香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

Algorithmic Discrimination in On-Demand Food-Delivery Platforms using Multi-Agent Reinforcement Learning

Presenter: Zijian Zhao

Ph.D. Student in Civil Engineering (Scientific Computation)

Ph.D. Qualifying Exam, December 2, 2025

Supervisor: Prof. Sen Li

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Part I: Background of Algorithm Discrimination in Food-Delivery Markets

Background: Algorithm Discrimination in Food-Delivery Platforms

- **Learning algorithm in food-delivery platforms:**
  - profit maximization: Platforms aim to achieve optimal control over couriers while offering the lowest payment.
  - overuse of privacy data: Platforms utilize personal data to differentiate between various types of couriers.
- **Discriminatory algorithmic labor management:**
  - personalized payment: Couriers have varying expected salaries for the same work (referred to as 'reservation value') → *violating principle of 'equal pay for equal work'*
  - dispatching preference: Couriers with high willingness to work present higher certainty and lower costs for platforms → prioritize couriers with more favorite working pattern
- **Couriers controlled by algorithms:**

Background: Data Privacy Regulation (DPR)

- **Take GDPR (Europe) as an example:**

GDPR

PIPL

CCPA

  - Article 6.1.a : Data processing is lawful only when the data subject has provided consent for one or more specific purposes.
  - Article 7.2: Individuals have the right to withdraw consent at any time, and the process for withdrawal must be as simple as providing consent.

- **Basic principles of DPR:**
  - data minimization
  - user content, individual control over personal information

Research Questions

- **Research questions:**
  - Q1: How does the discriminatory algorithm influence platforms and couriers? (Project 1)
  - Q2: How does data privacy regulation policy influence food delivery market? (Project 2)
- **Challenges:**
  - The dispatching strategy and payment strategy of platforms are coupled. (Project 1)
  - Couriers' behaviors occur on different time scales: (Project 2)
    - long-term: decide whether to work on the platform and whether to share their personal information
    - short-term: determine whether to accept assigned orders
  - Complex interaction and training process between platform and couriers (Project 1&2)

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Part II: Platform Modeling via Hybrid-Action MARL

Zijian Zhao, Sen Li*, "Discriminatory Order Assignment and Payment-Setting of On-Demand Food-Delivery Platforms: A Multi-Action and Multi-Agent Reinforcement Learning Framework" (under thrid-round review, Transportation Research Part E: Logistics and Transportation Review (TR_E))

System Workflow

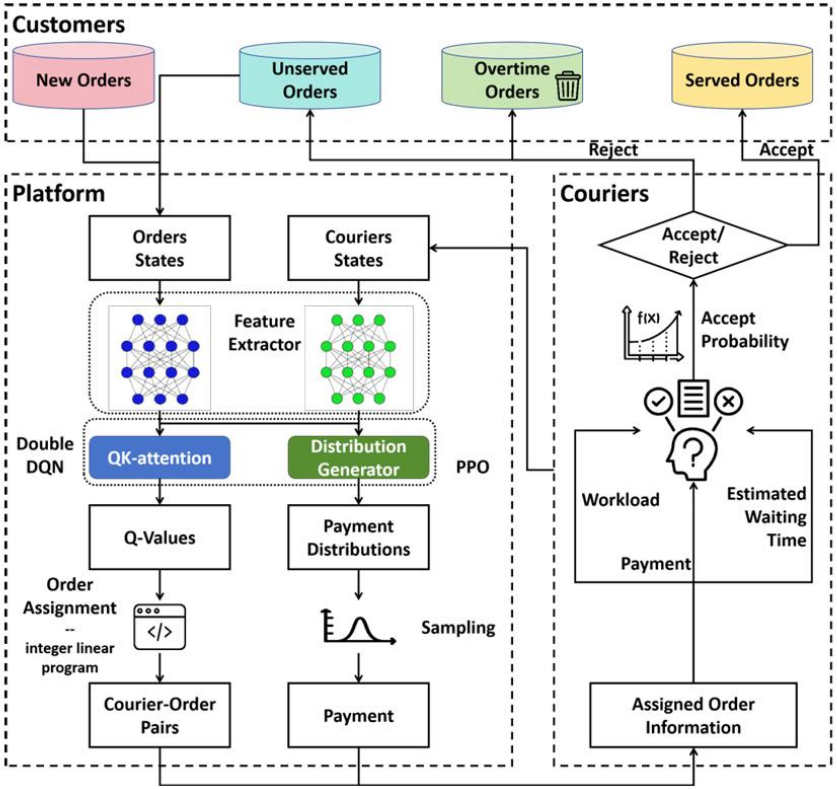


Figure 1: Workflow: At each time step, the platform first performs order assignment using DQN and sets payments using PPO. Couriers then decide whether to accept the assignment based on their own reservation values. Orders that are rejected will be declined by customers if they exceed the maximum waiting time.

- **Platform**: determine *order assignment* and *payment setting*
- **Couriers**: determine *whether to accept the assigned order* based on payment and individual reservation value
- **Customers & Restaurants:**
  - *set orders* at arbitrary times
  - *cancel the order* if not confirmed within a maximium time threshold

Model for Platform: MDP

- **State:** $S_t = \{s_{i,t} | i \in I\}$
  - t : time step; I : set of courier indices;
  - $s_{i,t}$: state of courier i at time t , defined as: $s_{i,t} = [t, C_{i,t}, O_t^w, G]$
    - $C_{i,t}$: courier information, including location, en route orders, and remaining capacity ...
    - O_t^w: informations of orders to be assigned at time t
    - G: global information, such as the number of couriers who choose to opt-in and opt-out ...

Order $j \in \mathcal{J}$		Courier $i \in \mathcal{I}$	
Parameter	**Symbol**	**Parameter**	**Symbol**
Origin Location	$o_j = (o_j^x, o_j^y) \in \mathbb{R}^2$	Current Location	$l_{i,t} = (l_{i,t}^x, l_{i,t}^y) \in \mathbb{R}^2$
Destination Location	$d_j = (d_j^x, d_j^y) \in \mathbb{R}^2$	Remaining Capacity	$c_{i,t}^r \in \mathbb{N}$
Service Request Time	$q_j \in \mathbb{R}^*$	En-route Orders	$O_{i,t}^e \in \mathbb{R}^{k_{i,t},6}$
Expected Arrival Time	$z_j \in \mathbb{R}^*$	Scheduled Route	$u_{i,t} \in \mathbb{R}^{g_{i,t},2}$

Table 1: Parameters of order j and courier i at time t: In this table, x, y represent the latitude and longitude coordinates, respectively, $k_{i,t}$ denotes the number of onboard orders, and $g_{i,t}$ denotes the number of intermediate nodes in scheduled route.

Model for Platform: MDP

❑ How platforms record courier individual information in state space?

- Platforms can record courier behavioral data in the state space to distinguish their types. However, recording all couriers' behavioral data seems unfeasible, since the data volume is often too large.

- We propose an alternative approach: recording the average unit payment (payment per minute of work) related to accepting and rejecting behaviors for each courier.

$$h^a_{i,t+1} = h^a_{i,t} + \eta(p^u_{i,t} - h^a_{i,t})\zeta_{i,t} \ ,$$
$$h^r_{i,t+1} = h^r_{i,t} + \eta(p^u_{i,t} - h^r_{i,t})(1 - \zeta_{i,t}) \ ,$$

  - $h^a_{i,t}, h^r_{i,t}$: average unit payment about accepting/rejecting behavior of courier *i* at time *t*
  - $p^u_{i,t}$: unit payment for courier *i* at time *t*
  - $\zeta_{i,t}$: behavior of courier *i* at time *t* (*1* represents acceptance and *0* represents rejection)
  - η: update rate

- Platforms can classify couriers based on the mean value of h^a and h^r (i.e. $\frac{h^a+h^r}{2}$).

Model for Platform: MDP

- **Action:** $A_t = \{a_{i,t} | i \in I\}$
  - $a_{i,t} = [\kappa_{i,t}, p_{i,t}]$ represents the action for courier i at time t
  - $\kappa_{i,t}$: decision vector indicating which orders are assigned to courier *i* at time *t*
    - $\kappa_{i,t}[j] = 1$ indicates order *j* is assigned to courier *i*
  - $p_{i,t}$: payment to courier *i* at time *t*

- **Reward function:**

$$\mathcal{R}^{t+1}(S_t, A_t) = \sum_{i \in \mathcal{I}^w} r_{i,t+1} = \sum_{i \in \mathcal{I}^w} \mathcal{R}_i^{t+1}(s_{i,t}, a_{i,t} | \zeta_{i,t})$$

$$\mathcal{R}_i^{t+1}(s_{i,t}, a_{i,t} | \zeta_{i,t}) = \begin{cases} \beta_1 + \beta_2 \tau_j - \beta_3 p_{i,t} - \beta_4 \chi_{i,j,t} - \beta_5 \rho_{i,j,t} , & \kappa_{i,t}[j] = 1, \zeta_{i,t} = 1 \\ 0 , & \text{otherwise} \end{cases},$$

  - R^{t+1}: global reward at time *t*; R_i^{t+1}: reward for courier *i* at time *t*; I^w: set of working courier indices
  - β: positive weights
  - τ_j: trip distance of oredr *j*; $\chi_{i,j,t}$: the number of en-route orders for courier *i* at time *t* that will exceed their expected time; $\rho_{i,j,t}$: additional travel time of all en-route orders for courier *i* at time *t*

Model for Couriers: Logit Model

- **Logit model:**

$$\mathcal{P}_i^a(p_{i,t}) = \frac{1}{1 + e^{-(u_{i,t}^a - u_i^r)}} ,$$

- P_i^a: order acceptance probability of courier *i*
- $u_{i,t}^a$: utility of accepting the order, proportional to payment $p_{i,t}$, inversely proportional to reservation value p_i^r
- u_i^r: utility of rejecting the order

Solution Methodology for MDP of Platform

❑ DDQN for order assignment:

- ➢ Summary: A neural network is employed to estimate the long-term reward (Q-value) for each courier selecting each order at each time step.
- ➢ Utilization: Find the optimal dispatching strategy to maximize the global Q-value (long-term reward) at each step through bipartite matching.

$$
\begin{aligned}
&\max_{x \in X} \sum_{i \in \mathcal{I}^w} \mathrm{W}_t(i,j) \cdot x_{i,j}, \\
\text{s.t.} \quad &\sum_{i \in \mathcal{I}^w} x_{i,j} \le 1, \quad \forall j \in \mathcal{J}_t^w, \\
&\sum_{j \in \mathcal{J}_t^w} x_{i,j} \le 1, \quad \forall i \in \mathcal{I}^w, \\
&\sum_{i \in \mathcal{I}^w} x_{i,j} \cdot \mathcal{D}(l_{i,t}, o_j) \le D_{max}, \quad \forall j \in \mathcal{J}_t^w, \\
&x_{i,j} \in \{0,1\}, \quad \forall i \in \mathcal{I}^w, j \in \mathcal{J}_t^w,
\end{aligned}
$$

- $W_t(i,j)$: Q-value representing the value of assigning order *j* to courier *i* at time *t*
- $x_{i,j}$: binary indicator for whether order *j* is assigned to courier *i*
- $D(l_{i,t}, o_j)$: distance between courier *i* and the origion of order *j*
- D_{max}: maximium matching distance
- J_t^w: indices of orders to be assigned at time *t*

- ➢ Training: gradient descent to minimize Temporal Difference (TD) error

$$
\bar{L}_Q = \mathbb{E}_{\pi_\theta^A, \pi_\theta^P} \left[\mathcal{R}^{t+1} + \gamma \mathrm{Q}_{\pi_{\theta^-}^A}^{DDQN}(S_{t+1}, \underset{\kappa_{t+1} \in \Phi_{t+1}}{\arg\max}\ \mathrm{Q}_{\pi_\theta^A}^{DDQN}(S_{t+1}, \kappa_{t+1}; \theta | \pi_\theta^P); \theta^- | \pi_{\theta^-}^P) - \mathrm{Q}_{\pi_\theta^A}^{DDQN}(S_t, \kappa_t; \theta | \pi_\theta^P) \right]
$$

Solution Methodology for MDP of Platform

❑ PPO for payment setting:

- ➢ Summary: An actor network is employed to directly generate payments for each courier, while a critic network estimates the V-value to guide the updating of the critic network.
- ➢ Finding: In our scenario, the PPO can directly utilize the Q-network of DDQN as the critic network.

$$
\begin{aligned}
V_{\pi^P}^{PPO}(s_{i,t}|\pi^A) &= \mathbb{E}_{\pi^A,\pi^P}\left[\sum_{\tau\in\mathcal{T}}\gamma^\tau\cdot\mathcal{R}_i^{t+1+\tau}(s_{i,t+\tau},a_{i,t+\tau})\mid s_{i,t}\right] \\
&= \mathbb{E}_{\pi^A,\pi^P}\left[\sum_{\tau\in\mathcal{T}}\gamma^\tau\cdot\mathcal{R}_i^{t+1+\tau}(s_{i,t+\tau},a_{i,t+\tau})\mid s_{i,t},\pi^A(s_{i,t})\right] \\
&= \mathbb{E}_{\pi^A,\pi^P}\left[\sum_{\tau\in\mathcal{T}}\gamma^\tau\cdot\mathcal{R}_i^{t+1+\tau}(s_{i,t+\tau},a_{i,t+\tau})\mid s_{i,t},\kappa_{i,t}^*\right] \\
&= Q_{\pi^A}^{DDQN}(s_{i,t},\kappa_{i,t}^*|\pi^P)\ ,
\end{aligned}
$$

- π_A: policy of DDQN (assignment)
- π_P: policy of PPO (payment)
- $\kappa_{i,t}^*$: optimal order assignment action of courier *i* at *t*

- ➢ Training of actor: policy gradient to maximize the objective function

$$
\mathrm{J}^{PPO}(\pi_\theta^P|\pi_\theta^A) = \mathbb{E}_{\pi_\Theta^A,\pi_\Theta^p}\left[\frac{\pi_\theta(P_t|S_t)}{\pi_\Theta(P_t|S_t)}\mathcal{A}_{\pi_\theta^P}^{PPO}(S_t,P_t;\theta|\pi_\theta^A)\right]
$$

Neural Network Architecture

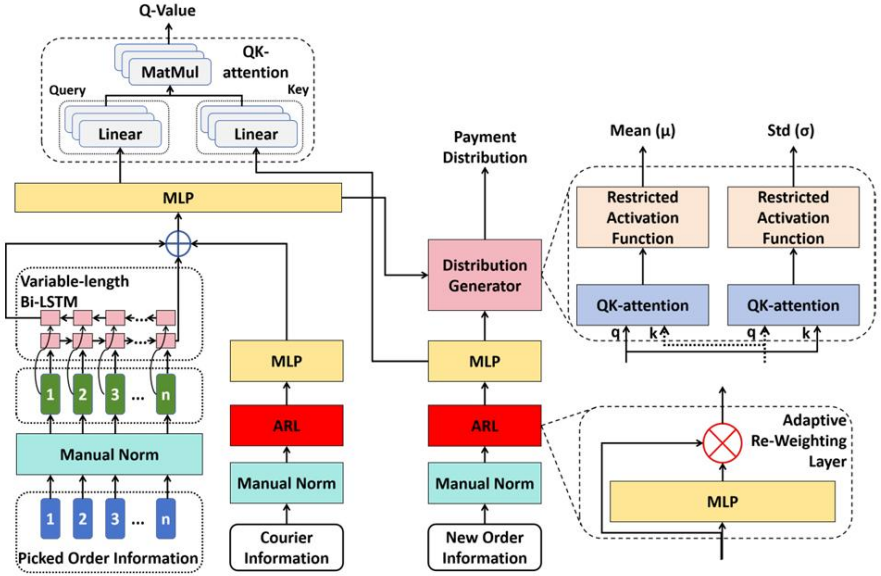


Figure 2: Architecture of the neural network: The proposed neural network consists of two upstream stems and two downstream heads. (1) One upstream stem employs an MLP-based network to extract features from new orders to be assigned. The other upstream stem includes two small branches: one utilizes a Bi-LSTM to process the sequence information of assigned orders for each courier, while the other uses an MLP to process additional courier information. The outputs from these two branches are then fused using an MLP. (2) The two downstream heads have a similar structure, all based on QK-attention, which efficiently captures the relationships between courier information embeddings and new order information embeddings. One head employs a QK-attention mechanism to directly predict the Q-value, while the other one generate the distribution of payment for each courier-order pair.

- **Upstream layer: dual encoder**
  - Bi-LSTM for sequential enroute order information compression
  - ARL-enhanced[1,2] MLP for non-sequential feature extractor
    - Rescale Input: $x \circ \mathrm{MLP}(x)$
  - MLP for sequential and non-sequential feature fusion

- **Downstream haed: QK-Attention[3]**
  - reduce multiplicative complexity to additive complexity
  - QK-Attention: $y = \mathrm{NN}(x_1) \cdot \mathrm{NN}(x_2)$
  - Conventional method: $y = \mathrm{NN}([x_1; x_2])$

[1] Chen T, Wang Y, Chen H, et al. Modelling the 5g energy consumption using real-world data: Energy fingerprint is all you need[C]//2025 IEEE Globecom Workshops (GC Wkshps). IEEE, 2025: 1-6.
[2] Zhao Z, Meng F, Lyu Z, et al. Csi-bert2: A bert-inspired framework for efficient csi prediction and classification in wireless communication and sensing[J]. arXiv preprint arXiv:2412.06861, 2024.
[3] Zhao Z, Chen T, Cai Z, et al. Crossfi: A cross domain wi-fi sensing framework based on siamese network[J]. IEEE Internet of Things Journal, 2025.

Alternative Training with Rollback Mechanism

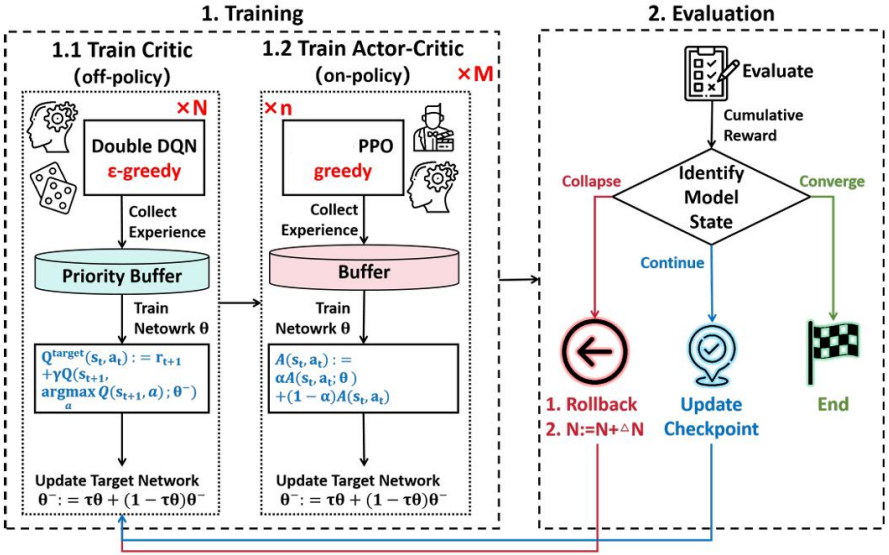


Figure 3: Training Process: The workflow illustrates the training process for both DQN and PPO. The two modules are trained separately, with DQN utilizing an off-policy method and PPO employing an on-policy method. The entire model is evaluated regularly to monitor its state. If the state is normal, a checkpoint is saved. If the model collapses, we rollback to the last checkpoint and update the training strategy by increasing the number of training episodes for DQN. Training will stop once the model has converged.

- **Alternative training strategy**
  - Training DQN and PPO simultaneously can be unstable. It's difficult to determine whether changes in rewards result from one's own actions or those of the other algorithm.
- **Rollback mechanism**
  - At times, the PPO policy may collapse due to the critic (DQN) not aligning with updates from the actor.
  - When a collapse occurs, we
    - roll back to the last checkpoint
    - reduce the frequency of PPO training updates

Case Studies

- **Food delivery in Hong Kong, China**

Configuration	Our Setting
Batch Size	512
Optimizer	Adam
Learning Rate α	0.0005
Exploration Rate of DQN ϵ	$0.99 \rightarrow 0.0005$
Decay Rate of α and ϵ	0.99
Updating Rate of Target Network τ	0.005
$[N, n, M, m, \triangle N, K]$ in Algorithm 2	[4,1,2,10,1,10]
Discount Factor γ	0.99
Agent (Courier) Amount	1,000
Total Number of Parameters	149K
$\beta_1 \sim \beta_5$ in Platform Reward Function (Eq. 6)	[3,5,4,3,1]

Table 2: A brief summary of the model parameters.

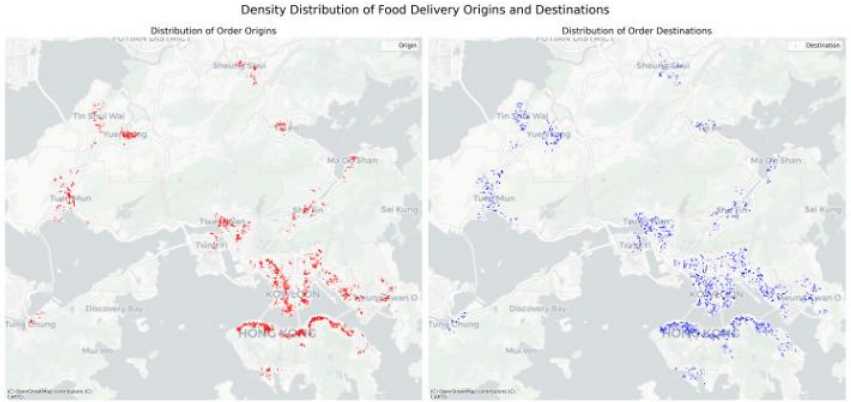


Figure 4: Order spatial distribution map: the red points represent the order origins, while the blue points indicate the destinations.

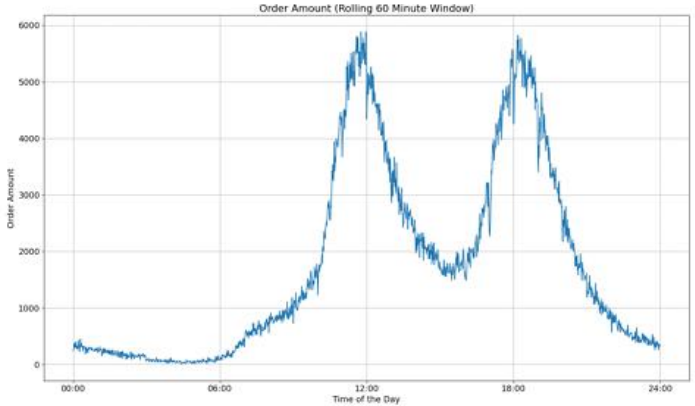
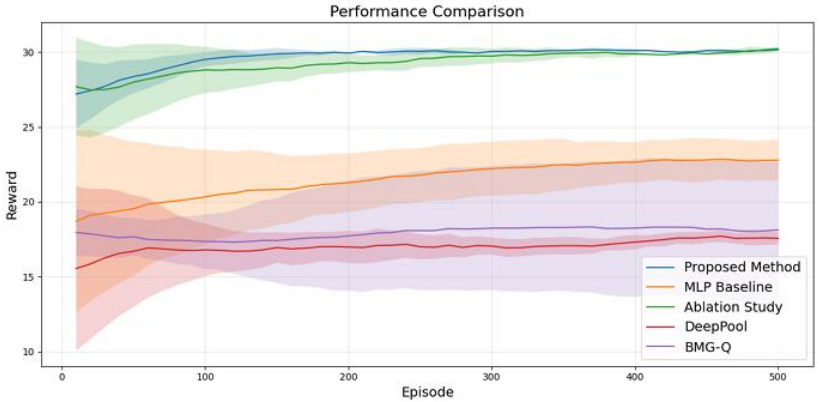


Figure 5: Order Temporal Distribution.

Case Studies

- **Performance comparison**

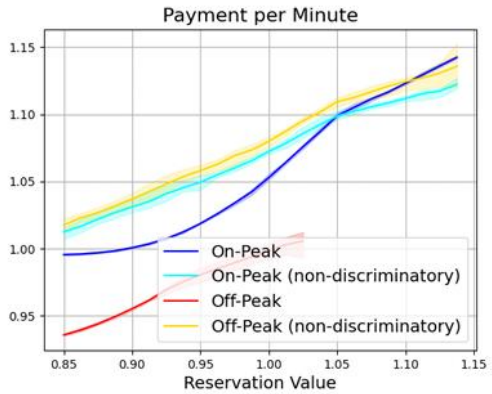


Metrics \ Methods	**Off-Peak** (1960 orders in total)		
	Ours	**DeepPool**	**BMG-Q**
Reward (Profits)	**19.60**	17.87	19.30
Order Pickup	1954	**1958**	1943
Average Delivery Time	17.84	**15.45**	17.57
Overtime Order Amount	34	**9**	36
Total Payment (10^3)	46.10	50.06	**41.37**
Metrics \ Methods	**On-Peak** (5880 orders in total)		
	Ours	**DeepPool**	**BMG-Q**
Reward (Profits)	**30.74**	17.26	21.37
Order Pickup	**3686**	3272	3445
Average Delivery Time	**15.56**	16.08	16.82
Overtime Order Amount	57	**36**	75
Total Payment (10^3)	90.67	82.55	**77.84**

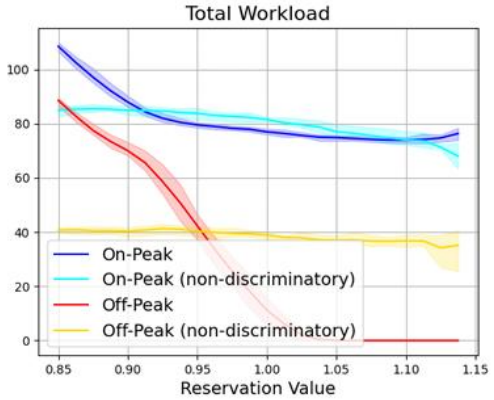
Metrics \ Methods	**Off-Peak** (1960 orders in total)		
	Ours	**Ablation Study**	**MLP Baseline**
Reward (Profits)	**19.60**	17.37	13.30
Order Pickup	**1954**	**1954**	1896
Average Delivery Time	17.84	**15.95**	18.17
Overtime Order Amount	34	**25**	36
Total Payment (10^3)	**46.10**	47.87	51.15
Metrics \ Methods	**On-Peak** (5880 orders in total)		
	Ours	**Ablation Study**	**MLP Baseline**
Reward (Profits)	**30.74**	30.36	20.59
Order Pickup	**3686**	3654	3502
Average Delivery Time	**15.56**	15.65	15.58
Overtime Order Amount	57	68	**49**
Total Payment (10^3)	**90.67**	95.21	99.88

Case Studies

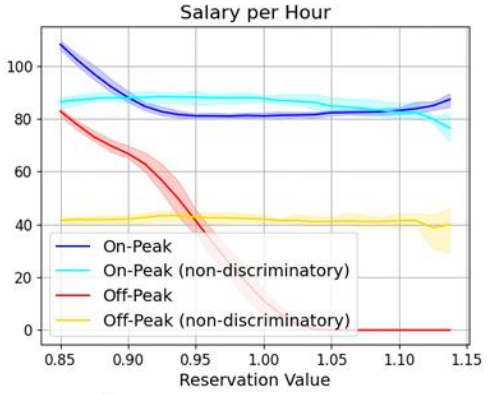
- **Comparison between discriminatory algorithm and non-discriminatory algorithm**



(c) Payment per minute for the orders assigned to the courier



(b) Total trip time (minutes) of all order assigned to the courier per hour



(d) Salary per hour for the courier

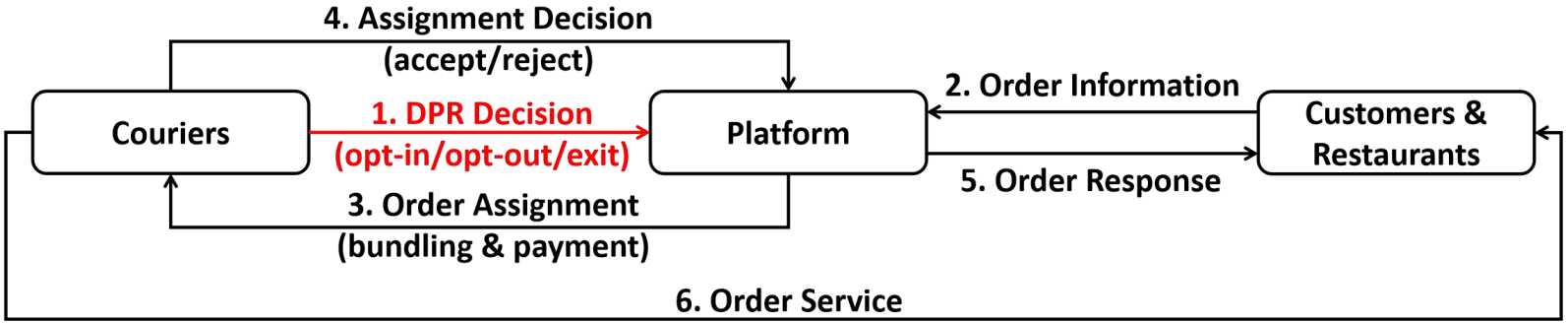
- **How platforms discriminate different types of couriers?**
  - Higher reservation couriers receive:
    - higher unit payments
    - lower workloads
    - lower total earnings

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Part III: Courier Modeling via MA-CMAB

Zijian Zhao, Sen Li*, "The Impacts of Data Privacy Regulations on Food-Delivery Platforms", Transportation Research Part C: Emerging Technologies (TR_C), 2025

System Workflow



- **Platform**: determine *order assignment* and *payment setting*
- **Couriers**:
  - long-term: choose from *opt-in* (work with providing personal information), *opt-out* (work without providing personal information), or *exit* (do not work) at the beginning of each time period
  - short-term: determine *whether to accept the assigned order* based on payment and individual reservation value
- **Customers & Restaurants:**
  - *set orders* at arbitrary times
  - *cancel the order* if not confirmed within a maximium time threshold

Model for Couriers: Logit Model & CMAB

❑ Short-term decision -- logit Model:

$$\mathcal{P}_i^a(p_{i,t}) = \frac{1}{1 + e^{-(u_{i,t}^a - u_i^r)}} ,$$

- P_i^a: order acceptance probability of courier *i*
- $u_{i,t}^a$: utility of accepting the order, proportional to payment $p_{i,t}$, inversely proportional to reservation value p_i^r
- u_i^r: utility of rejecting the order

❑ Long-term decision -- Contextual Multi-Armed Bandits (CMAB):

$$\tilde{\mathcal{R}}_i^k(\tilde{S}_i^k, \tilde{A}_i^k) = \begin{cases} \sum_{t \in \mathcal{T}} p_{i,t} & \text{if } \tilde{A}_i^k \text{ is opt-in or opt-out} \\ u_0 & \text{if } \tilde{A}_i^k \text{ is exit} \end{cases}$$

- $\tilde{S}_i^k$: information of courier *i* at episode *k*, including the reservation value, initial location, and platform strategy
- $\tilde{A}_i^k$: action of courier *i* at episode *k*, including opt-in, opt-out, and exit
- Opt-in/Opt-out reward: the accumulative income over the entire period (episode)
- Exit reward: a fixed utility (varying among periods)

Solution Methodology for CMAB of Couriers

❑ Thompson sampling method:

- ➢ Summary: Utilize a neural network to estimate the reward distribution for each action within a given state.
- ➢ Utilization: Select the action with the highest sampled reward.
- ➢ Training: Model the reward distribution as a Gaussian distribution and train the network using Maximum Likelihood Estimation (MLE).

$$\log P(r_{i,h}|\mu_{i,h},\sigma_{i,h}) = -\frac{1}{2}\left[\log(2\pi) + \log((\sigma_{i,h})^2) + \frac{(r_{i,h}-\mu_{i,h})^2}{(\sigma_{i,h})^2}\right]$$

$$L_{\text{CMAB}} = \frac{1}{|\mathcal{B}|}\sum_{(\tilde{S}_i,h,r_{i,h})\in\mathcal{B}}\left[\log(\sigma_{i,h}) + \frac{(r_{i,h}-\mu_{i,h})^2}{2(\sigma_{i,h})^2}\right]$$

- h: action index, corresponding to opt-in, opt-out, and exit
- u, σ: mean and standard deviation of the estimated reward distribution, respectively
- B: experience buffer

PPO-KL-CLIP for Platform Payment Setting

- **Original PPO —— unstable chanllenge[4,5]:**
  - PPO-CLIP: If new policy θ have already deviated original policy Θ significantly, the gradient becomes 0, preventing the policy from recovering.
  - PPO-KL: The performance can be influenced by β significantly, necessitating a balance between KL-restriction and policy updates.

$$J^{PPO}_{CLIP}(\pi^P_\theta|\pi^A_\theta) = \mathbb{E}_{\pi^A_\Theta,\pi^p_\Theta}[\min(\frac{\pi_\theta(P_t|S_t)}{\pi_\Theta(P_t|S_t)}\mathcal{A}^{PPO}_{\pi^P_\theta}(S_t,P_t;\theta|\pi^A_\theta), \text{clip}(\frac{\pi_\theta(P_t|S_t)}{\pi_\Theta(P_t|S_t)},1-\epsilon,1+\epsilon)\mathcal{A}^{PPO}_{\pi^P_\theta}(S_t,P_t;\theta|\pi^A_\theta))]\ ,$$

$$J^{PPO}_{KL}(\pi^P_\theta|\pi^A_\theta) = \mathbb{E}_{\pi^A_\Theta,\pi^p_\Theta}[\frac{\pi_\theta(P_t|S_t)}{\pi_\Theta(P_t|S_t)}\mathcal{A}^{PPO}_{\pi^P_\theta}(S_t,P_t;\theta|\pi^A_\theta) - \beta\text{KL}(\pi_\Theta||\pi_\theta)]\ ,$$

- **Our solution: PPO-KL-CLIP**
  - inspired by GRPO[6]
  - add KL normalization to PPO-CLIP, helping recover the deviated policy.

$$J^{PPO}_{KL-CLIP}(\pi^P_\theta|\pi^A_\theta) = \mathbb{E}_{\pi^A_\Theta,\pi^p_\Theta}[\min(\frac{\pi_\theta(P_t|S_t)}{\pi_\Theta(P_t|S_t)}\mathcal{A}^{PPO}_{\pi^P_\theta}(S_t,P_t;\theta|\pi^A_\theta), \text{clip}(\frac{\pi_\theta(P_t|S_t)}{\pi_\Theta(P_t|S_t)},1-\epsilon,1+\epsilon)\mathcal{A}^{PPO}_{\pi^P_\theta}(S_t,P_t;\theta|\pi^A_\theta)) - \beta\text{KL}(\pi_\Theta||\pi_\theta)]\ ,$$

[4] Hsu C C Y, Mendler-Dünner C, Hardt M. Revisiting design choices in proximal policy optimization[J]. arXiv preprint arXiv:2009.10897, 2020.
[5] Engstrom L, Ilyas A, Santurkar S, et al. Implementation Matters in Deep Policy Gradients: A Case Study on PPO and TRPO[C]//International Conference on Learning Representations. 2020.
[6]Shao Z, Wang P, Zhu Q, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models[J]. arXiv preprint arXiv:2402.03300, 2024.

Neural Network Architecture

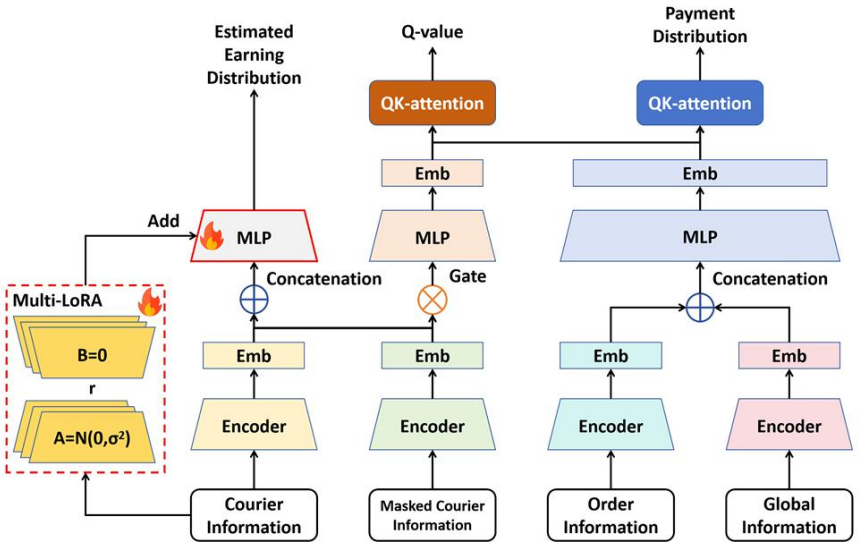


Fig. 2. Neural network architecture: The network is divided into three parts: upstream, downstream, and bypass layers. In the upstream layers, four different encoders are employed to process information from heterogeneous couriers, orders, and the global context. In the downstream layers, MLPs are initially used to fuse features from the upstream network. Subsequently, the QK-Attention module generates the platform's strategy (Q-value and payment distribution), while another MLP estimates the couriers' earnings. Additionally, the Multi-LoRA module serves as a bypass layer to facilitate efficient training of the courier strategy. In the figure, the 'fire' symbol indicates the trainable modules during the training of the courier strategy, while other components remain frozen.

- **Shared heterogeneous encoder**
  - shared structure enhances the extraction of robust features
  - heterogeneous structure facilitates simultaneous processing of opt-in and opt-out couriers.
- **Triple decision heads**
  - for payment setting, order assignment, and courier long-term choice, respectively
- **Multi-LoRA for couriers CMAB**
  - shared encoder enables couriers to understand the platform's strategy
  - multiple lora prevents homogenous strategies among couriers, promoting diverse decision-making

Training Process

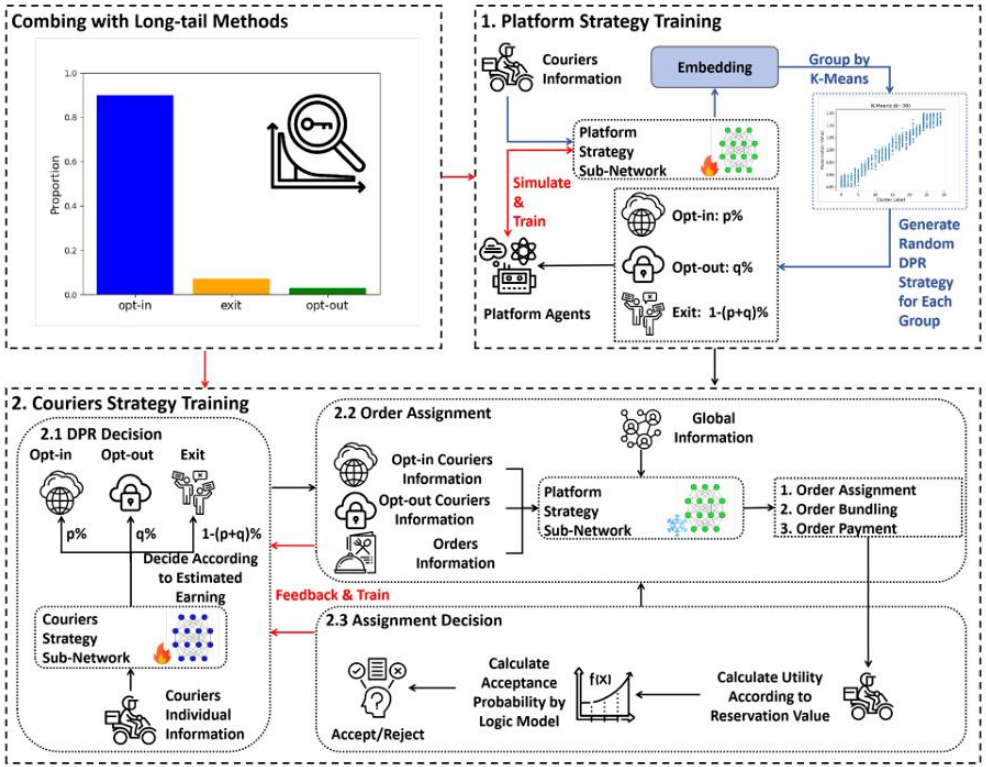


Fig. 3. Training process: The training consists of two stages: platform strategy training and courier strategy training. During platform training, we randomly set the order amount and courier strategies, aiming to enable the platform to respond optimally under various circumstances. In the courier training phase, we fix the optimal strategy learned by the platform and train the couriers to find their own optimal strategies to maximize individual earnings. Additionally, to mitigate the effects of label imbalance–primarily caused by the significant differences in the number of couriers choosing to opt-in and opt-out in each episode–we incorporate long-tail methods into the loss function.

- **Phase1: platform training**
  - train an adaptive policy that optimally responds to all possible decisions made by couriers
- **Phase2: couriers rraining**
  - given the fixed well-trained platform strategy, train couriers to identify the optimal long-term decisions aimed at maximizing cumulative earnings

Case Studies

❑ Food delivery in Hong Kong, China

Table 2
A brief summary of the model parameters.

Configuration	Our Setting
Batch Size	512
Optimizer	Adam
Scheduler	ExponentialLR
Exploration Rate of DDQN ϵ	$0.99 \rightarrow 0.0005$
Learning Rate α	0.0005
Decay Rate of α and ϵ	0.99
Updating Rate of Target Network ψ	0.005
Discount Factor γ	0.9
Number of Potential Courier	1000
Total Number of Parameters	321K
$\beta_1 \sim \beta_5$ in Platform Reward Function (Eq. 12)	[3,5,4,3,1]
$a, b - u^r$ in Courier Logit Model (Eqs. 13 and 14)	50, -0.95

❑ Comparative policies

Option / Policy	DPR	Benchmark	MASK
Opt-in	✓	✓	
Opt-out	✓		✓
Exit	✓	✓	✓

Table 3: Comparison between different policies.

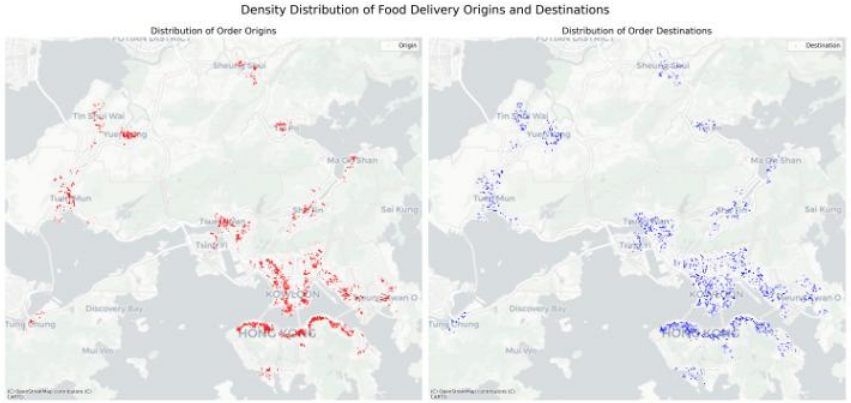


Figure 4: Order spatial distribution map: the red points represent the order origins, while the blue points indicate the destinations.

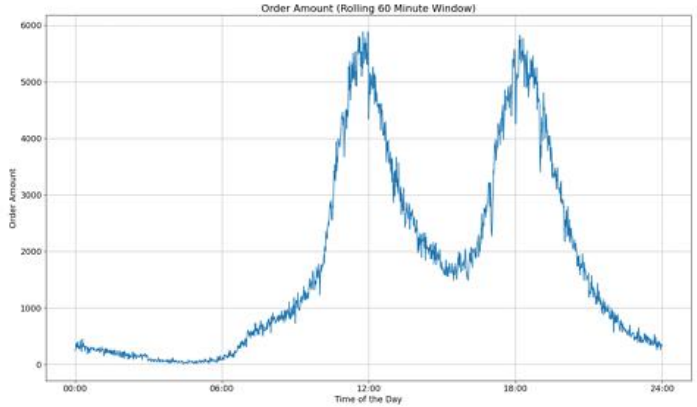
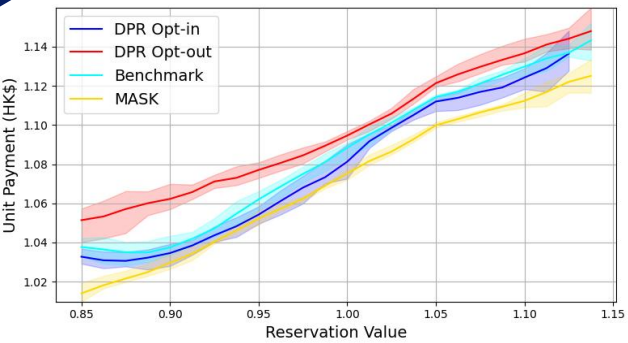
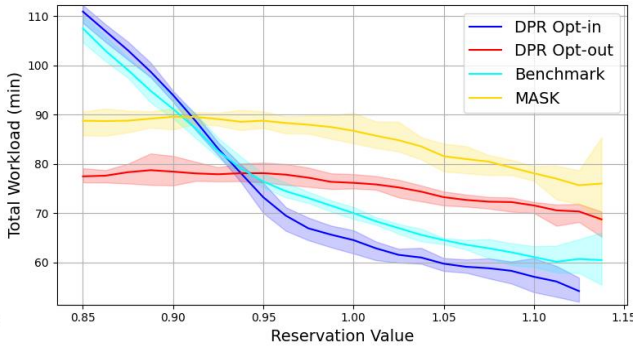


Figure 5: Order Temporal Distribution.

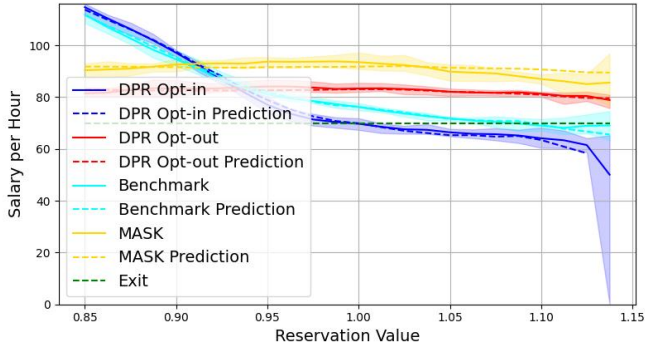
Case Studies



Unit Payment (payment per minute)



Total Workload (measured by time)



Total Earning (whole preiod)

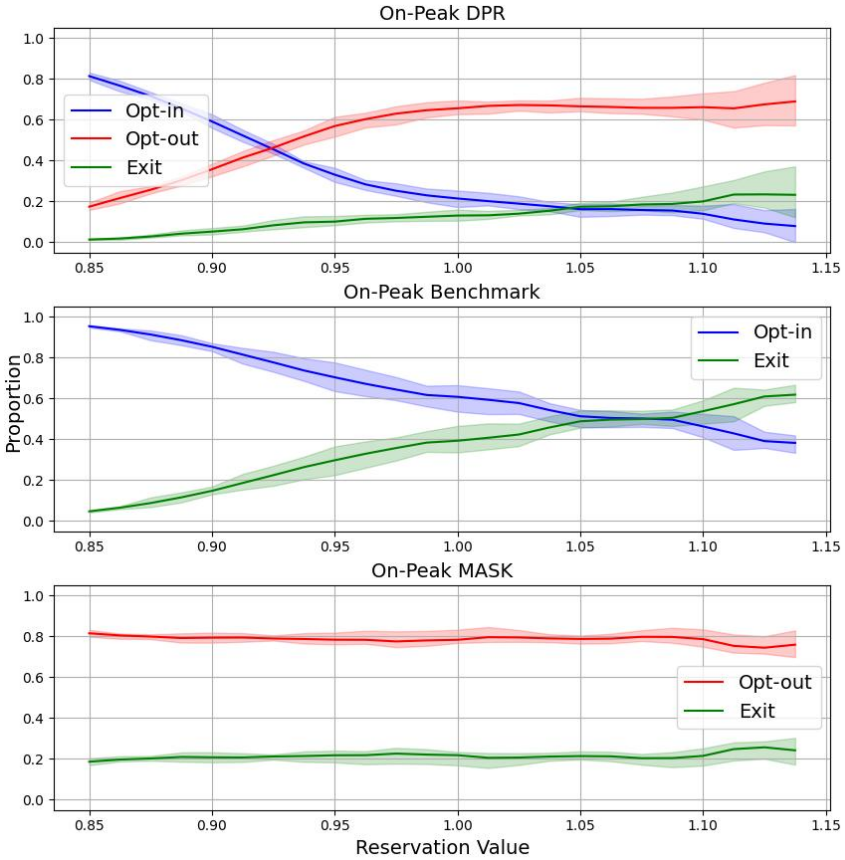
❑ How platforms discriminate different types of couriers?

- ➢ For opt-in couriers (blue curves):
  Higher reservation couriers receive:
  - higher unit payments
  - lower workloads
  - lower total earnings

- ➢ For opt-out couriers (red curves):
  Higher reservation couriers experience:
  - higher unit payments
  - slightly lower workloads
  - similar earnings compaerd to low reservation couriers

Case Studies

- **How DPR affects couriers behavior?**
  - More couriers are attracted to choose opt-out option on the platform, leading to a lower exiting rate.
  - DPR effectively protects the rights of high-reservation couriers, allowing them to choose opt out to access more work opportunities.



Proportion of Choices Among Couriers

Case Studies

- **How DPR affects platform profits?**

Metrics	DPR	Benchmark	Mask
	Off-Peak (1960 orders in total)		
Reward (Profits)	**4.57±0.28**	4.42±0.25	4.31±0.30
Order Pickup	**1921.20±6.20**	1882.9±26.80	1882.60±15.60
Service Rate (%)	**98.02±0.32**	96.06±1.37	96.05±0.80
Platform Payment (10^3)	**48.76±1.12**	54.51±1.83	49.48±1.42
	On-Peak (5880 orders in total)		
Reward (Profits)	**12.31±0.77**	11.17±0.71	11.36±0.72
Order Pickup	**3362.40±43.60**	2679.40±120.40	2910.20±65.80
Service Rate (%)	**57.18±0.74**	45.57±2.05	49.49±1.12
Platform Payment (10^3)	89.07±1.08	**86.58±0.77**	86.91±0.99

Table 5: Comparison of three strategies from the platform's perspective: The bold text indicates the best result for each stakeholder, a convention that will be maintained in the following tables.

- The platform gets higher profits since the increased number of active couriers helps fulfill more orders.

Case Studies

❑ How DPR affects customers satisfication?

Metrics	DPR	Benchmark	Mask
	Off-Peak (1960 orders in total)		
Overtime Order Amount	**100.00±17.00**	117.80±5.80	104.40±15.40
Overtime Rate (%)	**5.21±0.89**	6.26±0.32	5.55±0.80
Average Delivery Time	**19.24±0.29**	20.64±0.40	19.47±0.42
	On-Peak (5880 orders in total)		
Overtime Order Amount	192.40±15.40	**172.60±16.60**	196.60±16.40
Overtime Rate (%)	**5.72±0.53**	6.45±0.73	6.75±0.47
Average Delivery Time	**20.70±0.22**	21.17±0.57	21.15±0.26

Table 6: Comparison of three strategies from the customers' perspective.

➢ Customers get better experience due to the increased number of active couriers, which allows the platform to optimize planning better, leading to a lower order delivery times.

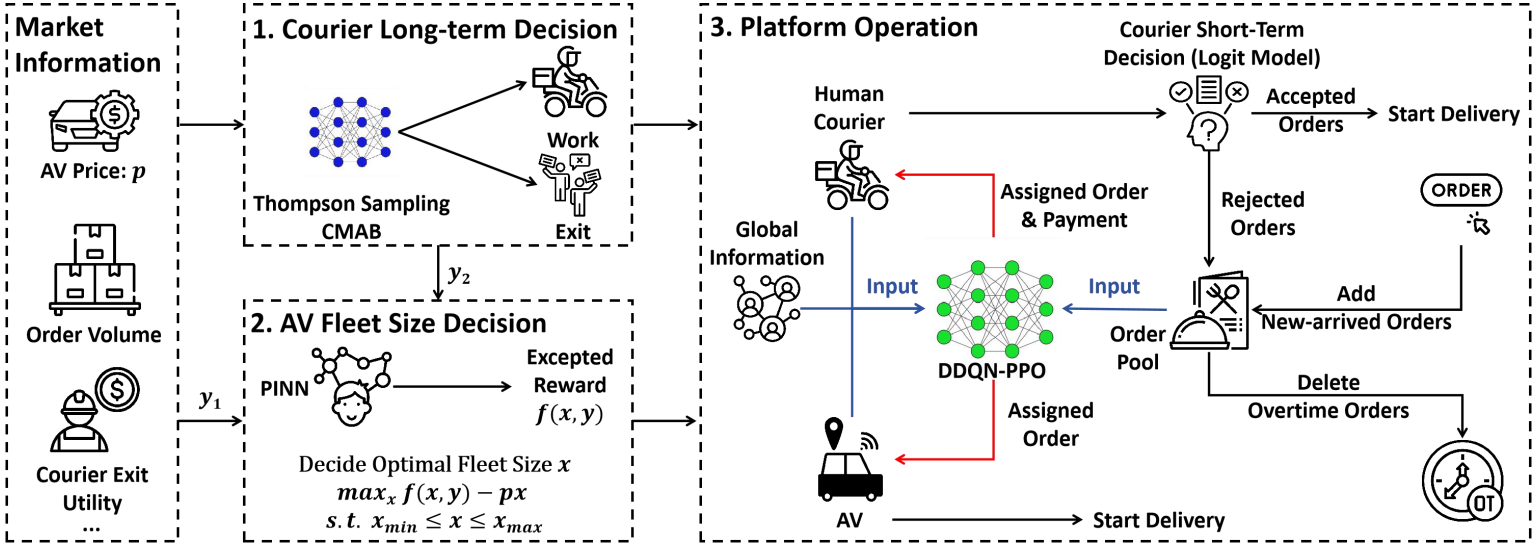
oi

THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

Part IV: Future Works

Future Works

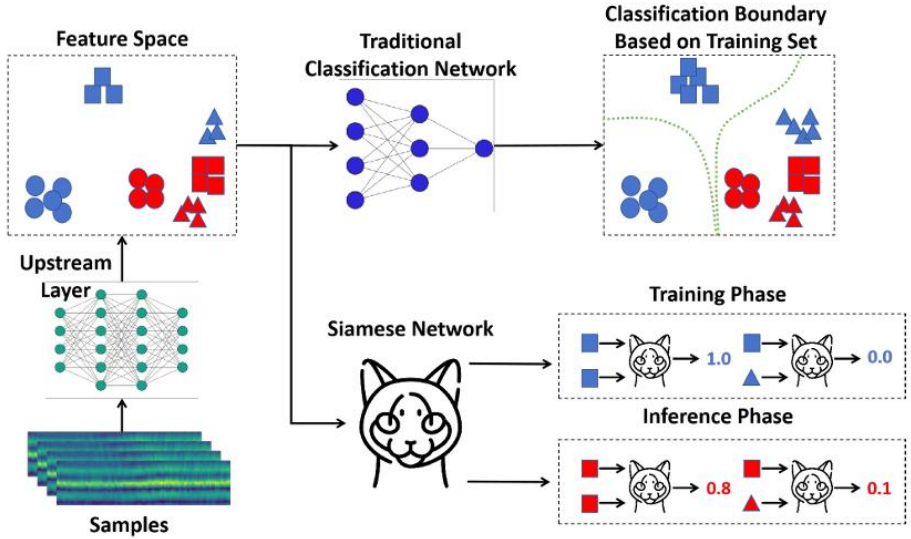
- **Impact of AV in discriminatory food delivery platform** [7]
  - How to determine optimal AV fleet size?
  - What's the impact of AV to human couriers?
  - How might the spatial distribution differ between AVs and human couriers?



[7] Ye A, Zhang K, Bell M G H, et al. Modeling an on-demand meal delivery system with human couriers and autonomous vehicles in a spatial market[J]. Transportation Research Part C: Emerging Technologies, 2024, 168: 104723.

Future Works

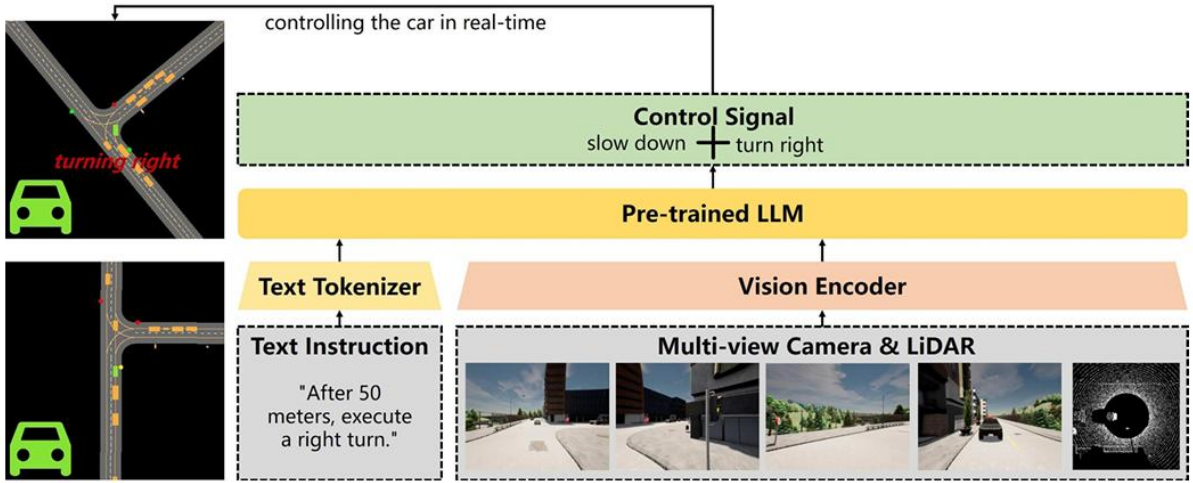
- **Cross-domain challenge in transportation**
  - domain shift in Offline-to-Online (O2O) RL [8]
  - traffic prediction task in cross-domain scenario (cross region & cross time ...) [9]



[8] Ye A, Zhang K, Bell M G H, et al. Modeling an on-demand meal delivery system with human couriers and autonomous vehicles in a spatial market[J]. Transportation Research Part C: Emerging Technologies, 2024, 168: 104723.
[9] Ma J, Wu F. Effective traffic signal control with offline-to-online reinforcement learning[C]//2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2023: 1-7.

Future Works

- **LLM & Transportation**
  - LLM4Transportation[10]: LLM-assisted decision (LLM as encoder & LLM as planner)
  - Transportation Foundation Model[11]: inject transportation knowledge into LLM (RAG, SFT, RL post-training)



[10] Lyu T, Feng S, Liu H, et al. LLM-ODDR: A Large Language Model Framework for Joint Order Dispatching and Driver Repositioning[J]. arXiv preprint arXiv:2505.22695, 2025.
[11] Wang J, Dong Z, Bai B, et al. FoodGPT: Reinforcement Post-Training of Large Language Models in the Food Delivery Domain[C]//Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2. 2025: 4966-4974.

oi

THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

Part V: Conclusion

Concluding Remarks

- **Project 1: Algorithm discrimination of food delivery platform**
  - Developed a Hybrid-Action MARL framework to model platform behavior.
  - Proposed a novel network structure and training strategy for effective policy training.
  - Simulation results indicate the platform favors low-reservation couriers by offering more work opportunities.

- **Project 2: Impact of data privacy regulation to food delivery market**
  - Developed a CMAB framework to model the long-term behavior of couriers under DPR.
  - Introduced MLE-based Thompson Sampling, PPO-KL-CLIP, and a multi-stage training process as methodological solutions.
  - Simulation findings demonstrate that DPR fosters a win-win scenario for platforms, couriers, and customers.

Publication List

❑ Journal

1. **Zijian Zhao**, Sen Li*, "The Impacts of Data Privacy Regulations on Food-Delivery Platforms", Transportation Research Part C: Emerging Technologies (TR_C), 2025
2. **Zijian Zhao**, Sen Li*, "Discriminatory Order Assignment and Payment-Setting of On-Demand Food-Delivery Platforms: A Multi-Action and Multi-Agent Reinforcement Learning Framework" (under thrid-round review, Transportation Research Part E: Logistics and Transportation Review (TR_E))
3. **Zijian Zhao**, Sen Li*, "The Impacts of Automatic Vehicle on Discriminatory Food Delivery Platform: A Multi-Stage Framework via Hybrid Learning Technologies" (in preparation)

❑ Conference

1. **Zijian Zhao**, "Towards Fairness in Transportation Gig Markets: Identifying, Imitating, and Mitigating Algorithm Discrimination via Deep Reinforcement Learning", 2026 Annual AAAI Conference on Artificial Intelligence (AAAI) / Special Interest Group on Artificial Intelligence (SIGAI) Doctoral Consortium, 2026
2. **Zijian Zhao**, Jing Gao*, Sen Li, "Ride-Hailing Order Dispatching with A Mixture of On-Demand and Pre-Booked Requests via Reinforcement Learning", 2026 COTA International Conference of Transportation Professionals (CICTP), 2026
3. **Zijian Zhao**, Sen Li*, "Multi-Agent Reinforcement Learning for Order Assignment and Payment Setting on Food-Delivery Platforms: The Implicit Algorithmic Biases", 2025 International Symposium on Transportation Data & Modelling (ISTDM), 2025
4. **Zijian Zhao**, Sen Li*, "Triple-BERT: Do We Really Need MARL for Ride-Sharing Order Dispatch?" (under review, 2026 International Conference on Learning Representations (ICLR))
5. **Zijian Zhao**, Sen Li*, "One Step is Enough: Multi-Agent Reinforcement Learning Based on One-Step Policy Optimization for Order Dispatch on Ride-Sharing Platforms" (under revision, to be submitted to 2026 International Conference on Machine Learning (ICML))
6. **Zijian Zhao**, Yitong Shang, Sen Li*, "AutoFed: A Personalized Federated Travel Demand Prediction Framework via Global Representation" (in preparation, to be submitted to 2026 International Conference on Machine Learning (ICML))

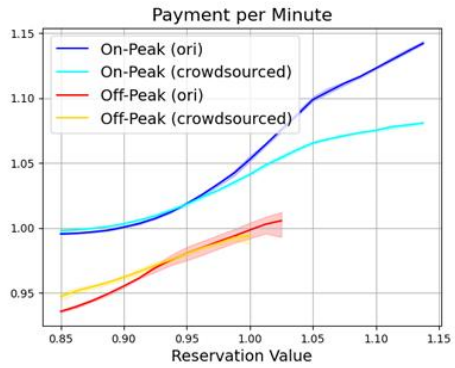
香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

Thanks for your attention!

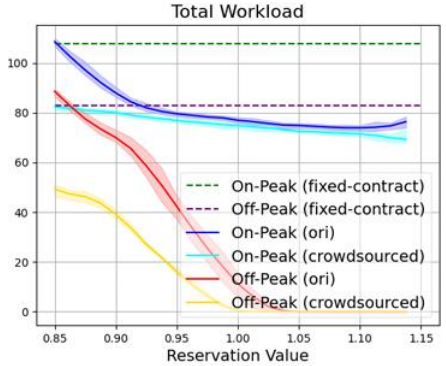
Presented by Zijian Zhao
zzhaock@connect.ust.hk

Case Studies

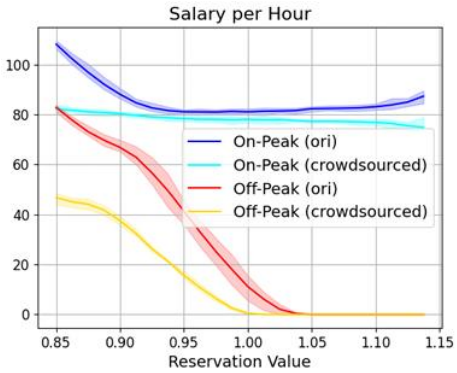
- **Comparison between crowdsourced-only market and mixed market**



(e) Payment per minute for the orders assigned to the courier



(d) Total trip time (minutes) of all order assigned to the courier per hour



(f) Salary per hour for the courier

- **What if we further introduce fixed-contract couriers?**
  - Fixed-contract couriers receive the highest preference.
  - Crowdsourced couriers get less workload and earning.