPPO	47.91±0.26	-	74.63±0.25	61.87±0.21	69.34±0.30	71.94±0.21	54.39±0.16	82.15±0.21
	MAPPO	50.83±0.29	88.66±0.15	77.17±0.19	60.12±0.15	70.48±0.22	70.52±0.15	56.13±0.12	83.56±0.18
Comm. -based	HIT-MAC	49.34±0.34	-	75.87±0.27	64.23±0.22	68.95±0.29	75.12±0.22	55.95±0.19	84.34±0.22
	Tom2C	51.15±0.31	84.45±0.18	79.56±0.24	59.78±0.18	69.89±0.26	72.89±0.18	58.88±0.17	80.16±0.15
Structure -based	DyPS	53.08±0.27	83.16±0.16	81.31±0.18	55.37±0.14	72.23±0.20	65.47±0.14	61.78±0.11	75.48±0.10
	CoopRide	53.26±0.12	80.89±0.15	80.79±0.21	58.22±0.11	71.87±0.25	68.01±0.12	59.41±0.15	77.89±0.11
	PG-DyRA	**55.49**±0.21	**72.54**±0.13	**89.35**±0.24	**52.38**±0.13	**77.84**±0.27	**62.31**±0.22	**65.70**±0.14	**72.03**±0.15

Figure 4 extends this analysis by comparing energy consumption against DyPS from 50% to 100% TCR. PG-DyRA consistently requires less energy at all stages, with the advantage widening at higher coverage rates. This validates our method's effectiveness in simultaneously optimizing task coverage and energy efficiency, both critical for practical deployment where achieving service guarantees quickly impacts operational viability.



(a) Happy Valley (b) SYN

Figure 4: Task Coverage Efficiency Comparison on One-Day Task Horizons (Happy Valley and SYN).



(a) Happy Valley (b) NYCTaxi

Figure 5: The effectiveness and efficiency of PG-DyRA under different agent initialization settings on Happy Valley and NYCTaxi. DyPS is chosen as the competitor, which performs well in the previous experiment.

5.3 Effect of Different Agent Initialization

To evaluate robustness under varying operational constraints, we conduct parameter sensitivity analysis on agent initialization configurations across three parameters (energy budget, movement speed and service capacity). We select DyPS as the comparison baseline due to its strong adaptability across different resource constraints. Figure 5 presents the performance across different parameter settings on HappyValley and NYCTaxi datasets.

Results reveal dataset-specific sensitivity patterns reflecting environmental characteristics. HappyValley's large sparse grid shows highest sensitivity to energy, as agents must travel longer distances to serve dispersed demand. Conversely, NYCTaxi's compact dense environment exhibits greatest sensitivity to capacity, where throughput becomes the limiting factor. PG-DyRA consistently outperforms DyPS across all parameters while showing lower sensitivity to variations. Crucially, our advantage amplifies under resource constraints, PG-DyRA achieves substantially higher completion

rates despite identical limited resources. This demonstrates that PG-DyRA provides inherent robustness through marginal contribution optimization that naturally adapts to resource availability.

The lower absolute performance on HappyValley compared to NYCTaxi reveals a fundamental tension between coordination quality and physical reachability. Even when the marginal contribution signal correctly identifies a high-demand region, an agent with limited movement speed may be unable to reach it within the remaining planning horizon, forcing a suboptimal local choice. The sensitivity results in Figure 5(a) support this interpretation. For example, PG-DyRA's advantage over DyPS narrows at lower movement speeds in HappyValley, where physical reachability rather than coordination quality becomes the binding constraint.

Table 2: Ablation Study on TCR of Happy Valley and NYCTaxi

Variation	Component		Happy Valley	NYCTaxi
w/o	Phase1		51.89±0.29	82.67±0.28
Replace	Inference →	MLP	46.67±0.28	67.45±0.16
	Infer-Vis →	MLP	48.34±0.15	75.89±0.12
	Infer-Invis →	MLP	49.78±0.10	77.56±0.13
PG-DyRA	/		**55.49**±0.21	**89.35**±0.24

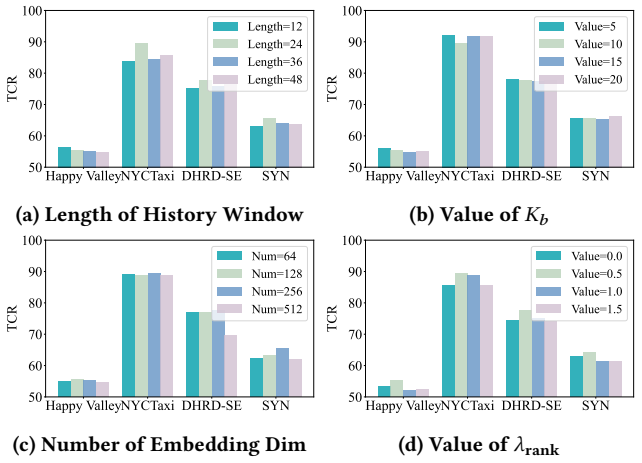
5.4 Ablation Study

To assess the impact of each component on the performance of PG-DyRA, we conduct an ablation study by comparing it with its variants, as detailed in Table 2. The variants include: (i) **w/o Phase 1** eliminates the pre-training phase, directly training the policy without inference network initialization. (ii) **Replace Inference with MLP** substitutes the entire intention inference module with a MLP. (iii) **Replace Infer-Vis with MLP** keeps the overall reasoning structure but uses MLP for visible agent allocation inference. (iv) **Replace Infer-Invis with MLP** similarly replaces only the invisible agents' inference component with MLP.

The pre-training phase proves essential, with its removal leading to notable performance decline, validating that Nash equilibrium solutions provide valuable initialization for subsequent policy learning. Replacing the inference network with MLPs consistently degrades performance across all variants, with the complete replacement showing the most significant impact. Interestingly, replacing either the visible or invisible inference branch alone causes moderate drops, indicating both components contribute meaningfully.

5.5 Hyperparameter Study

We analyze the impact of key architectural and training hyperparameters across four datasets: historical observation window size, state bank size K_b, embedding dimension d, and ranking consistency weight λ_{rank}. Results shown in Figure 6 demonstrate robust performance across wide parameter ranges, with TCR varying by less than 6 percentage points. The historical window exhibits diminishing returns beyond 24 timesteps, indicating recent observations sufficiently capture agent intentions. State bank size shows relative insensitivity, with 5-10 bases achieving comparable performance, suggesting a compact set of prototypical patterns effectively captures behavioral diversity. Embedding dimensionality maintains

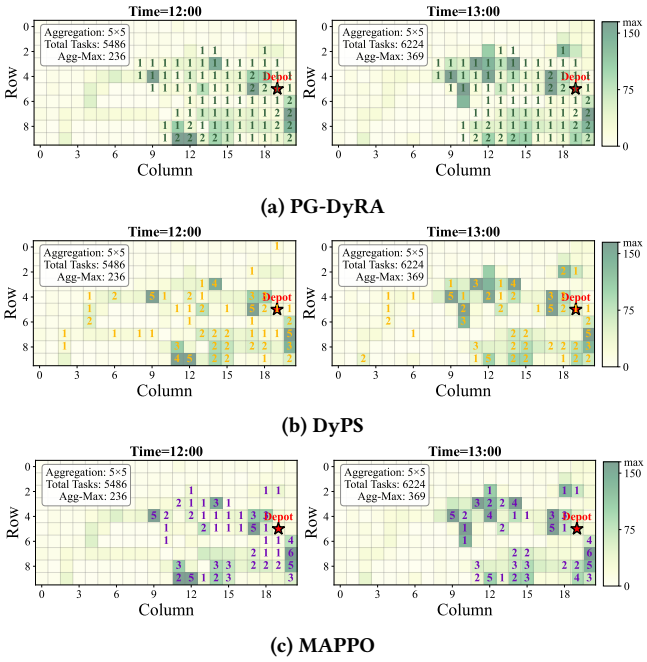


(a) Length of History Window (b) Value of K_b

(c) Number of Embedding Dim (d) Value of λ_{rank}

Figure 6: The results of varying the number of history window, K_b, embedding dim, and λ_{rank} on four datasets.

stable performance within 128-256, while extreme values may underfit or introduce unnecessary complexity. The ranking loss weight λ_{rank} demonstrates effectiveness when balanced appropriately, with excessive values degrading performance as the model overfits individual predictions at the expense of overall capacity accuracy.



(a) PG-DyRA

(b) DyPS

(c) MAPPO

Figure 7: The agent distribution and task density heatmap of Happy Valley from 12:00 to 13:00 on January 1, 2018. The numbers in the grid represent the agent count, while the color intensity of the grid indicates the task density.

5.6 Case Study

As shown in Figure 7, the agent distribution patterns reveal fundamental differences in coordination mechanisms across the three methods. Our PG-DyRA method exhibits spatially dispersed coverage with agents uniformly distributed across the service area, maintaining low local density. This dispersion pattern emerges from the marginal contribution mechanism: when multiple agents consider the same location, the diminishing marginal value (each additional agent receives diminishing marginal reward) creates a natural economic incentive for spatial separation. Consequently, agents autonomously spread across available regions to maximize their individual utilities, achieving broad service coverage without explicit anti-clustering constraints. The resulting distribution demonstrates that decentralized decision-making based on marginal returns effectively prevents resource concentration and promotes balanced allocation. In contrast, both DyPS and MAPPO produce concentrated deployment with pronounced clustering in high-demand areas, though through different mechanisms. For example, DyPS clustering results from role-based grouping where agents with similar behavioral characteristics converge to common targets. The comparison illustrates how marginal contribution-based coordination naturally enforces spatial diversity, while parameter sharing without diminishing returns leads to consensus-driven clustering.

6 Related Work

Dynamic Urban Resource Allocation. Dynamic resource allocation problems arise across urban mobility [14, 18, 34], delivery logistics [7, 26, 37], and emergency response [4, 22, 25], requiring coordination of multiple mobile units under spatiotemporal demand heterogeneity and coupled physical constraints. Centralized optimization methods [1, 3, 6] compute globally optimal assignments under complete information but scale exponentially at urban scale. RL approaches [21] recover scalability through adaptive policy learning, yet rely on a centralized controller aggregating global state. To enable distributed decision-making, MARL methods have been extensively studied. Communication-based methods [9, 24, 31, 33] improve coordination by exchanging learned messages during execution. Structure-based methods such as DyPS [27] and CoopRide [28] leverage role similarity or graph-based cooperation to coordinate without explicit messaging. However, these approaches either depend on information access that is unrealistic under practical sensing and bandwidth constraints, or lack theoretical guarantees that decentralized policies converge to globally optimal allocations.

Game-Theoretic Coordination. Game theory addresses the above gap through incentive alignment. Stackelberg games [35] have been applied to hierarchical resource allocation where a leader commits first and followers best-respond, but require a designated leader, which is impractical when all agents are symmetric mobile units. Coalition formation games [8, 17] allow agents to self-organize into cooperative groups via merge-split dynamics, but the negotiation overhead scales with agent count and is difficult to sustain under real-time mobility constraints. Potential games offer an alternative by aligning individual incentives with a global objective. In networked systems, potential games have been applied to distributed resource contention where each agent independently optimizes a private performance metric, e.g., processing delay [10], deployment cost [11], or data freshness [30], and the aggregate of individual payoffs serves as the potential function, enabling convergent distributed algorithms. These formulations, however, assume agents have full knowledge of the contention structure and operate at fixed infrastructure positions, without mobility constraints or partial observability.

In the urban computing domain, the most directly related work is GSTRL [13], which formulates collaborative resource allocation as a potential game and trains a centralized spatio-temporal RL policy using the coverage-based potential function as a shared reward. Different from GSTRL, PG-DyRA is a decentralized framework where each agent observes only nearby peers and acts without communication under per-step mobility and energy constraints. Most critically, instead of using the coverage objective as a shared reward signal, PG-DyRA constructs per-agent marginal contribution utilities that satisfy the exact potential game definition. Under partial observability, computing marginal contributions requires knowing C_r^{-i}, which is unobservable, and this gap motivates the dual-branch inference network and two-phase training of PG-DyRA.

7 Conclusion

In this paper, we propose PG-DyRA, a decentralized coordination framework for dynamic urban resource allocation. By formulating the problem as a potential game with marginal contribution utilities, we prove that distributed agents can in principle achieve globally optimal coordination through independent optimization, eliminating the exponential complexity of centralized planning while providing game-theoretic convergence guarantees to Nash equilibria. The critical innovation lies not merely in the theoretical framework, but in making it operational under real-world constraints. Our dual-branch inference network resolves the fundamental tension between partial observability and the global knowledge required for equilibrium computation, enabling agents to approximate marginal contributions from local observations alone. This synthesis of rigorous game theory with learned prediction addresses a key challenge in multi-agent systems, enabling theoretically grounded coordination without communication overhead or centralized control. Experiments on four urban datasets covering transportation, logistics, and infrastructure scenarios confirm that PG-DyRA consistently outperforms existing methods in both task coverage and energy efficiency.

Acknowledgments

The above work was supported in part by the National Natural Science Foundation of China (U25A20436, 62372047, 62406029), Guangxi Key Research & Development Program (FN2504240036, 2025FN96441087), Guangdong S&T Programme (2025B0101120006), the Supplemental Funds for Major Scientific Research Projects of Beijing Normal University, Zhuhai (ZHPT2023002), the Fundamental Research Funds for the Central Universities (2025JKZG069), Higher Education Research Topics of Guangdong Association of Higher Education in the 14th Five-Year Plan (24GYB207), Guangdong Provincial General Grant of 2026A1515010196, and AIRH, BNBU, No. 2020KSYS007.

References

[1] Akshay Ajagekar, Kumail Al Hamoud, and Fengqi You. 2022. Hybrid Classical-Quantum Optimization Techniques for Solving Mixed-Integer Programming Problems in Production Scheduling. *IEEE Transactions on Quantum Engineering* 3 (2022), 1–16.

[2] Yernat Assylbekov, Raghav Bali, Luke Bovard, and Christian Klaue. 2023. Delivery Hero Recommendation Dataset: A Novel Dataset for Benchmarking Recommendation Algorithms. In *Proceedings of the 17th ACM Conference on Recommender Systems* (Singapore, Singapore) *(RecSys '23)*. Association for Computing Machinery, New York, NY, USA, 1042–1044.

[3] Gaurav Baranwal, Dinesh Kumar, and Deo Prakash Vidyarthi. 2022. BARA: A blockchain-aided auction-based resource allocation in edge computing enabled industrial internet of things. *Future Generation Computer Systems* 135 (2022), 333–347.

[4] Souvik Basu, Siuli Roy, Somprakash Bandyopadhyay, and Sipra Das Bit. 2020. A Utility Driven Post Disaster Emergency Resource Allocation System Using DTN. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 50, 7 (2020), 2338–2350.

[5] Richard Bellman. 1966. Dynamic Programming. *Science* 153, 3731 (1966), 34–37.

[6] Dimitris Bertsimas, Patrick Jaillet, and Sébastien Martin. 2019. Online vehicle routing: The edge of optimization in large-scale applications. *Operations Research* 67, 1 (2019), 143–162.

[7] Giovanni Calabrò, Michela Le Pira, Nadia Giuffrida, Martina Fazio, Giuseppe Inturri, and Matteo Ignaccolo. 2023. A spatial agent-based model of e-commerce last-mile logistics towards a delivery-oriented development. *Transportation Research Interdisciplinary Perspectives* 21 (2023), 100895. doi:10.1016/j.trip.2023.100895

[8] Jiaxin Chen, Qihui Wu, Yuhua Xu, Nan Qi, Xin Guan, Yuli Zhang, and Zhen Xue. 2021. Joint Task Assignment and Spectrum Allocation in Heterogeneous UAV Communication Networks: A Coalition Formation Game-Theoretic Approach. *IEEE Transactions on Wireless Communications* 20, 1 (2021), 440–452.

[9] Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Mike Rabbat, and Joelle Pineau. 2019. TarMAC: Targeted Multi-Agent Communication. In *Proceedings of the 36th International Conference on Machine Learning (ICML '19, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, New York, NY, USA, 1538–1546.

[10] Wenhao Fan, Mingyu Hua, Yaoyin Zhang, Yi Su, Xuewei Li, Bihua Tang, Fan Wu, and Yuan'an Liu. 2023. Game-Based Task Offloading and Resource Allocation for Vehicular Edge Computing With Edge-Edge Cooperation. *IEEE Transactions on Vehicular Technology* 72, 6 (2023), 7857–7870.

[11] Xiangqiang Gao, Rongke Liu, and Aryan Kaushik. 2022. Virtual Network Function Placement in Satellite Edge Computing With a Potential Game Approach. *IEEE Transactions on Network and Service Management* 19, 2 (2022), 1243–1259.

[12] Richard M. Karp. 2010. *Reducibility Among Combinatorial Problems.* Springer Berlin Heidelberg, Berlin, Heidelberg, 219–241.

[13] Songxin Lei, Qiongyan Wang, Yanchen Zhu, Hanyu Yao, Sijie Ruan, Weilin Ruan, Yuyu Luo, Huaming Wu, and Yuxuan Liang. 2025. A Game-Theoretic Spatio-Temporal Reinforcement Learning Framework for Collaborative Public Resource Allocation. *arXiv preprint arXiv:2510.26184* (2025).

[14] Kaixiang Lin, Renyu Zhao, Zhe Xu, and Jiayu Zhou. 2018. Efficient Large-Scale Fleet Management via Multi-Agent Deep Reinforcement Learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 1774–1783.

[15] Dov Monderer and Lloyd S. Shapley. 1996. Potential Games. *Games and Economic Behavior* 14, 1 (1996), 124–143.

[16] Georgios Papoudakis, Filippos Christianos, and Stefano Albrecht. 2021. Agent Modelling under Partial Observability for Deep Reinforcement Learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS '21, Vol. 34)*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.). Curran Associates, Inc., Red Hook, NY, USA, 19210–19222.

[17] Nan Qi, Zanqi Huang, Fuhui Zhou, Qingjiang Shi, Qihui Wu, and Ming Xiao. 2023. A Task-Driven Sequential Overlapping Coalition Formation Game for Resource Allocation in Heterogeneous UAV Networks. *IEEE Transactions on Mobile Computing* 22, 8 (2023), 4439–4455.

[18] Zhiwei Tony Qin, Hongtu Zhu, and Jieping Ye. 2022. Reinforcement learning for ridesharing: An extended survey. *Transportation Research Part C: Emerging Technologies* 144 (2022), 103852.

[19] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2018. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML '18, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 4295–4304.

[20] Sijie Ruan, Jie Bao, Yuxuan Liang, Ruiyuan Li, Tianfu He, Chuishi Meng, Yanhua Li, Yingcai Wu, and Yu Zheng. 2020. Dynamic Public Resource Allocation Based on Human Mobility Prediction. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT '20, Vol. 4)*. Association for Computing Machinery, New York, NY, USA, Article 25, 22 pages.

[21] Soheil Sadeghi Eshkevari, Xiaocheng Tang, Zhiwei Qin, Jinhan Mei, Cheng Zhang, Qianying Meng, and Jia Xu. 2022. Reinforcement Learning in the Wild: Scalable RL Dispatching Algorithm Deployed in Ridehailing Marketplace. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (Washington DC, USA) *(KDD '22)*. Association for Computing Machinery, New York, NY, USA, 3838–3848.

[22] Deepshikha Sarma, Amrit Das, Pankaj Dutta, and Uttam Kumar Bera. 2022. A Cost Minimization Resource Allocation Model for Disaster Relief Operations With an Information Crowdsourcing-Based MCDM Approach. *IEEE Transactions on Engineering Management* 69, 5 (2022), 2454–2474.

[23] Alexey Skrynnik, Anton Andreychuk, Konstantin Yakovlev, and Aleksandr I. Panov. 2024. When to Switch: Planning and Learning for Partially Observable Multi-Agent Pathfinding. *IEEE Transactions on Neural Networks and Learning Systems* 35, 12 (2024), 17411–17424.

[24] Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. 2016. Learning multiagent communication with backpropagation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NeurIPS '16)*. Curran Associates Inc., Red Hook, NY, USA, 2252–2260.

[25] Hao Wang, Chi Harold Liu, Zipeng Dai, Jian Tang, and Guoren Wang. 2021. Energy-Efficient 3D Vehicular Crowdsourcing for Disaster Response by Distributed Deep Reinforcement Learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (Virtual Event, Singapore) *(KDD '21)*. Association for Computing Machinery, New York, NY, USA, 3679–3687.

[26] Hai Wang, Shuai Wang, Yu Yang, and Desheng Zhang. 2023. GCRL: Efficient Delivery Area Assignment for Last-mile Logistics with Group-based Cooperative Reinforcement Learning. In *2023 IEEE 39th International Conference on Data Engineering (ICDE '23)*. IEEE, 3522–3534.

[27] Jingwei Wang, Qianyue Hao, Wenzhen Huang, Xiaochen Fan, Zhentao Tang, Bin Wang, Jianye Hao, and Yong Li. 2024. DyPS: Dynamic Parameter Sharing in Multi-Agent Reinforcement Learning for Spatio-Temporal Resource Allocation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Barcelona, Spain) *(KDD '24)*. Association for Computing Machinery, New York, NY, USA, 3128–3139.

[28] Jingwei Wang, Qianyue Hao, Wenzhen Huang, Xiaochen Fan, Qin Zhang, Zhentao Tang, Bin Wang, Jianye Hao, and Yong Li. 2025. CoopRide: Cooperate All Grids in City-Scale Ride-Hailing Dispatching with Multi-Agent Reinforcement Learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (Toronto ON, Canada) *(KDD '25)*. Association for Computing Machinery, New York, NY, USA, 1457–1468.

[29] Jingyuan Wang, Jiawei Jiang, Wenjun Jiang, Chao Li, and Wayne Xin Zhao. 2021. LibCity: An Open Library for Traffic Prediction. In *Proceedings of the 29th International Conference on Advances in Geographic Information Systems* (Beijing, China) *(SIGSPATIAL '21)*. Association for Computing Machinery, New York, NY, USA, 145–148.

[30] Weizheng Wang, Gautam Srivastava, Jerry Chun-Wei Lin, Yaoqi Yang, Mamoun Alazab, and Thippa Reddy Gadekallu. 2023. Data Freshness Optimization Under CAA in the UAV-Aided MECN: A Potential Game Perspective. *IEEE Transactions on Intelligent Transportation Systems* 24, 11 (2023), 12912–12921.

[31] Yuanfei Wang, fangwei zhong, Jing Xu, and Yizhou Wang. 2022. ToM2C: Target-oriented Multi-agent Communication and Cooperation with Theory of Mind. In *International Conference on Learning Representations (ICLR '22)*.

[32] Laurence A Wolsey. 2020. *Integer programming.* John Wiley & Sons.

[33] Jing Xu, Fangwei Zhong, and Yizhou Wang. 2020. Learning Multi-Agent Coordination for Enhancing Target Coverage in Directional Sensor Networks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (Neurips '20, Vol. 33)*. Curran Associates, Inc., 10053–10064.

[34] Zhe Xu, Zhixin Li, Qingwen Guan, Dingshui Zhang, Qiang Li, Junxiao Nan, Chunyang Liu, Wei Bian, and Jieping Ye. 2018. Large-Scale Order Dispatch in On-Demand Ride-Hailing Platforms: A Learning and Planning Approach. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 905–913.

[35] Boran Yang, Dapeng Wu, Honggang Wang, Yishuang Gao, and Ruyan Wang. 2021. Two-Layer Stackelberg Game-Based Offloading Strategy for Mobile Edge Computing Enhanced FiWi Access Networks. *IEEE Transactions on Green Communications and Networking* 5, 1 (2021), 457–470.

[36] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and YI WU. 2022. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (Neurips '22, Vol. 35)*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.). Curran Associates, Inc., 24611–24624.

[37] Yuxiang Zeng, Yongxin Tong, and Lei Chen. 2019. Last-mile delivery made practical: an efficient route planning framework with theoretical guarantees. *Proc. VLDB Endow.* 13, 3 (Nov. 2019), 320–333. doi:10.14778/3368289.3368297

Appendices

A Theoretical Analysis

The proof of THEOREM 1.

Proof. We reduce from Partition, which is NP-complete [12]. A Partition instance is a multiset of positive integers $\{s_1, \ldots, s_N\}$, with $\sum_{k=1}^{N} s_k = 2W$. The question is whether some subset $A \subseteq \{1, \cdots, N\}$ satisfies $\sum_{k\in A} s_k = W$.

Given such an instance, set the number of timesteps to one (hereafter we suppress the superscript t, as the instance comprises a single decision step). Place two regions $\mathcal{R} = \{1, 2\}$ with demands $D_1 = D_2 = W$, and introduce N agents with capacities $c_k = s_k$. Set speed v and energy budgets large enough that every agent can reach either region, so $\mathcal{A}_i = \{1, 2\}$ for all i.

For any assignment $\mathbf{a}$, let $A = \{i : a_i = 1\}$ and $x = \sum_{i\in A} c_i$. Because every agent is assigned to exactly one of the two regions, $\sum_{i\notin A} c_i = 2W - x$, and the objective becomes

$$\Phi(\mathbf{a}) = \min(W,\, x) + \min(W,\, 2W - x).$$

Elementary case analysis gives

$$\Phi(\mathbf{a}) = \begin{cases} W + x, & 0 \le x \le W, \\ 3W - x, & W < x \le 2W. \end{cases}$$

The function is strictly increasing on $[0, W]$ and strictly decreasing on $(W, 2W]$, with a unique maximum of $2W$ attained if and only if $x = W$.

Since $\{c_k\} = \{s_k\}$, the value x ranges precisely over the subset sums of the Partition instance.

- If Partition is *yes*, pick any A with $\sum_{k\in A} s_k = W$ and assign those agents to region 1. Then $x = W$, so $\Phi(\mathbf{a}) = 2W$, hence $\Phi^* = 2W$.
- If Partition is *no*, then $x \neq W$ for every feasible assignment, so $\Phi(\mathbf{a}) < 2W$ for all $\mathbf{a}$, hence $\Phi^* < 2W$.

Thus $\Phi^* = 2W$ if and only if the Partition instance is yes. Any polynomial-time algorithm for computing Φ^* would therefore decide Partition in polynomial time, which is impossible unless P = NP. Since the constructed instance satisfies all constraints of the general problem, the general problem is NP-hard. □

Theorem 4. *Under partial observability, computing agent i's optimal action requires reasoning over exponentially many configurations of invisible agents.*

Proof. Consider agent i evaluating the utility of selecting region r. Agent i's marginal contribution to the system objective is:

$$u_i(r, \mathbf{a}_{-i}) = \min(D_r, C_r^{-i} + c_i) - \min(D_r, C_r^{-i}), \tag{16}$$

where $C_r^{-i} = \sum_{j\neq i: a_j = r} c_j$ represents the total capacity allocated to region r by all other agents. Computing $u_i(r, \mathbf{a}_{-i})$ requires knowing the joint action $\mathbf{a}_{-i}$ of all other agents.

Under partial observability, agent i directly observes only agents in $\mathcal{N}_i^{\text{vis}}$ but not those in $\mathcal{N}_i^{\text{inv}}$. Even if we assume the actions of visible agents are known, agent i must reason about how $|\mathcal{N}_i^{\text{inv}}|$ invisible agents distribute their capacities across R regions. Each invisible agent has R possible region choices, yielding $R^{|\mathcal{N}_i^{\text{inv}}|}$ possible joint action configurations for the invisible agents. For each configuration, agent i must compute C_r^{-i} for all $r \in \mathcal{R}$ and evaluate its marginal contribution, resulting in $O(R \cdot R^{|\mathcal{N}_i^{\text{inv}}|})$ computational complexity. This exponential dependence on the number of invisible agents renders exact computation intractable in large-scale systems. □

Remark 2. *The combination of NP-hardness and exponential complexity under partial observability renders traditional optimization approaches, such as integer programming, impractical for real-time deployment in city-scale scenarios with hundreds of agents and regions. This computational intractability motivates our method in Section 3, where we reformulate the problem as a potential game. This game-theoretic framework provides provable Nash equilibrium convergence guarantees while enabling efficient decentralized learning algorithms that scale to large agent populations.*

The proof of THEOREM 2.

Proof. We need to prove that for any agent $i \in \mathcal{N}$, any $\mathbf{a}_{-i} \in \mathcal{A}_{-i}$, and any $a_i, a'_i \in \mathcal{A}_i$, the equality holds

$$u_i(a_i, \mathbf{a}_{-i}) - u_i(a'_i, \mathbf{a}_{-i}) = \Phi(a_i, \mathbf{a}_{-i}) - \Phi(a'_i, \mathbf{a}_{-i}). \tag{17}$$

Let $\mathbf{a} = (a_i, \mathbf{a}_{-i})$ and $\mathbf{a}' = (a'_i, \mathbf{a}_{-i})$ denote the joint actions before and after agent i changes its strategy. Without loss of generality, assume $a_i = r \in \mathcal{R}$ and $a'_i = r' \in \mathcal{R}$.

First, we establish capacity relationships. (i) For region r, under $\mathbf{a}$, the set of agents selecting region r is $\mathcal{N}_r(\mathbf{a}) = \mathcal{N}_r(\mathbf{a}') \cup \{i\}$. The total capacity is $C_r(\mathbf{a}) = C_r(\mathbf{a}') + c_i$, where $C_r(\mathbf{a}) = \sum_{j:a_j=r} c_j$. (ii) For region r', under $\mathbf{a}'$, the set of agents selecting region r' is $\mathcal{N}_{r'}(\mathbf{a}') = \mathcal{N}_{r'}(\mathbf{a}) \cup \{i\}$. The total capacity is $C_{r'}(\mathbf{a}') = C_{r'}(\mathbf{a}) + c_i$. (iii) For other regions $h \neq r$ and r', we could get $\mathcal{N}_h(\mathbf{a}) = \mathcal{N}_h(\mathbf{a}')$, $C_h(\mathbf{a}) = C_h(\mathbf{a}')$

Second, we compute individual utility difference. According to the definition of marginal contribution utility (DEFINITION 9),

$$\begin{aligned} \Delta u_i &= u_i(a_i, \mathbf{a}_{-i}) - u_i(a'_i, \mathbf{a}_{-i}) \\ &= [\min(D_r, C_r(\mathbf{a})) - \min(D_r, C_r(\mathbf{a}'))] \\ &\quad - [\min(D_{r'}, C_{r'}(\mathbf{a}')) - \min(D_{r'}, C_{r'}(\mathbf{a}))]. \end{aligned} \tag{18}$$

Third, we compute potential function difference:

$$\begin{aligned} \Delta\Phi &= \Phi(\mathbf{a}) - \Phi(\mathbf{a}'), \\ &= \sum_{g\in\mathcal{R}} \min(D_g, C_g(\mathbf{a})) - \sum_{g\in\mathcal{R}} \min(D_g, C_g(\mathbf{a}')). \end{aligned} \tag{19}$$

Since only agent i changes its strategy, capacities in regions other than r and r' remain unchanged. Thus,

$$\begin{aligned} \Delta\Phi &= [\min(D_r, C_r(\mathbf{a})) + \min(D_{r'}, C_{r'}(\mathbf{a}))] \\ &\quad - [\min(D_r, C_r(\mathbf{a}')) + \min(D_{r'}, C_{r'}(\mathbf{a}'))]. \end{aligned} \tag{20}$$

Comparing Step 2 and Step 3, we have $\Delta u_i = \Delta\Phi$. Therefore, Φ satisfies the definition of potential game. □

The proof of COROLLARY 1.

Proof. The existence of pure strategy Nash equilibrium follows from the finite improvement property of exact potential games [15]. Starting from any joint action, any sequence of utility-improving unilateral deviations must terminate at a pure strategy Nash equilibrium, since each improvement strictly increases the potential function Φ and the action space is finite.

For optimality, let $\mathbf{a}^*$ be a global maximizer of the system objective Φ over the feasible joint action space. By definition of the maximizer, for any agent i and any $a_i \in \mathcal{A}_i$:

$$\Phi(a_i^*, \mathbf{a}_{-i}^*) \geq \Phi(a_i, \mathbf{a}_{-i}^*) \tag{21}$$

By Theorem 2:

$$u_i(a_i^*, \mathbf{a}_{-i}^*) \geq u_i(a_i, \mathbf{a}_{-i}^*) \tag{22}$$

This holds for all agents i and all their feasible actions, implying that no unilateral deviation can increase any agent's utility. Hence $\mathbf{a}^*$ is a pure strategy Nash equilibrium, i.e., the global optimum is attained at a Nash equilibrium. □

Theorem 5. *Adding individual movement costs to the marginal contribution utility preserves the potential game property.*

Proof. Define the modified utility as $\tilde{u}_i(a_i, \mathbf{a}_{-i}) = u_i(a_i, \mathbf{a}_{-i}) - \beta \cdot d_i(a_i)$, where $d_i(a_i) = d(p_i, a_i)$ is the movement distance incurred by agent i's action a_i, independent of $\mathbf{a}_{-i}$. Define the modified potential function:

$$\tilde{\Phi}(\mathbf{a}) = \Phi(\mathbf{a}) - \sum_{j=1}^{N} \beta \cdot d_j(a_j). \tag{23}$$

We verify the potential game condition. For any agent i and any two feasible actions a_i, a'_i, with $\mathbf{a}_{-i}$ fixed:

$$\begin{aligned}
&\tilde{u}_i(a'_i, \mathbf{a}_{-i}) - \tilde{u}_i(a_i, \mathbf{a}_{-i}) \\
&\quad = [u_i(a'_i, \mathbf{a}_{-i}) - \beta d_i(a'_i)] - [u_i(a_i, \mathbf{a}_{-i}) - \beta d_i(a_i)] \\
&\quad = [u_i(a'_i, \mathbf{a}_{-i}) - u_i(a_i, \mathbf{a}_{-i})] - \beta[d_i(a'_i) - d_i(a_i)] \\
&\quad = [\Phi(a'_i, \mathbf{a}_{-i}) - \Phi(a_i, \mathbf{a}_{-i})] - \beta[d_i(a'_i) - d_i(a_i)].
\end{aligned} \tag{24}$$

Now expand $\tilde{\Phi}$ on both joint actions. Since only agent i's action differs, all other agents' movement costs cancel:

$$\begin{aligned}
&\tilde{\Phi}(a'_i, \mathbf{a}_{-i}) - \tilde{\Phi}(a_i, \mathbf{a}_{-i}) \\
&\quad = [\Phi(a'_i, \mathbf{a}_{-i}) - \sum_{j \neq i} \beta d_j(a_j) - \beta d_i(a'_i)] \\
&\qquad - [\Phi(a_i, \mathbf{a}_{-i}) - \sum_{j \neq i} \beta d_j(a_j) - \beta d_i(a_i)] \\
&\quad = [\Phi(a'_i, \mathbf{a}_{-i}) - \Phi(a_i, \mathbf{a}_{-i})] - \beta[d_i(a'_i) - d_i(a_i)].
\end{aligned} \tag{25}$$

Comparing the two results, we have

$$\tilde{u}_i(a'_i, \mathbf{a}_{-i}) - \tilde{u}_i(a_i, \mathbf{a}_{-i}) = \tilde{\Phi}(a'_i, \mathbf{a}_{-i}) - \tilde{\Phi}(a_i, \mathbf{a}_{-i}), \tag{26}$$

which satisfies Definition 10. Thus the modified game with utilities $\tilde{u}_i$ remains an exact potential game with potential function $\tilde{\Phi}$. □

B Ranking Consistency Loss

Agent i's region selection is governed by comparing marginal utilities across all feasible regions. This comparison is structurally determined by the relative ordering of C_r^{-i} across regions. A region whose aggregate capacity already approaches or exceeds D_r yields near-zero marginal return, while underserved regions offer substantially higher utility. Consequently, a capacity predictor that preserves the correct rank ordering produces better-aligned region comparisons than one that minimises pointwise error alone.

Let $\hat{\mathbf{c}} \in \mathbb{R}^R$ denote the predicted aggregate capacity allocation across all R regions, and $\mathbf{c}^* \in \mathbb{R}^R$ the corresponding ground-truth aggregate allocation. We measure ordering agreement via the Spearman rank correlation. Let $\mathbf{r}_{\text{pred}}$ and $\mathbf{r}_{\text{tgt}}$ be the rank vectors of $\hat{\mathbf{c}}$ and $\mathbf{c}^*$, respectively, where the k-th entry records the rank of region k among all R regions in ascending order. The ranks are produced by a differentiable soft-ranking operator, so that $\mathcal{L}_{\text{rank}}$ propagates gradients to the predicted allocation. By normalizing rank vectors, we obtain $\tilde{\mathbf{r}}_{\text{pred}}$ and $\tilde{\mathbf{r}}_{\text{tgt}}$. The Spearman rank correlation between predicted and ground-truth orderings is:

$$\rho = \frac{1}{R-1} \sum_{k=1}^{R} \tilde{r}_{\text{pred}}[k] \cdot \tilde{r}_{\text{tgt}}[k], \tag{27}$$

which takes values in $[-1, 1]$. The ranking consistency loss is accordingly defined as $\mathcal{L}_{\text{rank}} = 1 - \rho$.

C Datasets

Happy Valley [20]. This dataset is derived from Tencent's location-based population density heatmaps, capturing crowd movement within and around Beijing Happy Valley from January 1 to October 31, 2018. Each location (a $10m \times 10m$ grid) has a crowd flow observation each hour. The area of interest is about $5.5 \times 10^5 m^2$, containing $51 \times 108 = 5,508$ uniform grids. The average aggregated hourly crowd flow across all grids is approximately 6,470, and the average daily total reaches about 77,650, which is derived from the 12 hours of daily operation.

NYCTaxi [29]. This dataset is derived from New York Taxi & Limousine Commission. From this dataset, we can obtain information on the entry and exit of taxi passengers to the grids, and we followed the usual grid division method ($10 \times 20 = 200$). We used the data from January 1 to March 31, 2014. The average aggregated hourly crowd flow across all grids is approximately 218, and the average daily total reaches about 5,247.

DHRD-SE [2]. DHRD-SE is a food delivery dataset in Stockholm, which spans a period of three months. Due to privacy requirements, the geographical location of this dataset has been encrypted using geohash. After decrypting the location, we divided it into a $1km \times 1km$ grid, resulting in $45 \times 40 = 1,800$ grids. The average aggregated hourly orders across all grids is approximately 538, and the average daily total reaches about 9,139.

SYN. SYN is the square grid world we have generated, containing $25 \times 25 = 625$ grids. On average, 100 tasks are generated every hour (with a fluctuation of 20%), randomly distributed across 625 grids. Approximately 2,400 tasks are generated each day.

Table 3: Statistics of experimental scenarios

Scenarios	Happy Valley	NYCTaxi	DHRD-SE	SYN
Grids	51×108	10×20	45×40	25×25
Demands	77.7K	5.2K	9.1K	2.4K
Agents	95	42	30	8
Capacity	100	5	30	15
Energy	250	100	300	500
Speed	20	10	30	15
Vision	20	10	30	15
Time Intv.	60min	60min	60min	30min