HONOR: O-RAN-compliant Handover Optimization in Vehicular Networks via Offline RL

Yizhou Wang*, Yushan Yang*, Mingyu Ma*†, Mahdi Attawna‡, Frank H.P. Fitzek†‡, Giang T. Nguyen*†
*Haptic Communication Systems, TU Dresden, Germany
†Centre for Tactile Internet with Human-in-the-Loop (CeTI)
‡Deutsche Telekom Chair of Communication Networks, TU Dresden, Germany
E-mails: yizhou.wang@mailbox.tu-dresden.de,
{yushan.yang | mingyu.ma | mahdi.attawna | frank.fitzek | giang.nguyen}@tu-dresden.de

***Abstract*—Safety-critical vehicle-to-everything (V2X) services require stringent latency and reliability quality of service (QoS) under fast channel dynamics and dense traffic. In millimeter wave (mmWave) vehicular networks, analog beamforming mitigates path loss but enforces beam exclusivity, which couples multi-vehicle association decisions and complicates efficient handover. This paper proposes *Handover Optimization in Vehicular Networks via Offline Reinforcement Learning* (HONOR), an O-RAN-compliant handover xApp. Specifically, we formulate multi-vehicle handover as a Constrained Markov Decision Process (CMDP) subject to latency and reliability and handle the constraints via Lagrangian relaxation. To enforce one-to-one matching, an assignment-constrained factorized deep Q-learning network with validity masking is designed to obtain a feasible association via assignment-based action selection. Using a fixed dataset generated by O-RAN-integrated ns-3 simulations, the policy is learned offline and implemented as the HONOR xApp. Extensive evaluations with multiple random seeds show that HONOR reduces handover frequency by 18%–73% and suppresses ping-pong events by 41%–81% relative to representative baselines, while improving latency and slightly increasing packet delivery reliability.**

***Index Terms*—handover, V2X, mmWave, offline reinforcement learning, O-RAN, ns-3.**

I. Introduction

Vehicle-to-everything (V2X) communication is a key paradigm for enabling real-time data exchange between vehicles and roadside infrastructure [1]. By facilitating this interconnected ecosystem, V2X promises a future defined by autonomous driving and intelligent transportation systems that can fundamentally reshape road traffic safety and efficiency [2]. However, realizing V2X vision requires a communication substrate capable of stringent quality of service (QoS). Specifically, these safety-critical applications require the underlying cellular networks to deliver ultra-low end-to-end latency and near-perfect packet delivery reliability, even amidst time-varying traffic loads inherent to high-mobility environments [3–5].

Current cellular networks, including 5G, struggle with consistently satisfying stringent latency and reliability requirements for V2X applications under fast channel dynamics. While 5G supports operation in mmWave bands to access wider bandwidth [6, 7], this alone does not guarantee service-level performance due to blockage sensitivity and high-frequency propagation losses. To cope with the increased path loss, beamforming is commonly used to concentrate signal power into narrow, directional beams, which unintentionally introduces complex resource coupling, restricting a beam to serving a vehicle (i.e. one-to-one roadside unit (RSU)-vehicle matching) simultaneously [8, 9]. This limitation incurs the fierce competition for RSU resources among vehicles, especially when a RSU employs analog beamforming and can form only one beam at a time. In multi-vehicle scenarios, a handover or association decision for one vehicle directly affects the resource availability of others, calling for system-wide joint optimization driven by real-time measurements. However, such fast-timescale closed-loop control is difficult to implement in conventional monolithic radio access network (RAN) due to limited openness, intelligence, and programmability of the control plane.

Open RAN (O-RAN) represents a revolutionary shift for future cellular networks, transitioning from closed, monolithic systems to a disaggregated and fully programmable framework [10–12]. It brings unprecedented intelligence to the network edge through the near real-time RAN intelligent controller (near-RT RIC), which supports customized xApps for near-real-time radio resource management. By leveraging E2-compliant measurements and interfaces, O-RAN enables data-driven closed-loop control [10, 13]. Despite these architectural advantages, comprehensive and deployable O-RAN-based handover optimization for multi-vehicle scenarios remains under explored. Early solutions employing conventional threshold-based strategies lack this holistic coordination, tend to trigger unnecessary handovers and ping-pong effects, increasing handover overhead and causing transient disruptions that degrade latency and reliability QoS [14]. This motivates a deployable handover xApp that fuses high-dimensional real-time context, including mobility dynamics, beam exclusivity, and traffic load, to balance these competing objectives under latency and reliability QoS constraints. To avoid risky online exploration, we learn the policy offline from logged traces generated by existing handover rules. This design also enables iterative improvement by retraining on newly collected logs.

In this paper, we propose HONOR, a handover xApp based on offline reinforcement learning (RL) for mmWave vehicular networks, which minimizes handover frequency under multi-vehicle contention while satisfying latency and reliability QoS

ISBN 978-3-903176-82-9

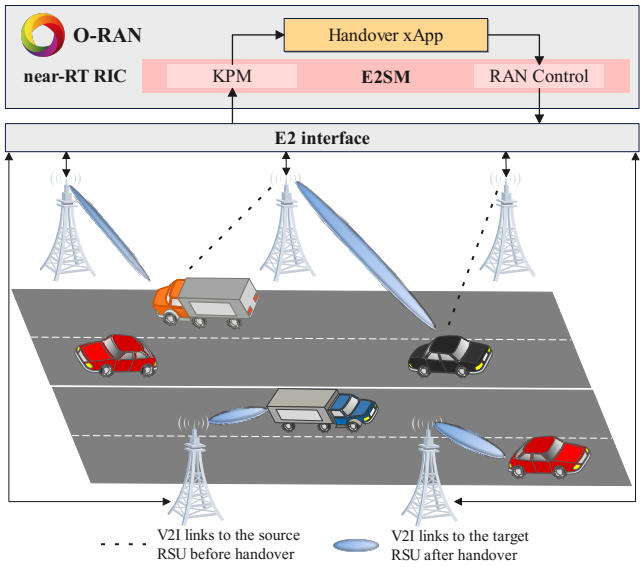

Fig. 1: System description.

constraints. Our main contributions are: (i) We formulate the multi-vehicle handover in mmWave networks as a CMDP subject to latency and reliability under beam-exclusivity-induced coupling, and relax it via a Lagrangian approach to enable learning-based optimization; (ii) An assignment-constrained factorized deep Q-learning network (DQN) with validity masking is designed to jointly determine RSU-vehicle association with one-to-one matching; (iii) For implementation, an agent is trained offline and deployed on the near-RT RIC as a handover xApp termed HONOR; (iv) We conduct extensive multi-vehicle mmWave network simulations with multiple random seeds in ns-3 integrated with O-RAN, and the results show HONOR consistently reduces handover frequency by 18%–73% and suppresses ping-pong events by 41%–81% compared with representative baselines, while improving latency and slightly increasing packet delivery reliability.

II. System Model and Problem Formulation

In this section, we describe the considered beamforming-enabled vehicular network under the O-RAN architecture and specify the vehicle–RSU association and handover model. Based on these definitions, we formulate a long-term optimization problem that minimizes the handover frequency subject to QoS constraints.

A. System Model

Fig. 1 depicts a beamforming-enabled vehicular network in which RSUs are uniformly deployed on both sides of the roadway, and vehicles travel in both directions. An xApp running on a near-RT RIC manages the handover of all vehicles in this area. Time is slotted with index $t = 0, 1, 2, \ldots$. In slot t, each vehicle $v \in \mathcal{V}(t)$ is associated with one serving RSU selected from $\mathcal{U}$. Each RSU is assumed to employ analog beamforming and thus can serve at most one vehicle per slot. Let a binary $x_{v,u}(t)$ denote the association indicator, where $x_{v,u}(t) = 1$ if vehicle v is associated with RSU u in slot t, and $x_{v,u}(t) = 0$ otherwise.

At the beginning of each slot, the near-RT RIC gathers the performance metrics from each RSU via key performance measurement (KPM) report service [15], and forwards the relevant information to the handover xApp. Based on the information, the xApp decides on handover and executes the decision through RAN control (RC) service [16]. Let $H_v(t) \in \{0, 1\}$ denote the handover indicator of vehicle v at slot t, where $H_v(t) = 1$ if the serving RSU of vehicle v differs from that in slot $t-1$, and $H_v(t) = 0$ otherwise. In slot $t = 0$, vehicle v will establish a connection with a RSU, so no handover event will occur, which is denoted as $H_v(0) = 0$. Therefore, the per-vehicle average number of handovers during slot t is presented as $\bar{H}(t) = \frac{\sum_{v\in\mathcal{V}(t)} H_v(t)}{|\mathcal{V}(t)|}$, where $|\mathcal{V}(t)|$ denotes the number of active vehicles in the vehicular network in slot t. In all considered scenarios, at least one vehicle is in each slot, which means $|\mathcal{V}(t)| \geq 1, \forall t \geq 0$.

B. Problem Formulation

To mitigate the overhead induced by redundant handovers, the goal is to minimize the average handover frequency per vehicle over the long term, subject to QoS constraints on latency and reliability. In slot t, the latency and reliability of vehicle v are denoted as $L_v(t)$ and $P_v(t)$, respectively. $L_v(t)$ is measured by the downlink user-plane latency, whereas $P_v(t)$ is measured by packet delivery rate (PDR). For all vehicles, the maximum latency tolerance is $L_{\max}$ and the minimum reliability tolerance is $P_{\min}$ Therefore, the constrained optimization problem is formulated as

$$\min_{x_{v,u}(t)\in\{0,1\}} \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{\sum_{v\in\mathcal{V}(t)} H_v(t)}{|\mathcal{V}(t)|} \tag{1a}$$

$$\text{s.t.} \quad \sum_{u\in\mathcal{U}} x_{v,u}(t) = 1, \ \forall t \geq 0, \ \forall v \in \mathcal{V}(t) \tag{1b}$$

$$\sum_{v\in\mathcal{V}(t)} x_{v,u}(t) \leq 1, \ \forall t \geq 0, \ \forall u \in \mathcal{U} \tag{1c}$$

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{\sum_{v\in\mathcal{V}(t)} L_v(t)}{|\mathcal{V}(t)|} \leq L_{\max}, \tag{1d}$$

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{\sum_{v\in\mathcal{V}(t)} P_v(t)}{|\mathcal{V}(t)|} \geq P_{\min}, \tag{1e}$$

where constraint (1b) ensures that each vehicle is associated with exactly one RSU per slot, whereas constraint (1c) restricts each RSU to serving at most one vehicle per slot under the assumed analog beamforming setting. Constraints (1d) and (1e) impose the QoS requirements on latency and reliability in the long term, respectively.

Due to the long-term average constraints and the binary one-to-one association variables, the optimization problem in (1) is a mixed-integer and generally nonconvex optimization, which is computationally prohibitive to solve directly. Therefore, we employ the Lagrangian duality method to decouple the constraints. By introducing Lagrangian relaxation to incorporate the latency and reliability constraints into the objective function, we transform the original constrained optimization into an unconstrained max-min dual problem, which is given by (2),

$$\max_{\lambda\geq 0,\mu\geq 0}\min_{x_{v,u}(t)\in\{0,1\}}\limsup_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\Big[\frac{\sum_{v\in\mathcal{V}(t)}H_v(t)}{|\mathcal{V}(t)|}+\lambda\Big(\frac{\sum_{v\in\mathcal{V}(t)}L_v(t)}{|\mathcal{V}(t)|}-L_{\max}\Big)+\mu\Big(P_{\min}-\frac{\sum_{v\in\mathcal{V}(t)}P_v(t)}{|\mathcal{V}(t)|}\Big)\Big] \tag{2a}$$

$$\text{s.t.}\quad \sum_{u\in\mathcal{U}}x_{v,u}(t)=1,\ \forall\, t\geq 0,\ \forall\, v\in\mathcal{V}(t) \tag{2b}$$

$$\sum_{v\in\mathcal{V}(t)}x_{v,u}(t)\leq 1,\ \forall\, t\geq 0,\ \forall\, u\in\mathcal{U}. \tag{2c}$$

where λ and μ are nonnegative multipliers for the latency and reliability constraints associated with (1d) and (1e), respectively.

III. Handover optimization via Offline RL

Since the optimization problem (2) can be viewed as a sequential decision process operating in decision epochs indexed by time slots, it can formulated as a Markov Decision Process (MDP). Based on the formulation, we design an assignment-constrained factorized DQN, and train it offline, particularly in conservative Q-learning (CQL) method. The Lagrangian multipliers are updated during training. The detail is presented as follows.

A. MDP Formulation

We model the handover association procedure as a discrete-time sequential decision process, where each interaction at slot t is represented by a transition tuple (S, A, R, S', d), where S is the state of the vehicular network, A is the action indicating the association between vehicles and RSUs, R is the reward caused by the action A in State S, S' denotes the next state of S, and d indicates whether a vehicle has entered or left the vehicular network. Taking slot t as an example, the elements in the tuple are detailed as follows:

1) State: The state $S(t) = \{S_u(t)\,|\,u \in \mathcal{U}\}$ is used to characterize the dynamic vehicular network environment and is constructed from telemetry collected at all RSUs via KPM reports generated at the beginning of slot t. Each component $S_u(t)$ summarizes the measurements reported by RSU u, including (i) the signal-to-interference-plus-noise ratio (SINR) between the RSU u and its serving vehicle as well as the SINR between the vehicle and other RSUs; (ii) the user-plane latency experienced by the vehicle in the previous slot; (iii) the number of transport blocks scheduled for initial transmission and retransmission from RSU u to the vehicle in the previous slot; (iv) the number of transport blocks transmitted using QPSK, 16QAM and 64 QAM modulation rates from RSU u to the vehicle in the last slot. If RSU u is not associated with any vehicle, we set $S_u(t) = \mathbf{0}$.

2) Action: Let $A(t) = \{A_u(t)\,|\,u \in \mathcal{U}\}$ denote the RSU-side handover action determined by the xApp at the beginning of slot t. For a given RSU $u \in \mathcal{U}$, three cases are considered: (i) $A_u(t) = 0$, indicating that RSU u is idle at the beginning of slot t and thus does not initiate any handover action; (ii) $A_u(t) = u$, indicating that RSU u maintains its current association during slot t, i.e., no handover is triggered; (iii) $A_u(t) = n$ with $n \in \mathcal{U}$ and $n \neq u$, indicating that a handover is triggered at the beginning of slot t and the vehicle currently served by RSU u is reassociated to the target RSU n for slot t.

3) Reward: The relaxed objective in (2a) minimizes the long-term time-average cost consisting of the handover term and multiplier-weighted QoS-related penalties, whereas RL maximizes the expected cumulative reward. A direct reward definition as the negative instantaneous Lagrangian cost would assign positive rewards when the latency and reliability constraints are satisfied with slack, since the corresponding penalties $\frac{\sum_{v\in\mathcal{V}(t)}L_v(t)}{|\mathcal{V}(t)|}-L_{\max}$ and $P_{\min}-\frac{\sum_{v\in\mathcal{V}(t)}P_v(t)}{|\mathcal{V}(t)|}$ become negative. These slack-induced rewards may encourage overly conservative behavior by rewarding constraint slack, even though only constraint satisfaction is required. To ensure that constraint slack is not rewarded, we penalize only violations by using the positive-part operator $[\cdot]^+ \triangleq \max\{\cdot, 0\}$ [17]. Accordingly, the per-slot reward is defined as

$$\begin{aligned}R(t) =& -\Big[\frac{\sum_{v\in\mathcal{V}(t)}H_v(t)}{|\mathcal{V}(t)|}+\lambda\Big[\frac{\sum_{v\in\mathcal{V}(t)}L_v(t)}{|\mathcal{V}(t)|}-L_{\max}\Big]^+\\&+\mu\Big[P_{\min}-\frac{\sum_{v\in\mathcal{V}(t)}P_v(t)}{|\mathcal{V}(t)|}\Big]^+\Big].\end{aligned} \tag{3}$$

The reward is observed at the end of slot t after executing action $A(t)$. The nonnegative multipliers λ and μ are updated during training to balance the handover objective and the QoS constraints, and the update procedure is provided in Sec. III-C.

4) Terminal Flag: We use a terminal indicator $d(t) \in \{0, 1\}$ to mark whether the transition at slot t terminates a trajectory segment. In particular, we set $d(t) = 1$ at trace boundaries, including the end of the mobility trace and any slot where the active vehicle set changes between t and $t + 1$ due to vehicle arrivals or departures, $d(t) = 0$ otherwise. This prevents bootstrapping across discontinuous system configurations.

B. Deep Q-learning Network

Building upon the above MDP formulation, we resort to RL to learn a policy that maps the observed state to per-slot association decisions. Since the action space is discrete and decisions must be produced with low inference latency in the near-RT RIC, we adopt a value-based approach and implement it using DQN.

However, directly applying an off-the-shelf monolithic DQN to enumerate association actions is impractical, as the feasible action space grows combinatorially with the number of active links. Moreover, the one-to-one association constraint couples concurrent decisions through resource competition, which necessitates globally consistent inference. Therefore, we design an assignment-constrained factorized DQN to account for the

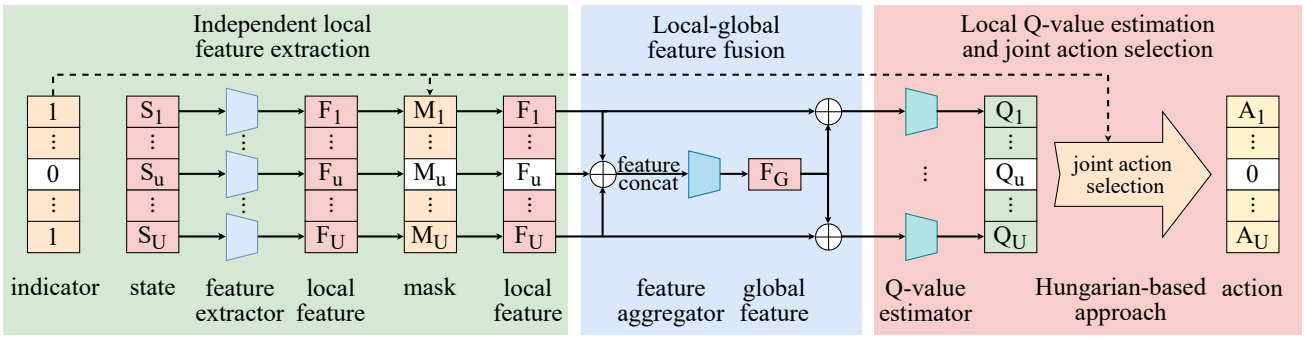

Fig. 2: Our DQN architecture design. The indicator is a binary validity mask: 1 denotes an active RSU serving a vehicle, and 0 denotes an inactive RSU, for which all associated state features and action options are masked as invalid.

decision coupling induced by the one-to-one association constraint, while avoiding explicit enumeration of the combinatorial joint action space. As shown in Fig. 2, the proposed DQN can be decomposed into three functional components, which are detailed below.

1) Independent Local Feature Extraction: The network takes as input the per-RSU state component S_u together with a binary validity indicator $m_u \in \{0, 1\}$ that specifies whether RSU u is active in the current slot. When $m_u = 0$, all state features and action options associated with RSU u are masked as invalid. Each S_u is processed by an independent local feature extractor to produce a local feature embedding F_u. The embedding is subsequently gated by m_u so that an inactive RSU yields an all-zero feature vector, preventing idle branches from injecting invalid information into the downstream.

2) Local-global Feature Fusion: We concatenate the masked embeddings of all RSUs into a fixed-length vector. Inactive branches contribute zero blocks after masking, which preserves dimensional consistency and removes irrelevant information. The concatenated vector is fed into an aggregation network to generate a global feature. This global feature is broadcast to each RSU branch and concatenated with its masked local embedding for Q-value estimation.

3) Local Q-value Estimation and Joint Action Selection: For each RSU, the corresponding Q-value estimator outputs an action-value vector over candidate association actions based on the fused local-global feature embedding, while the mask disables inactive branches so that only active ones contribute. To satisfy the one-to-one association constraint without enumerating the combinatorial joint action space, we select the per-slot joint action by solving an assignment problem that maximizes the aggregated estimated Q-value, which is efficiently computed via the Hungarian algorithm. The resulting assignment yields a feasible and globally consistent joint action with tractable inference complexity for near-RT RIC deployment.

Therefore, the system-wide Q-value is computed as the sum of masked per-RSU Q-values

$$Q^{\theta}(S, A) = \sum_{u \in \mathcal{U}} m_u Q_u(S, A_u), \tag{4}$$

where $Q_u(\cdot)$ denotes the per-RSU action-value function output by the corresponding DQN branch, and θ collects the trainable parameters of the proposed structured DQN. According to the Bellman equation, the DQN is trained by minimizing the temporal difference (TD) loss

$$\mathcal{L}_{\mathrm{TD}}(\theta) = \mathbb{E}\big[\big(Q^{\theta}(S, A) - y\big)^2\big], \tag{5}$$

where $y(t)$ is the TD target and the expectation is taken over transitions sampled for training. In practice, we adopt double DQN (DDQN) to improve training stability [18]. Specifically, we maintain an online network with parameters θ for action selection and a target network with parameters θ^- for target evaluation. The target network is updated by soft updates rather than a hard delayed copy, so that θ^- evolves smoothly and provides more stable targets. Accordingly,

$$y = R + \gamma(1 - d)\, Q^{\theta^-}\big(S', \arg\max_{A' \in \mathcal{A}(S')} Q^{\theta}(S', A')\big). \tag{6}$$

where $\gamma \in [0, 1)$ is the discount factor, $\mathcal{A}(S')$ denotes the set of feasible joint association actions at state S' under the one-to-one association constraint, and S' represents the next state of S.

C. Offline Training with Handover-triggered Sampling and Weighted Loss

To ensure safe deployment and avoid impairments induced by online exploration, the proposed RL agent is trained offline in accordance with the O-RAN artificial intelligence (AI)/ machine learning (ML) workflow [13]. Accordingly, we construct an offline dataset $\mathcal{D}$ consisting of transition tuples generated from V2X simulations. The data collection procedure is detailed in Sec. IV-B. Using $\mathcal{D}$, we present the offline training objective and the corresponding updates as follows.

1) Offline Objective with Conservative Q-learning: In offline training with a fixed dataset $\mathcal{D}$, the maximization in the TD target can select out-of-distribution (OOD) actions, which leads to overestimated Q-values. Therefore, we adopt CQL to inhibit such overestimation by combining the TD loss with a conservative regularization [19]. The total loss function is denoted as

$$\mathcal{L}(\theta) = \mathcal{L}_{\mathrm{TD}}(\theta) + \alpha \mathcal{L}_{\mathrm{CQL}}(\theta), \tag{7}$$

where $\alpha > 0$ controls the strength of the regularization, and the corresponding loss $\mathcal{L}_{\mathrm{CQL}}(\theta)$ is defined as

$$\begin{aligned}\mathcal{L}_{\mathrm{CQL}}(\theta) =& \mathbb{E}_{S\sim\mathcal{D}}\big[\log \sum_{A'\in\mathcal{A}(S)} \exp\big(Q^\theta(S,A')\big) \\ &- \mathbb{E}_{A\sim\mathcal{D}(\cdot\,|\,S)}[Q^\theta(S,A)]\big].\end{aligned} \tag{8}$$

This term encourages conservative Q-values for actions that are unlikely under the dataset policy, which improves robustness to distribution shift in offline learning.

2) Handover-aware Sampling: Handover events are sparse in $\mathcal{D}$. As a result, too few handover-related transitions can be acquired through uniform sampling, which limits the learning of handover decisions. To mitigate this imbalance, we partition $\mathcal{D}$ into a handover subset $\mathcal{D}_{\mathrm{H}}$ and a non-handover subset $\mathcal{D}_{\mathrm{nH}}$, and increase the sampling ratio of $\mathcal{D}_{\mathrm{H}}$ during training. Specifically, for training, each transition tuple in the batch $\mathcal{B}(i)$ is drawn from $\mathcal{D}_{\mathrm{H}}$ with probability $\rho(i)$ and from $\mathcal{D}_{\mathrm{nH}}$ with probability $1-\rho(i)$, where $i = 0, 1, \ldots, I-1$ indicates the training step index. We schedule $\rho(i)$ using a cosine annealing rule that decays from an initially large value to the original dataset distribution $\frac{|\mathcal{D}_{\mathrm{H}}|}{|\mathcal{D}|}$. This schedule intentionally biases the batch composition toward $\mathcal{D}_{\mathrm{H}}$ in the early training period, which increases the exposure to scarce handover transitions and accelerates the learning of handover decisions. As training progresses, $\rho(i)$ decreases, and the sampling distribution approaches the dataset distribution, thereby limiting long-term sampling bias and avoiding overemphasizing handover transitions.

3) Handover-weighted Loss: Although handover-aware sampling increases the occurrence of handover transitions during training, optimizing the expectation-based loss can still bias the updates toward the abundant non-handover transitions. Therefore, we apply handover-aware weights to emphasize decision-critical transitions in both the TD and the CQL losses. For a specific transition $\tau = (S, A, R, S', d)$, we assign a handover-aware weight $w(\tau) = 1 + \delta\,\beta(\tau)$, where $\beta(\tau) \in \{0, 1\}$ indicates whether τ is a handover transition, with $\beta(\tau) = 1$ if $\tau \in \mathcal{D}_{\mathrm{H}}$ and $\beta(\tau) = 0$ otherwise, and $\delta > 0$ controls the additional emphasis on handover transitions. Given the batch $\mathcal{B}(i)$,the normalized weighted TD loss and CQL loss are respectively denoted as

$$\begin{aligned}\mathcal{L}^{\mathrm{w}}_{\mathrm{TD}}(\theta) &= \frac{1}{N_i}\sum_{\tau\in\mathcal{B}(i)} w(\tau)\big(Q^\theta(S,A) - y\big)^2, \\ \mathcal{L}^{\mathrm{w}}_{\mathrm{CQL}}(\theta) &= \frac{1}{N_i}\sum_{\tau\in\mathcal{B}(i)} w(\tau)\big[\log \sum_{A'\in\mathcal{A}(S)} \exp\big(Q^\theta(S,A')\big) \\ &\quad - \mathbb{E}_{A\sim\mathcal{D}(\cdot|S)}[Q^\theta(S,A)]\big].\end{aligned} \tag{9}$$

where $N_i = \sum_{\tau\in\mathcal{B}(i)} w(\tau)$. Accordingly, the weighted total loss is

$$\mathcal{L}^{\mathrm{w}}(\theta) = \mathcal{L}^{\mathrm{w}}_{\mathrm{TD}}(\theta) + \alpha\,\mathcal{L}^{\mathrm{w}}_{\mathrm{CQL}}(\theta). \tag{10}$$

D. Lagrangian Multiplier Update

The multipliers λ and μ act as adaptive weights for the aggregated QoS constraint residuals in (2a) and (3). They are updated via projected dual subgradient ascent using batch residual estimates, so that persistent violations increase the penalties in subsequent updates, while sustained slack reduces them. Since handover-aware sampling changes the batch distribution, we update λ and μ using an auxiliary batch $\widetilde{\mathcal{B}}(i)$ sampled from $\mathcal{D}$ under the original dataset distribution. At step i, we estimate the constraint residuals as

$$\begin{aligned}\widehat{g}_L(i) &= \mathbb{E}_{\tau\sim\widetilde{\mathcal{B}}(i)}\Big[\frac{\sum_{v\in\mathcal{V}^\tau} L_v^\tau}{|\mathcal{V}^\tau|} - L_{\max}\Big], \\ \widehat{g}_P(i) &= \mathbb{E}_{\tau\sim\widetilde{\mathcal{B}}(i)}\Big[P_{\min} - \frac{\sum_{v\in\mathcal{V}^\tau} P_v^\tau}{|\mathcal{V}^\tau|}\Big]\end{aligned} \tag{11}$$

where $\mathcal{V}^\tau$ is the active vehicle set associated with the transition τ, L_v^τ and P_v^τ The multipliers are updated by

$$\begin{aligned}\lambda(i{+}1) &= [\lambda(i) + \eta_\lambda \widehat{g}_L(i)]^+, \\ \mu(i{+}1) &= [\mu(i) + \eta_\mu \widehat{g}_P(i)]^+,\end{aligned} \tag{12}$$

where $\eta_\lambda > 0$ and $\eta_\mu > 0$ are dual stepsizes.

The handover-aware offline RL training with Lagrangian multiplier update is summarized in Algorithm 1.

Algorithm 1 Offline Training with Dual Multiplier Update

1: Construct the dataset $\mathcal{D}$ consisting of transition tuples and partition it into handover subset $\mathcal{D}_{\mathrm{H}}$ and non-handover subset $\mathcal{D}_{\mathrm{nH}}$.
2: Randomly initialize the online and target RL networks with parameters θ and $\theta^- \leftarrow \theta$, respectively.
3: Initialize the soft update weight ξ and discount γ for TD loss, weight α for CQL loss, weight δ for an transition involving handover, initial sampling ratio $\rho_{\max}$, initial multipliers λ_0, μ_0 and corresponding dual stepsizes η_λ, η_μ.
4: **for** $i = 0, 1, \ldots, I-1$ **do**
5: Compute sampling ratio $\rho(i)$ by cosine annealing from $\rho_{\max}$ to $\frac{|\mathcal{D}_{\mathrm{H}}|}{|\mathcal{D}|}$
6: Acquire batch $\mathcal{B}(i) \leftarrow \rho(i)\mathcal{D}_{\mathrm{H}} + (1-\rho(i))\mathcal{D}_{\mathrm{nH}}$.
7: Sample auxiliary batch $\widetilde{\mathcal{B}}(i)$ from $\mathcal{D}$
8: Update the online RL network θ by minimizing Eq. (10) with $\mathcal{B}(i)$
9: Update the target RL network: $\theta^- \leftarrow \xi\theta + (1-\xi)\theta^-$.
10: Calculate the constraint residuals based on $\widetilde{\mathcal{B}}(i)$ via Eq. (11), and update Lagrangian multipliers via Eq. (12)
11: **end for**

IV. Implementation and Performance Evaluation

After offline training, we encapsulate the learned policy as a handover xApp termed HONOR and validate it via an end-to-end, trace-driven evaluation in a beamforming-enabled mmWave highway scenario with O-RAN-compliant closed-loop control. Sec. IV-A describes the simulation platform, O-RAN integration, and scenario configuration, and Sec. IV-B presents the baselines, offline RL setup, and evaluation metrics. Sec. IV-C reports the results and analyzes performance trade-offs across vehicle-density regimes.

A. Scenario and Simulation Settings

1) Simulation Platform and O-RAN Integration: To simulate realistic operating conditions and enable closed-loop

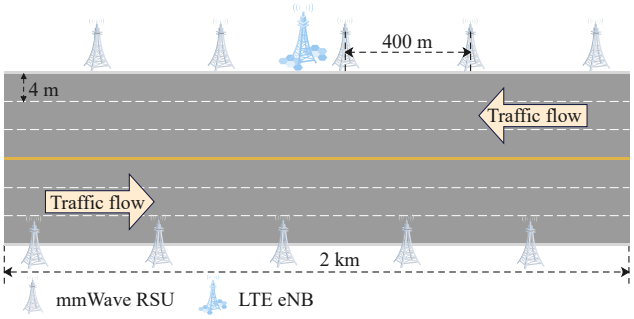

Fig. 3: Simulation scenario, which is a bidirectional six-lane highway compliant to 3GPP TR 38.913.

handover control, we implement beamforming-enabled vehicular handover simulations in ns-3 using the `mmWave` module, and integrate the `oran-interface` module to support O-RAN-compliant E2 message exchange between the simulator and a containerized near-RT RIC from the O-RAN Software Community (OSC) [20]. Vehicular mobility traces are generated in SUMO and imported into ns-3 via `ns3::WaypointMobilityModel`. The near-RT control loop operates with a control interval of $\Delta t = 100$ ms, which is also used as the decision interval for all strategies.

2) Scenario Topology and Mobility Model: We consider a highway scenario aligned with 3GPP specifications [6, 21, 22]. The highway segment is 2 km long and has two directions with three $4-$m-wide lanes per direction. We assume a non-standalone (NSA) deployment with one Long-term Evolution (LTE) eNB and 10 RSUs, as shown in Fig. 3. The eNB is located at the roadside midpoint with height 35 m. The RSUs are placed alternately on both sides at height 10 m with a 400 m spacing on each side.

Vehicular mobility is generated in SUMO using the Intelligent Driver Model (IDM), where each vehicle adapts its acceleration to the preceding vehicle. Vehicles are inserted independently in each direction according to a Poisson arrival process during the first 10 s, with rate r_{arr} (veh/s). To vary traffic density, we set $r_{\text{arr}} \in \{0.1, 0.2, 0.3, 0.4\}$. The maximum vehicle speed is capped at 180 km/h, and lane speed limits are 100 km/h, 150 km/h, and 180 km/h for the inner, middle, and outer lanes, respectively. Since the evaluation is trace-driven, the simulation duration follows the imported trace and varies across random seeds and density settings.

3) Simulation Implementation and Configuration: Following the NSA architecture, the eNB provides wide-area control-plane connectivity, while the RSUs deliver user-plane traffic over mmWave links at 28 GHz [7]. Each RSU employs a 2×2 uniform planar array (UPA) and forms a single directional beam, whereas each vehicle uses a single antenna element [6]. Thus, each RSU serves at most one vehicle per decision slot, inducing one-to-one association coupling across vehicles and RSUs. The RSU and vehicle transmit powers are 30 dBm and 23 dBm, respectively [23]. We configure a 50 MHz carrier bandwidth with numerology index 2 and use the `MmWaveFlexTtiMacScheduler` with HARQ enabled.

Moreover, we adopt the 3GPP highway propagation loss and channel condition models, and update the channel condition every 100 ms in accordance with mmWave-V2X specifications [21, 22]. We configure per-vehicle downlink user-plane traffic, where each vehicle is served by an independent UDP flow generated at the remote host using an `OnOffApplication`. Each flow uses a payload size of 300 bytes and an on-state rate of 5 Mbps. The on-state duration is exponentially distributed with mean 10 ms and the off-state duration is set to zero, yielding back-to-back downlink transmissions [6, 20]. To align traffic demand with vehicle presence, each flow is activated when the vehicle enters the road segment and terminated once it leaves.

TABLE I: Offline RL hyperparameters and their values.

Hyperparameters	**Value**
DQN Agent (Offline)	
Training steps I	20000
Batch size $\lvert\mathcal{B}(i)\rvert$ and $\lvert\widetilde{\mathcal{B}}(i)\rvert$	128
Discount factor γ	0.99
Weight of soft update ξ	0.005
Weight of CQL loss α	1
Initial mixture probability of sampling ρ_0	0.5
Weight of handover sample in loss δ	1
Optimizer	
Optimizer	Adam
Learning rate	0.00005
Neural Network(Fig. 2)	
Feature extractor	$16\times32\times32$ neurons
Q-value estimator	$160\times64\times10$ neurons
Feature aggregator	$320\times256\times128$ neurons
Lagrangian Multipliers	
Initial values λ_0, μ_0	$0, 0$
Stepsizes η_λ, η_μ	$0.001, 0.001$

B. Evaluation Procedure

1) Baseline Strategies: To construct the offline dataset and provide benchmarking references, we run ns-3 simulations with two SINR-driven handover schemes.

Threshold: A handover is triggered when the SINR of a neighboring RSU exceeds that of the serving RSU by 3 dB [23]. The handover is executed immediately once the condition holds at a decision epoch.

DT: A handover is triggered when the neighboring RSU maintains higher SINR than the serving RSU for a time-to-trigger (TTT) that is adjusted based on the SINR difference [24]. This temporal filtering mitigates short-term fluctuations and reduces ping-pong effects.

Both baselines update decisions every $\Delta t = 100$ ms, consistent with the near-RT control-loop interval.

2) Offline RL Configuration: We train the RL agent described in Sec. III. The agent instantiates 10 RSU-specific feature-extractors and Q-value estimators. The latency and reliability QoS constraints are set to 10 ms and 99%, respectively. Training is performed for a fixed number of gradient steps on the logged dataset without online interaction. The complete hyperparameters are summarized in TABLE I.

3) Metrics: We evaluate HONOR and the baselines using metrics that reflect handover efficiency and user-plane service

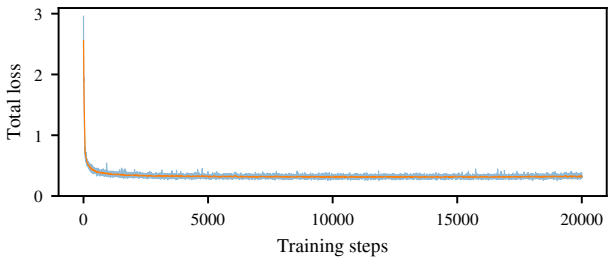

Fig. 4: Offline training loss versus training steps. The curve shows the evolution of the total loss during offline learning.

quality. All metrics are computed per simulation run and then aggregated across 100 seeds for each density setting.

Total handover frequency: It captures the rate of association changes and directly quantifies the effectiveness of the handover optimization objective. For each decision slot, we count the total number of handover events across all active vehicles and normalize it by the number of active vehicles in that slot. We then average this per-slot per-vehicle quantity over the entire run duration, yielding the average handover frequency per vehicle (Hz).

Ping-pong handover frequency: It measures handover stability by counting immediate reversal events. Specifically, a ping-pong handover is recorded for a vehicle when its next handover returns to the RSU it most recently left (A→B→A). We compute the ping-pong handover frequency using the same per-slot normalization and time averaging as the total handover frequency, but counting only ping-pong events.

User-plane latency: It quantifies the user-plane packet delay under mobility and contention. We measure latency as the downlink packet data convergence protocol (PDCP)-to-PDCP delay between the serving RSU and the vehicle for successfully delivered packets, and report its distribution (e.g., median and quartiles) as well as the mean across random seeds.

Reliability: It captures packet delivery performance and is quantified by the downlink PDR. For each run, PDR is computed as the ratio of the total number of successfully received downlink packets to the total number of downlink packets sent, aggregated over all vehicles and the entire simulation duration.

C. Result Analysis

In this subsection, we examine the offline training behavior of the RL agent. Fig. 4 shows the total training loss versus gradient steps. The loss decreases in the initial phase and then stabilizes, indicating stable training on the logged dataset.

We evaluate the deployed HONOR handover xApp in the V2X scenario described in Sec. IV-A. Unless otherwise stated, results are averaged over 100 independent seeds. We report mean values with 95% confidence intervals for total handover frequency, ping-pong handover frequency, and reliability, and report user-plane latency as defined in the metrics.

1) Total Handover Frequency: We evaluate the average handover frequency per vehicle as the vehicle density increases. Fig. 5 shows that HONOR achieves the lowest handover

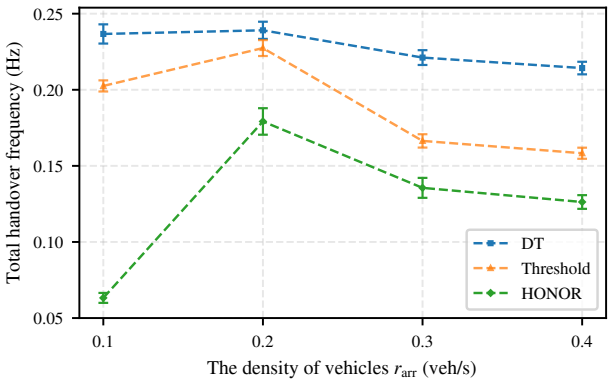

Fig. 5: Mean handover frequency per vehicle (Hz) versus vehicle density for HONOR, Threshold, and DT. Error bars indicate 95% confidence intervals computed over 100 independent seeds.

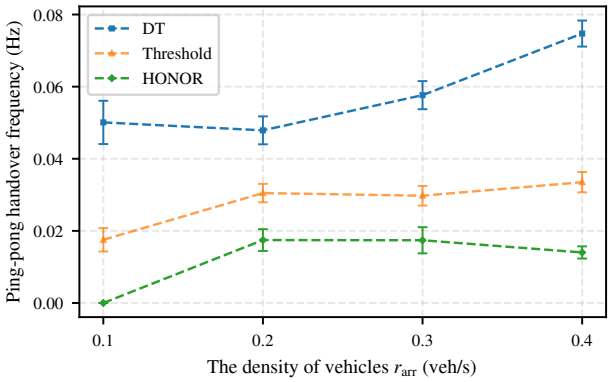

Fig. 6: Comparison of ping-pong handover frequency per vehicle across densities, showing the tendency of rapid back-and-forth reassociations.

frequency at all densities, and its 95% confidence intervals do not overlap with those of the baselines. Compared with Threshold, HONOR reduces the mean handover frequency by 68.8%, 21.2%, 18.5%, and 20.2% at densities 0.1 to 0.4, respectively. Relative to DT, the corresponding reductions are 73.3%, 25.1%, 38.7%, and 41.1%. These trends are consistent with the baseline control logic, where Threshold and DT react to short-term link-quality fluctuations with a 100 ms decision period. In contrast, HONOR accounts for switching costs under coupled scheduling induced by beam exclusivity, and therefore avoids handovers triggered by transient SINR advantages, leading to a lower handover frequency.

2) Ping-pong Handover Frequency: In addition to the total handover frequency, we assess handover stability via the average per-vehicle ping-pong handover frequency. Fig. 6 shows that HONOR consistently yields fewer ping-pong events across all densities. At density 0.1, no ping-pong events are observed under HONOR, while both baselines exhibit non-zero ping-pong frequencies. At higher densities, HONOR further reduces the ping-pong frequency by 42.8%, 41.5%, and 58.2% relative to Threshold, and by 63.5%, 69.8%, and 81.2% relative to DT at densities 0.2 to 0.4, respectively. This behavior is

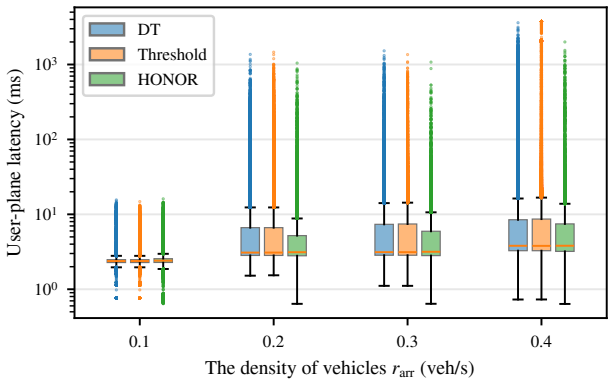

Fig. 7: User-plane latency distribution versus vehicle density for HONOR, Threshold, and DT. Boxplots report median and interquartile range; outliers capture tail-delay events.

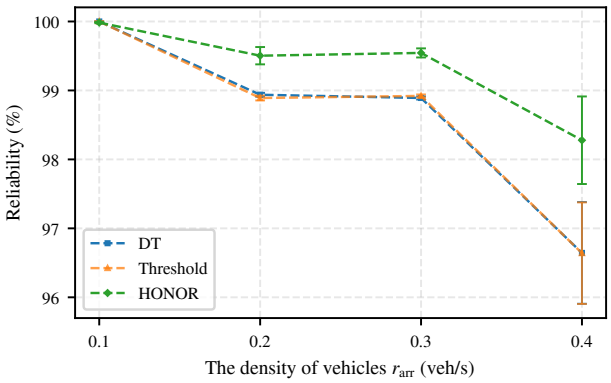

Fig. 8: Packet delivery ratio across densities under the evaluated handover policies, reflecting end-to-end delivery robustness during mobility.

consistent with the fact that contention can create short-lived and alternating SINR advantages across RSUs, which tends to trigger oscillatory decisions under SINR-reactive rules. By accounting for the longer-term cost of oscillations, HONOR favors association persistence unless the expected gain is sustained, thereby reducing ping-pong events. We also observe a mild non-monotonic trend, where ping-pong slightly increases from density 0.2 to 0.3 and then decreases at 0.4, suggesting that the policy remains responsive at moderate contention but becomes more conservative when contention dominates.

3) Latency Distribution: To assess service quality beyond handover-related KPIs, we report the distribution of per-slot user-plane latency, where one latency statistic is computed for each active vehicle in each decision slot. Fig. 7 summarizes latency using boxplots, highlighting the median/interquartile range and tail outliers. Across all densities, the median latency is similar for all methods, indicating that typical delay is largely governed by radio scheduling and is only weakly affected by the handover rule. Differences mainly appear in the upper tail and thus in the mean at medium-to-high densities. For densities 0.2–0.4, HONOR achieves a lower mean latency than Threshold and DT, with the largest gain at density 0.4 where the mean decreases from 8.68 ms (DT) and 11.71 ms (Threshold) to 6.52 ms. This improvement aligns with the handover results, as fewer unnecessary switches reduce transient disruptions and queue build-up that disproportionately affect the latency tail.

4) Reliability: Finally, we evaluate reliability (downlink PDR) as density increases. Fig. 8 shows that HONOR improves reliability over both baselines at medium-to-high densities. The gain is $+0.57$–$+0.65$ percentage points at densities 0.2–0.3 and $+1.64$ percentage points at density 0.4, with the reported 95% confidence intervals. Reliability degrades for all schemes as density increases due to contention and beam exclusivity. Under high load, overly reactive handovers can further increase disruption risk. By reducing handover frequency and maintaining more stable serving configurations, HONOR improves packet delivery performance in dense regimes.

V. Related Work

A. Handover Management in mmWave Networks

In mmWave vehicular networks, handover management has been extensively investigated, including switching between mmWave and LTE [25], beam-centric base station selection [26], and optimal beam selection within a base station [27]. These works generally optimize link-level triggers or introduce architecture-level mechanisms, so they often focus on a single user or consider multiple users taking turns using the same beam, as contention and queueing may violate stringent latency and reliability QoS requirements. We therefore adopt a conservative beam-exclusivity scenario where each RSU serves at most one vehicle per decision slot, so handover decisions must be coordinated across vehicles to avoid infeasible associations.

B. Closed-loop Control in O-RAN

O-RAN has attracted significant interest as an enabler of programmable, closed-loop network control [10]. Prior work develops platforms and workflows for implementing and orchestrating xApps, including experimental closed-loop control and policy-oriented optimization [20]. In vehicular networks, recent studies discuss O-RAN-empowered V2X architectures [28] and demonstrate policy-driven control operations [29], while early efforts explore O-RAN-driven beamforming [30] and mobility management for connected vehicles [23, 31]. These works primarily target high throughput and signal strength, and rarely focus on QoS in terms of latency and reliability.

C. Reinforcement Learning for RAN Control

Method-wise, RL has been widely applied to RAN control for vehicular networks by casting handover [32] or beam tracking [33] as a MDP and then training policies to improve different QoS. These works mainly adopt online RL, which relies on interaction and exploration to collect experiences and improve policies during training. However, such exploration is impractical for safety-critical V2X applications, since trial-and-error actions may cause unpredictable results. In contrast, our approach follows an offline RL paradigm, learning the

handover policy solely from a fixed logged dataset without interacting with the environment during training.

VI. Conclusion

This paper presents an O-RAN-compliant handover management for mmWave multi-vehicle networks. Specifically, we rigorously formulate the handover process as a constrained sequential decision problem with latency and reliability QoS requirements, and solve it through Lagrangian relaxation and an assignment-constrained factorized DQN with validity masking that enforces one-to-one RSU-vehicle matching. The handover policy is trained offline and deployed as HONOR xApp on the near-RT RIC. Extensive O-RAN-integrated ns-3 evaluations across vehicle density regimes demonstrate that HONOR consistently reduces handover frequency by 18%–73% and suppresses ping-pong events by 41%–81% relative to representative baselines, while improving latency performance and slightly increasing packet delivery reliability. Notably, HONOR is derived from baseline data yet outperforms the baselines, enabling iterative improvement with new logs.

Future work includes incorporating fairness across vehicles and extending the model beyond single-beam analog beamforming to support multi-user service with joint handover-resource allocation.

Acknowledgment

Funded in part by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy – EXC 2050/2 – Project ID 390696704 – Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of TUD Dresden University of Technology and the Federal Ministry of Research, Technology, and Space (BMFTR) for its support as part of the research program Communication Systems "Souverän. Digital. Vernetzt.". Joint project 6G-life, project identification number: 16KIS2413K, and supported in part by the European Union's Horizon Europe research and innovation programme, project TOAST under the grant agreement No. 101073465 (HORIZON-MSCA-2022-DN-01). The work of Yizhou Wang is supported by the China Scholarship Council.

References

[1] H. Zhou, W. Xu, J. Chen, and W. Wang, "Evolutionary V2X technologies toward the internet of vehicles: Challenges and opportunities," *Proc. IEEE*, vol. 108, no. 2, pp. 308–323, 2020.

[2] E. Ahmed and H. Gharavi, "Cooperative vehicular networking: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 996–1014, 2018.

[3] 3GPP, "Service requirements for V2X services," 3rd Generation Partnership Project (3GPP), 3GPP TS 22.185, Apr. 2024.

[4] 3GPP, "Service requirements for enhanced V2X scenarios," 3rd Generation Partnership Project (3GPP), 3GPP TS 22.186, Apr. 2024.

[5] S. Gyawali, S. Xu, Y. Qian, and R. Q. Hu, "Challenges and solutions for cellular based V2X communications," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 1, pp. 222–255, 2020.

[6] 3GPP, "Study on scenarios and requirements for next generation access technologies," 3rd Generation Partnership Project (3GPP), 3GPP TR 38.913, Mar. 2024.

[7] M. Mezzavilla *et al.*, "End-to-end simulation of 5G mmWave networks," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2237–2263, 2018.

[8] J. Choi, V. Va, N. Gonzalez-Prelcic, R. Daniels, C. R. Bhat, and R. W. Heath, "Millimeter-wave vehicular communication to support massive automotive sensing," *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 160–167, 2016.

[9] J. Tan *et al.*, "Beam alignment in mmWave V2X communications: A survey," *IEEE Commun. Surveys Tuts.*, 2024.

[10] M. Polese, L. Bonati, S. D'Oro, S. Basagni, and T. Melodia, "Understanding O-RAN: Architecture, interfaces, algorithms, security, and research challenges," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 2, pp. 1376–1411, 2023.

[11] M. Polese *et al.*, "Empowering the 6G cellular architecture with open RAN," *IEEE J. Sel. Areas Commun.*, 2023.

[12] P. Schwenteck, G. T. Nguyen, H. Boche, W. Kellerer, and F. H. P. Fitzek, "6G perspective of mobile network operators, manufacturers, and verticals," *IEEE Netw. Lett.*, vol. 5, no. 3, pp. 169–172, 2023.

[13] O-RAN WG 2, "AI/ML workflow description and requirements," O-RAN ALLIANCE e.V., Technical Report, Jul. 2021.

[14] M. Tayyab, X. Gelabert, and R. Jäntti, "A survey on handover management: From LTE to NR," *IEEE Access*, vol. 7, pp. 118 907–118 930, 2019.

[15] O-RAN WG 3, "O-RAN E2 service model (E2SM) KPM 6.0," O-RAN ALLIANCE e.V., Technical Report, Feb. 2025.

[16] O-RAN WG 3, "O-RAN E2 service model (E2SM) RAN control 7.0," O-RAN ALLIANCE e.V., Technical Report, Feb. 2025.

[17] J. Li, D. Fridovich-Keil, S. Sojoudi, and C. J. Tomlin, "Augmented lagrangian method for instantaneously constrained reinforcement learning problems," in *Proc. 60th IEEE Conf. Decis. Control (CDC)*, 2021, pp. 2982–2989.

[18] H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 2094–2100.

[19] A. Kumar, A. Zhou, G. Tucker, and S. Levine, "Conservative Q-learning for offline reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 1179–1191.

[20] A. Lacava *et al.*, "Programmable and customized intelligence for traffic steering in 5G networks using open RAN architectures," *IEEE Trans. Mobile Comput.*, vol. 23, no. 4, pp. 2882–2897, 2023.

[21] 3GPP, "Study on evaluation methodology of new vehicle-to-everything (V2X) use cases for LTE and NR," 3rd Generation Partnership Project (3GPP), 3GPP TR 37.885, Jun. 2019.

[22] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," 3rd Generation Partnership Project (3GPP), 3GPP TR 38.901, Mar. 2024.

[23] K. Suzuki, J. Nakazato, Y. Sasaki, K. Maruta, M. Tsukada, and H. Esaki, "Toward B5G/6G connected autonomous vehicles: O-RAN-driven millimeter-wave beam management and handover management," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, 2024, pp. 1–6.

[24] M. Polese, M. Giordani, M. Mezzavilla, S. Rangan, and M. Zorzi, "Improved handover through dual connectivity in 5G mmWave mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 2069–2084, 2017.

[25] S. Zang *et al.*, "Mobility handover optimization in millimeter wave heterogeneous networks," in *Proc. 17th Int. Symp. Commun. Inf. Technol. (ISCIT)*, 2017, pp. 1–6.

[26] A. Kose, C. H. Foh, H. Lee, and M. Dianati, "Beam-centric handover decision in dense 5G-mmWave networks," in *Proc. IEEE Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, 2020, pp. 1–6.

[27] S.-M. Oh, S.-Y. Kang, K.-C. Go, J.-H. Kim, and A.-S. Park, "An enhanced handover scheme to provide the robust and efficient inter-beam mobility," *IEEE Commun. Lett.*, vol. 19, no. 5, pp. 739–742, 2015.

[28] F. Linsalata, E. Moro, M. Magarini, U. Spagnolini, and A. Capone, "Open RAN-empowered V2X architecture: Challenges, opportunities, and research directions," in *Proc. IEEE Veh. Netw. Conf. (VNC)*, 2024, pp. 113–116.

[29] P. Sroka, Ł. Kulacz, S. Janji, M. Dryjański, and A. Kliks, "Policy-based traffic steering and load balancing in O-RAN-based vehicle-to-network communications," *IEEE Trans. Veh. Technol.*, 2024.

[30] S. Ozawa, Y. Sasaki, J. Nakazato, M. Tsukada, and K. Maruta, "Toward O-RAN-based cell-free architecture: Cooperative O-RU/V2X mmWave beam tracking," in *Proc. IEEE 99th Veh. Technol. Conf. (VTC2024-Spring)*, 2024, pp. 1–5.

[31] K. Maruta *et al.*, "Millimeter-wave fast beam tracking enabled by RAN/V2X cooperation," in *Proc. Int. Conf. Artif. Intell. Inf. Commun. (ICAIIC)*, 2024, pp. 388–392.

[32] K. Tan, D. Bremner, J. Le Kernec, Y. Sambo, L. Zhang, and M. A. Imran, "Intelligent handover algorithm for vehicle-to-network communications with double-deep Q-learning," *IEEE Trans. Veh. Technol.*, vol. 71, no. 7, pp. 7848–7862, 2022.

[33] J. Ye and H. Gharavi, "Deep reinforcement learning assisted beam tracking and data transmission for 5G V2X networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 9, pp. 9613–9626, 2023.