Received 7 August 2024; revised 12 August 2024; accepted 22 August 2024. Date of publication 30 August 2024; date of current version 12 September 2024.

Digital Object Identifier 10.1109/OJCOMS.2024.3452591

MVX-ViT: Multimodal Collaborative Perception for 6G V2X Network Management Decisions Using Vision Transformer

GHAZI GHARSALLAH [1] (Student Member, IEEE), AND GEORGES KADDOUM [1,2] (Senior Member, IEEE)

[1] Electrical Engineering Department, École de Technologie Supérieure, University of Quebec, Montreal, QC H3C 1K3, Canada

[2] Artificial Intelligence and Cyber Systems Research Center, Department of Computer Science and Mathematics, Lebanese American University, Beirut 797751, Lebanon

CORRESPONDING AUTHOR: G. GHARSALLAH (e-mail: ghazi.gharsallah.1@ens.etsmtl.ca)

ABSTRACT Advancements in sixth-generation (6G) networks, coupled with the evolution of multimodal sensing in vehicle-to-everything (V2X) networks, have opened avenues for transformative research into multimodal-based artificial intelligence (AI) applications for wireless communication and network management. However, this promising research direction is often constrained by the limited availability of suitable datasets. In response, this paper introduces a comprehensive configurable co-simulation framework that integrates the state-of-the-art CARLA and Sionna simulators to generate a multimodal multi-view V2X (MVX) dataset. We present novel AI-based models to predict future line-of-sight (LoS) blockages and optimal beam direction as well as an innovative antenna position optimization (APO) solution, all of which are underpinned by the multimodal dataset MVX. Our framework capitalizes on collaborative perception and significantly enhances V2X communication by integrating LiDAR and wireless data. Thorough evaluations demonstrate that our collaborative perception approach outperforms traditional methods of both beam and blockage prediction in terms of accuracy and efficiency. Additionally, we evaluate the importance of infrastructural elements in V2X systems and conduct a computational study to illustrate that our framework is suitable for various operational scenarios and can be used as a digital twin solution. This work not only contributes to the field of V2X wireless communications by providing a versatile framework for network management but also sets the stage for future research on multi-sensor fusion in AI applications for V2X wireless communication environments to enhance the efficiency and resilience of future 6G networks.

INDEX TERMS 6G, V2X, collaborative perception, vision transformer, network management.

I. INTRODUCTION

A. MOTIVATION

IN THE rapidly advancing field of wireless communications, the transition towards sixth-generation (6G) represents a fundamental transformation characterized by exploring high-frequency bands such as millimeter waves (mmWave) and terahertz (THz). However, these high frequencies are subject to significant penetration loss and attenuation, which make them vulnerable to physical blockages [1] that cause the received signal-to-noise ratio to fluctuate. These fluctuations become particularly problematic when physical obstructions disrupt the line-of-sight (LoS) between base stations and users and cause frequent communication channel interruptions. These interruptions degrade network reliability and lead to substantial delays when re-establishing LoS connections [2]. In addition, massive antenna arrays make it possible to form ultra-narrow beams. This technological advancement serves two purposes: 1) it substantially amplifies the received signal power at the intended users, and 2) it reduces interference, which is crucial for maintaining system integrity. However, this technology complicates beam management, particularly in scenarios involving high-mobility vehicles. The challenges they introduce must be overcome to achieve ultra-reliable low-latency

communication (URLLC) in 6G networks. Meanwhile, the evolution of multimodal sensing plays a crucial role in shaping the future of wireless systems, particularly in 6G networks. 6G vehicular networks have the potential to generate plenty of multimodal data given the various types of sensors that are used by autonomous vehicles [3], [4]. This characteristic has spurred innovative research directions, notably multimodal sensing-aided wireless communication and network management [5], [6], [7], to address the challenges associated with blockage and beam vector prediction using simulated datasets. This research has highlighted how efficiently artificial intelligence (AI) components can utilize the rich dataset to extract valuable insights for predicting blockages [5] and beam vectors [6] to achieve URLLC.

As AI research for autonomous driving progresses, a novel approach has emerged to achieve a more accurate and comprehensive understanding of the environment and AI decision-making capabilities: collaborative perception. Traditional single-agent perception systems often grapple with challenges such as limited sensor range, which can lead to potentially catastrophic outcomes at great distances, as highlighted in [8]. This limitation arises primarily from an individual vehicle's perception being confined to a single perspective with a restricted field of view. Several studies [9], [10], [11], [12] have explored overcoming these limitations by integrating multiple viewpoints of the same scene. This exploration focuses on vehicle-to-everything (V2X) collaborative perception [13], which encompasses both vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) interactions. As these studies demonstrate, the success of collaborative perception in AI-based autonomous driving is largely attributable to the availability of high-quality datasets that are configurable and scalable. However, in wireless communication, particularly in 6G V2X networks, there is currently a lack of comprehensive multimodal datasets that are configurable, multimodal, and multi-user. This absence of suitable datasets presents a significant challenge when it comes to applying proven multimodal collaborative perception methods to enhance the reliability and efficiency of 6G network management.

B. RELATED WORK

Thanks to the use of multimodal data, the evolution of AI-based network management solutions, particularly for beam and blockage prediction in wireless networks, has been significantly accelerated. This section reviews the key developments in this area.

Initially, solutions such as the multi-connectivity approach proposed in [14] laid the groundwork for managing connection links between users and multiple base stations. In that model, a centralized unit aggregates information from all base stations to evaluate the connection quality. However, a critical limitation of these solutions is that they are reactive - they detect disconnections after the fact, which means that users are temporarily disconnected, and communication is delayed until the base station re-establishes the link. It became evident that more predictive approaches were needed to address this challenge. The objective then shifted to enabling base stations to anticipate the state of non-line-of-sight (NLoS) connections and proactively selecting optimal beam vectors to prevent connection loss. The introduction of comprehensive datasets catalyzed the development of several AI-based network management solutions that incorporate multimodal data by using wireless simulators, such as the Blender Sensor simulator [15], the Simulation of Urban Mobility (SUMO) [16] traffic simulator, and Remcom's Wireless InSite simulator [17] for ray tracing, to generate datasets that combine wireless and multimodal data. The following are the prominent related datasets that integrated multimodal data with wireless data.

- LiDAR Data [18]: This dataset combines LiDAR data points and wireless data to aid with LoS detection and reduce the overhead for mmWave beam selection. It was generated using the Blender Sensor simulator [15], the SUMO simulator [16], and the Wireless InSite simulator [17].
- ViWi [19]: This dataset was produced using a virtual simulated multimodal data framework that generates images from fixed base station cameras, LiDAR data, and wireless data using the Wireless InSite simulator [17]. It underscores the importance of visual data in wireless communication datasets, with studies showing that such data enhances blockage prediction accuracy in V2X networks.
- DeepSense 6G [20]: This multimodal dataset is sourced from real-world scenes utilizing various sensors.

The research community has built upon these datasets by exploring the use of various data types to enhance prediction accuracy for static and dynamic blockage detection. Notably, the work presented in [21] leverages the ViWi dataset framework [19] and visual data from the perspective of the base stations to develop a machine learning-based blockage prediction framework. This approach was further expanded by [22], which introduced the use of LiDAR data to predict dynamic link blockages proactively, and [5], utilized base station-perspective images for blockage prediction in mmWave environments. The work presented in [6] further advances this field and uses both visual and positional data to predict the optimal beam indices, which is an innovative alternative to conventional beam sweeping approaches. These studies underscore the significant advantages of employing multimodal data to create an environment-aware AI solution. Notably, the vision-aided blockage prediction solution is highly accurate within a 500-ms prediction window. However, its accuracy degree is reduced when applied over larger prediction windows. Meanwhile, the beam prediction approach effectively reduces beam training overhead, which enhances the overall prediction accuracy and network management efficiency. However, these solutions can be used only with data collected from either a user or a base station perspective, which inherently limits the sensors'

range. This limitation is mainly due to the confinement of perception to a single perspective, which restricts the field of view and may compromise the accuracy of predictions. This constraint can lead to lower communication quality and potentially catastrophic outcomes. Therefore, there is growing recognition that a collaborative perception approach that encompasses a broader environmental understanding is needed. However, the development of such an approach is limited by a lack of suitable datasets. Existing datasets cannot be used to test collaborative perception solutions as they require data inputs from a wider range of perspectives to ensure comprehensive environmental awareness. This gap highlights the urgent need for new and more comprehensive datasets that can support the advancement of collaborative perception techniques for network management. While the focus on multimodal datasets in autonomous driving has led to significant advancements in collaborative perception, the integration of such approaches in wireless communication, particularly AI-based network management solutions, comes with a distinct set of challenges and opportunities.

The exploration of multimodal datasets designed for collaborative perception in autonomous driving has attracted considerable interest. A variety of datasets with distinct features have been developed to advance this field. CARLA [23] is an open-source realistic simulator that was designed to advance autonomous driving research. It incorporates urban layouts, various vehicle models, realistic building structures, dynamic pedestrian activity, and detailed street signs. What sets this simulator apart is its ability to accommodate a wide range of sensor setups comprising cameras and LiDAR sensors. Furthermore, it provides rich data for comprehensive experimentation, including Global Positioning System (GPS) coordinates, vehicle speed, and vehicle acceleration. The CARLA [23] simulator's exceptional quality has laid the foundation for the creation of the following datasets that have contributed significantly to collaborative perception and autonomous driving research.

- OPV2V [24]: This dataset utilizes the CARLA [23] simulator to generate multi-view autonomous driving data (images, LiDAR data points, and GPS coordinates) and focuses primarily on V2V communication.
- V2X-Sim [25]: This dataset comprises RGB images and infrastructure viewpoints. Its utility extends beyond V2V communication.
- DAIR-V2X [26]: This real-world connected autonomous driving dataset features images and point clouds.
- DOLPHINS [27]: This simulated large-scale multimodal multi-view autonomous driving dataset includes vehicles and roadside units (RSUs). It supports both V2V- and V2I-based collaborative perception.

While these datasets have facilitated the development of AI solutions for intelligent transportation, they lack the integration of wireless data, which limits their applicability to wireless communication and network management applications.

C. CONTRIBUTION

In this paper, we introduce a multimodal V2X (MVX) dataset, a groundbreaking dataset that is set to revolutionize 6G V2X AI applications. MVX is the world's first configurable and scalable multi-agent and multimodal dataset created using a co-simulation framework that incorporates differentiable accurate ray tracing simulation. MVX leverages the CARLA simulator [23], which is known for its realistic and high-quality environment modeling, precise three-dimensional (3D) maps, and sensor configurability. We also employ the state-of-the-art Sionna differentiable ray tracing simulator [28] for radio propagation modeling [29].

MVX outperforms traditional autonomous driving datasets by incorporating wireless data, which expands its utility to a wide range of AI-assisted network management and wireless communication applications. It is more useful than existing multimodal wireless datasets in three key ways: 1) it is the first to incorporate wireless data using differentiable ray tracing simulation, 2) it provides a highly configurable simulation that allows all aspects of the physical world and the wireless environment to be manipulated, and 3) it takes into account a wide variety of ground truth information, including object bounding boxes and semantic segmentation. This synthetic dataset has the potential to support large-scale generation, which is challenging and costly to accomplish using real-world data collection methods. Our primary contributions are as follows:

- *Co-Simulation Framework:* We introduce a co-simulation framework that grants complete control over physical world parameters such as sensors, antennas, the number of users, vehicles, pedestrians, and buildings. It also allows the wireless environment configurations, including transmitters, receivers, and the channel model, to be customized.
- *MVX Dataset:* We introduce MVX, a comprehensive multimodal set of data that was collected using a variety of sensors, including cameras and LiDAR sensors, and incorporates both the user and the base station perspective in different environments and scenarios. It includes crucial ground truth annotations like accurate 3D object bounding boxes, semantic segmentation, and differentiable ray tracing-based wireless data.
- *Multimodal Collaborative Perception for Blockage and Beam Prediction:* We propose a novel multimodal transformer-based collaborative perception AI framework for multi-user V2X blockage and beam vector prediction. This framework integrates collaborative perception into future LoS blockage and optimal beam vector prediction, which enables base stations to have a comprehensive understanding of the environment and significantly improves the accuracy and reliability of predictions.
- *Antenna Position Optimization (APO):* We present an architecture to optimize antenna placement in a variety of scenarios. Our architecture utilizes data from

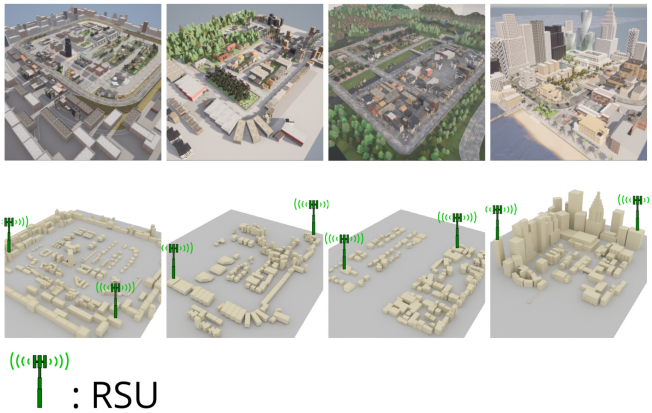


FIGURE 1. Considered CARLA maps and their equivalent in the Sionna simulator.

simulated V2X network traffic and communication scenarios to inform strategic antenna positioning and optimize network performance and reliability in different environmental contexts.

II. SYSTEM AND CHANNEL MODELS

In this section, we present the MVX co-simulation framework, including the settings used for the physical environment as well as the wireless environment's system model and channel model.

To create a realistic simulation environment, we configured the physical settings as follows:

A. SCENARIO AND PHYSICAL ENVIRONMENT SETTINGS

We chose four different representative autonomous driving scenarios and various weather conditions from the CARLA simulator's preset maps, which are presented in Figure 1. The diversity of the maps used is essential to validate the generalizability of AI solutions. Each scenario involves two RSUs and multiple vehicles equipped with LiDAR sensors. The vehicles are initialized in specific locations at each simulation round. Equipping the RSUs with cameras increases the amount of multimodal data from different perspectives that are available to facilitate the development of collaborative perception solutions for wireless communication and network management. Interconnected autonomous vehicles can expand their perceptual fields with the help of other agents and RSUs to improve invisible object detection. The comprehensive data collected provides more information than is typically available from RSU sensors and ensures more accurate and safety-focused decision-making.

B. WIRELESS SYSTEM MODEL

We consider a dynamic V2X network that includes multiple dynamic users (vehicles), dynamic pedestrians, and base stations, and use the Sionna simulator's equivalents of the chosen CARLA maps, which are depicted in Figure 1. Ray tracing is used to simulate physically accurate environment-specific wireless channel realizations for a given scene and user position. More specifically, we use the Sionna ray tracing simulator, which is differentiable as a result of TensorFlow's automatic gradient computation and thus yields channel impulse responses that are differentiable with respect to the ray tracing simulation parameters including material properties (conductivity, permittivity), antenna patterns, orientations, and positions. Transmitters and receivers are equipped with planar antenna arrays having M elements, which are defined by scene properties, to ensure enhanced beamforming capabilities in complex urban settings. The beamforming vector for a given planar antenna array is defined as

$$\mathbf{W} = \frac{1}{\sqrt{M}} \begin{bmatrix} 1 \\ e^{j\frac{2\pi d}{\lambda}(\sin(\theta)\cos(\phi))} \\ \vdots \\ e^{j\frac{2\pi d}{\lambda}((M-1)\sin(\theta)\cos(\phi))} \end{bmatrix}, \tag{1}$$

where d is the antenna element spacing, λ is the wavelength of the carrier frequency, and θ and ϕ are the beam steering angles in azimuth and elevation, respectively. We conduct ray tracing to compute the propagation paths between all transmitters and receivers. Each propagation path i is characterized by a channel coefficient a_i, a delay τ_i, and the angles $(\theta_{\mathrm{R},i}, \varphi_{\mathrm{R},i})$ and $(\theta_{\mathrm{T},i}, \varphi_{\mathrm{T},i})$, which represent the angles of arrival and departure, respectively, in both azimuth and elevation. The channel coefficient a_i is defined as

$$a_i = \frac{\lambda}{4\pi} \mathbf{C}_{\mathrm{R}}\left(\theta_{\mathrm{R},i}, \varphi_{\mathrm{R},i}\right)^{\mathrm{H}} \mathbf{T}_i \mathbf{C}_{\mathrm{T}}\left(\theta_{\mathrm{T},i}, \varphi_{\mathrm{T},i}\right), \tag{2}$$

where $\mathbf{C}_{\mathrm{T}}$ and $\mathbf{C}_{\mathrm{R}}$ are the antenna patterns, and $\mathbf{T}_i$ is a transfer matrix. The Paths object provides detailed information about each simulated path, allowing the generation of channel impulse responses and the application of Doppler shifts to model time-evolving wireless channels.

C. CHANNEL MODEL

In our study, we adopt a geometric channel model as is used in [18], [30], [31] to accurately simulate wireless communications in dynamic V2X networks. This model is particularly suitable for representing complex signal interactions in urban environments and taking into account the various scattering, reflection, and diffraction phenomena that occur in such settings [32]. This model considers multiple propagation paths, each characterized by specific path loss, delay, and arrival and departure angle values. When there are N propagation paths, the channel frequency response at frequency f for user u is mathematically represented as

$$H_u(f) = \sum_{i=1}^{N} a_i e^{-j2\pi f \tau_i}, \tag{3}$$

where τ_i represents the path delay.

D. CONFIGURABILITY

One of the most notable features of our simulation environment is its exceptional configurability. As depicted in

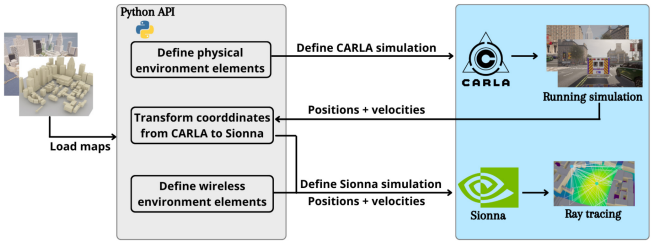


FIGURE 2. The workflow of MVX co-simulation framework.

Figure 2, users can easily modify its configuration using the CARLA Python API. They can add sensors and position them anywhere, remove or alter existing sensors, adjust the number and position of agent vehicles, specify their paths, relocate the RSUs, and even make changes to the buildings with a single line of code.

The wireless environment's configuration is equally flexible and enables users to modify various parameters, including the channel models, the wireless characteristics, material properties such as conductivity and permittivity, and the antenna patterns, orientations and positions of the transmitters and receivers. These adjustments can be made using the Sionna Python API. Therefore, our co-simulation framework is designed to be highly flexible and adaptable to user-specific requirements.

III. PROBLEM FORMULATION AND PROPOSED SOLUTION

In this paper, we address three interrelated challenges associated with dynamic V2X networks: blockage prediction, beam prediction, and antenna position optimization. To tackle these challenges, we introduce a collaborative perception solution using our rich multimodal dataset MVX, which encompasses LiDAR point clouds and beamforming vectors from the perspective of multiple users and infrastructure.

Traditional blockage and beam prediction approaches often rely on the perspective of a single network element, typically a base station equipped with a camera or a LiDAR sensor. We propose a novel approach that considers V2X perception as a heterogeneous multi-agent perception system in which various types of agents, such as base stations and vehicles, perceive their surrounding environment simultaneously and communicate with each other. This collaborative approach aims to provide a more comprehensive understanding of the environment and facilitate more accurate predictions. Our objective is to develop a robust fusion system that significantly enhances the base station's perceptual capabilities. We consider the blockage prediction and beam direction prediction problems independently, although both are inherently spatial. We train our proposed multimodal collaborative perception solution to independently predict future LoS blockages and optimal beam directions using environmental awareness. Our proposed framework's overall architecture is illustrated in Figure 3. The framework comprises four key components: 1) data sequence preparation, 2) feature extraction and sharing, 3) V2X-ViT [13], a dedicated vision transformer designed for V2X collaborative perception, and 4) a prediction head component. The subsections that follow detail each problem's formulation and the constraints and objectives that guide our efforts to enhance communication efficiency and reliability in complex urban environments.

A. COLLABORATIVE PERCEPTION

The focus of our approach to collaborative perception in dynamic V2X networks is synergistic interactions among the vehicles (agents) and base stations. Each vehicle in the network actively participates in data sharing by transmitting to the base station a rich set of features extracted from LiDAR data points. The full collection of data from multiple agents ensures that the base station has a holistic view of the network environment.

The extraction of features from the data collected is a pivotal component of our system. To this end, we adopt the PointPillars method [33], which is an anchor-based technique that is known for efficiently handling point cloud data, lessening computational demand, optimizing memory [24], and transforming LiDAR raw point clouds into a structured format. The structured data, which resembles a 2D pseudo-image, is then processed through the PointPillars backbone to yield informative feature maps. The feature maps are represented as $\{F_i[t]\}_{i=1}^{N_a}$ at a time step t for agent i ($i = -1$ for the base station), where N_a denotes the number of agents, and contains essential spatial information that is then transmitted to the base station.

At the core of our collaborative perception framework is the V2X-ViT solution [13] for V2X collaborative perception, which simultaneously covers V2V and V2I communication. The V2X-ViT solution introduces an innovative heterogeneous multi-agent self-attention module (HMSA) which was designed to adeptly learn and distinguish the varied interactions involved in V2V and V2I communication. It also incorporates a multi-scale window attention (MSwin) module that was specifically developed to effectively capture long-range spatial interactions, which is particularly critical in scenarios requiring high-resolution detection. The V2X-ViT solution employs an adaptive delay-aware positional encoding (DPE) module for the temporal alignment of features, which effectively addresses feature misalignments that may arise due to localization errors and time delays. Additionally, HMSA and MSwin modules facilitate the capture of both inter- and intra-agent interactions. The result is an enriched, aggregated fused feature map $I(t, i)$.

The final stage in our pipeline is the prediction head component, which receives the fused feature maps and applies the time series model gated recurrent unit (GRU) to predict the future LoS state and the optimal beam vector for the future time interval. We chose to use GRU over other time series models for several compelling reasons. GRUs' relatively simple architecture and efficient performance give them a distinct advantage over other recurrent neural network

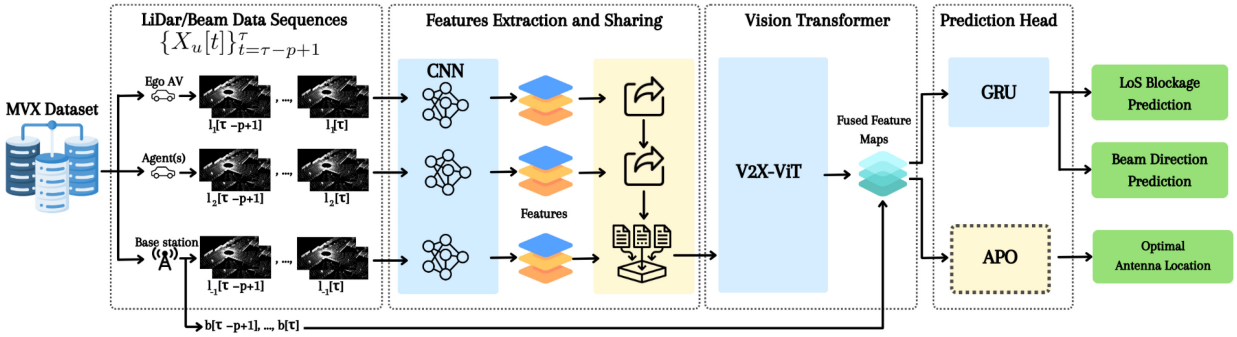


FIGURE 3. Overview of our proposed V2X collaborative perception-based network management system.

variants and enable them to be trained faster and be less computationally complex. Moreover, in recent related works [5], [21], [34], [35], [36], researchers have been able to directly adopt the GRU model for their time series predictions without needing to perform extensive model comparison.

B. BLOCKAGE PREDICTION

In this study, we address the challenge of predicting link blockages in dynamic V2X networks. Our approach involves developing a model that uses beam vectors and LiDAR point clouds collected from the perspective of various users and the base station to predict blockages.

For each link between user u and a base station in the network, we collect a sequence of data over time. At any given time instance t, this data is represented by $X_u[t] = (l_u[t], b_u[t])$, where $l_u[t]$ is the LiDAR data point that captures the environment from the user's perspective and $b_u[t]$ is the beam vector that corresponds to the communication link between the base station and user u. The objective is to use this data to accurately predict the occurrence of a link blockage in the future time steps considered. This problem is formulated as a binary classification problem, where the model predicts whether a blockage will happen in a predefined future time window. For each user u at any time instance t, the LoS state is defined as $s_u[t]$, which is a binary indicator. This indicator reflects whether the communication link between user u and the base station is in a LoS (0) or NLoS (1) state at time t.

The task is to predict the blockage status $y_u[\tau + f]$ at a future time $\tau + f$ given a sequence of data $\{X_u[t]\}_{t=\tau-p+1}^{\tau}$ for a user u over a past time window of size p. In this case, $y_u[\tau + f]$ is a binary variable indicating the occurrence (1) or absence (0) of a blockage in the future time interval f and is defined as

$$y_u[\tau + f] = \begin{cases} 0, & s_u[t] = 0, \forall t \in \{\tau + 1, \ldots, \tau + f\} \\ 1, & \text{otherwise} \end{cases}. \quad (4)$$

Therefore, the challenge lies in predicting $s_u[t + f]$ using the current and past data collected. We specifically seek to develop a predictive solution P such that

$$P\left(\{X_u[t]\}_{t=\tau-p+1}^{\tau}\right) \rightarrow y_u[\tau + f], \quad (5)$$

where P is a function that maps the current sequence of data $\{X_u[t]\}_{t=\tau-p+1}^{\tau}$ to the predicted future LoS state $y_u[\tau + f]$. We aim to forecast the likelihood of an LoS blockage occurring in the next f time steps by analyzing the data $X_u[t]$, which includes the LiDAR data point $l_u[t]$ and the beam vector $b_u[t]$. Our blockage prediction approach requires training a time series GRU model using the MVX dataset whose true future LoS states are known. This training will enable the model to lea'rn the complex patterns and dependencies in the fused feature map $I(t, i)$ provided by the V2X-ViT component that leads to blockages, to facilitate accurate predictions during real-time operation in the dynamic V2X network.

C. BEAM PREDICTION

Our approach to beam prediction involves employing our comprehensive dataset MVX which includes LiDAR data points and previously established beam vectors to anticipate the optimal beam configuration for future communications.

First, a sequence of data is collected for each user $X_u[t]$. The beam vector is an essential aspect of the data and indicates the direction of signal transmission between the user and the base station. The aim is to predict the optimal beam vector that allows the base station to target the user in communication to provide better signal coverage in the future time steps considered. This prediction is crucial for proactive network management because it maintains connectivity and prevents potential signal degradation. This can be formalized by saying future beam state $b_u[t + f]$ is the beam vector that will be most suitable for time $t + f$ considering current and past collected data sequences. The challenge lies in predicting $\theta_u[t+f]$ and $\phi_u[t+f]$, which are the beam steering angles of the beam vector $b_u[t+f]$ in azimuth and elevation, respectively, using the data collected up to the current time t, which is equivalent to predicting the beam vector $b_u[t+f]$. We formulate the beam prediction task as follows:

$$B\left(\{X_u[t]\}_{t=\tau-p+1}^{\tau}\right) \rightarrow \left(\theta_u[\tau + f], \phi_u[\tau + f]\right), \quad (6)$$

where B is a predictive function that maps the sequence of data $\{X_u[t]\}_{t=\tau-p+1}^{\tau}$ to the steering angles $(\theta_u[\tau + f]$ and $\phi_u[\tau + f])$ of the optimal future beam state $b_u[\tau + f]$. We solve this by training a model that is capable of

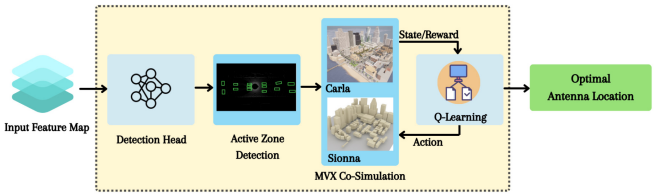


FIGURE 4. APO component architecture.

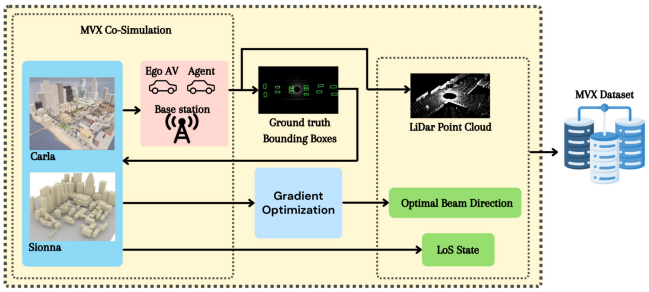


FIGURE 5. MVX data generation framework.

understanding the V2X network's temporal dynamics and spatial characteristics.

The solution leverages our multimodal dataset MVX and employs a GRU to predict the beam state. With this predictive modeling, we aim to obtain a communication system that dynamically adapts to the varying conditions of a V2X network, optimizes beam selection in real time, and enhances the overall network performance.

D. ANTENNA POSITION OPTIMIZATION

Since our co-simulation framework is highly configurable, we propose to optimize the positioning of the antenna in our system to ensure maximum coverage and communication reliability. We propose to use the APO framework illustrated in Figure 4 to do so.

The fused feature maps provided by the V2X-ViT component are fed into the APO component, which utilizes two 1×1 convolution layers to perform the critical tasks of box regression and classification. The regression outputs define the bounding boxes that precisely indicate the agents' position and determine the active area of communication. Once the active zones have been detected, the framework leverages co-simulation with the CARLA and Sionna simulators. This co-simulation makes it possible to realistically model both the urban environment (via CARLA) and the intricate details of wireless signal propagation (via Sionna). We evaluate various scenarios in the simulated environment to investigate the impact antenna positioning has on network performance. The forecasted active area is fed into a Q-learning algorithm that maximizes the coverage map within the target area. This process aims to optimize the antenna's positioning to ensure maximum efficiency and coverage. Formally speaking, the goal is to find the optimal location p^* within an area of potential antenna locations in the simulated environment that maximizes the received signal's coverage in the targeted area (active zone) of the scene. This objective can be quantified by constructing a grid-like coverage map representation that is subdivided into rectangular cells quantifying the signal reception quality at specific locations. The mathematical formulation of the average signal power in each grid cell (i, j) is:

$$\alpha_{i,j} = \frac{1}{|C|} \int_{C_{i,j}} |h(r)|^2 dr, \tag{7}$$

where $|h(r)|^2$ represents the squared magnitude of the path coefficients at position $r = (x, y)$, and the integral is computed over the cell area $C_{i,j}$, with dr being the differential element $dx \cdot dy$. The coverage map contains these values for each cell and provides a detailed signal strength profile for the entire area. To determine the optimal antenna location p^*, we solve the following maximization problem:

$$p^* = \arg\max_{p \in \mathcal{P}} \left(\sum_{i,j} \alpha_{i,j}(p) \right), \tag{8}$$

where $\mathcal{P}$ encompasses all potential antenna locations in the simulated environment, and $\alpha_{i,j}(p)$ denotes the computed average signal power for cell (i, j) given the antenna's position at p. This procedure requires determining the coverage map for different antenna locations and selecting the one that maximizes the cumulative signal coverage, thereby optimizing antenna placement for enhanced network performance.

IV. IMPLEMENTATION

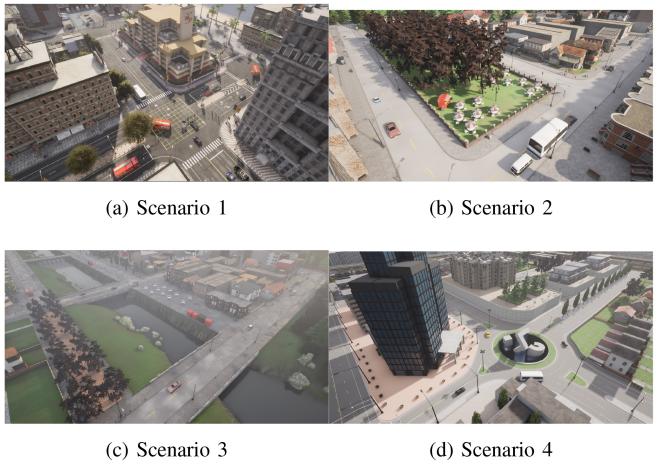
In this section, we present the details of implementing the MVX co-simulation framework for data generation, the collaborative perception framework proposed for beam and blockage prediction, and the APO component.

A. MVX CO-SIMULATION FRAMEWORK AND DATA GENERATION

To the best of our knowledge, there exists no public V2X collaborative perception multimodal datasets for wireless communication applications. We address this situation by providing the public dataset MVX, which was generated using the framework illustrated in Figure 5. It is a novel multimodal dataset suitable for V2X collaborative perception research in wireless communications. It is unique because it includes differentiable ray tracing simulations and high-resolution LiDAR data from the perspective of multiple vehicles and infrastructure to ensure that it is relevant and useful for wireless communication and collaborative perception network management applications.

1) PHYSICAL ENVIRONMENT CONFIGURATION

The MVX dataset was derived from co-simulations conducted in 60 unique scenes across four distinct maps or scenarios using the CARLA simulator. Each scenario lasted precisely 50 seconds and contained two base stations and



(a) Scenario 1 (b) Scenario 2

(c) Scenario 3 (d) Scenario 4

FIGURE 6. The scenarios considered in the co-simulation to generate the MVX dataset.

between two and six intelligent agents, all four of which are depicted in Figure 6. In each scenario, the ego vehicle follows a predefined route and synchronously collects sensor data every 10 seconds, at a rate equivalent to 2 frames per second. Once the simulation of one scenario is complete, the simulator is reset to start a new round. A random number of smart agents also simultaneously collect data, and other traffic participants, such as cars and pedestrians, are first randomly positioned and then controlled using CARLA's AI control mechanisms. LiDAR data is acquired using LiDAR sensors that have a 120-meter data range and have been configured as indicated in Table 1. These sensors are positioned one on each of the vehicles and on the top of each RSU. The sensors use ray-casting to emulate rotating LiDAR. To facilitate sensor calibration [27], the camera and LiDAR sensor are installed at the same location. We also collect ground truth labels provided by the CARLA simulator, which include the 3D bounding boxes of the vehicles participating in the simulation.

2) WIRELESS ENVIRONMENT CONFIGURATION

The previously mentioned physical world scenarios are then replicated in the Sionna simulator, as shown in Figure 1. The ray tracing simulation results are then tuned by adjusting the maximum number of interactions (or bounces) that rays can have with objects in the scene. The shoot-and-bounce ray tracing algorithm's stochastic nature means that multiple runs can yield different path calculations. We fixed the random seed in TensorFlow to ensure reproducibility. We use a site-specific link-level simulation of a 6G V2X network for our wireless environment and consider an OFDM MIMO system in the downlink direction. In line with the 3rd Generation Partnership Project's (3GPP)'s 5G standards for mmWave V2X communication, we configure our ray tracing simulation to have the planar array antenna parameters set out in Table 1 and generate the dataset needed for our framework, including the LoS state $s_u[t]$ and the optimal beam orientation $b_u[t]$ for communication with ego vehicle u.

TABLE 1. Architecture parameters.

Parameter	Value
Data Generation Parameters	
LiDAR Sensors Parameters	
Channels	64
Range	120m
Rotation Frequency	20
Points Generated per Second	10^6
Antennas Parameters	
Number of rows and columns	8×8
Carrier Frequency	140 GHz
Wavelength λ	2.14 mm
Vertical and Horizontal Antenna Spacing	1.07 mm
Antenna Pattern	3GPP TR 38.901 model
Type of Polarization	dual polarization "VH"
Link-Level Simulation Parameters	
Bandwidth	2 GHz
Modulation Order	256-QAM
Code Rate of Channel Coding	5/6
Sub-Carrier Spacing	240 kHz
Number of Sub-Carriers	≈ 8,333
V2X-ViT Architecture Parameters	
Fusion Method	Intermediate Fusion
Core Method	Point Pillar Transformer
Number of Filters	64
Number of Fusion Blocks per Encoder	1
Number of Encoder Layers	3
Optimizer	Adam
Learning Rate	10^{-3}
Learning Rate Scheduler Method	MultiStep
Gamma	0.1

We utilize the Sionna ray tracing module in our simulation to ensure the propagation environment is accurately modeled. This module performs deterministic ray tracing to capture the multipath effects present in the vehicular environment, such as reflections, diffractions, and scattering. We ensure the urban landscape's complex propagation conditions are realistically depicted by adjusting the number of bounces. When it comes to configuring the simulated OFDM MIMO system parameters, such as the number of subcarriers, the subcarrier spacing, and the cyclic prefix length, we tried to configure them in line with the characteristics that 6G V2X networks are expected to have [37], [38], [39], [40], [41]. We set the carrier frequency and the bandwidth for the simulations to 140 GHz and 2 GHz, respectively, which are typical values for the sub-THz frequency range in anticipated 6G networks [42]. This makes it possible to achieve high data transmission rates, which are crucial for V2X communication scenarios that incorporate high mobility and require rapid data exchange. This frequency range was chosen in order to evaluate how our framework performs with higher frequencies that are more susceptible to blockages.

Downlink transmission involves multiple antennas at both the transmitter and the receiver to support beamforming. Once the ego vehicle's and the connected agents' positions and velocities have been obtained from the CARLA simulator, we compute the coverage map for the designated target area. We then calculate the gradients of the average received signal power in this area with respect to the transmitter's orientation. These gradients, which are accessible because the Sionna ray tracing simulations are differentiable, are leveraged to refine the transmitter's orientation using gradient ascent and maximize signal coverage at the ego vehicle's location. The simulation generates time-domain channel coefficients that are converted into frequency-domain channel responses, which are essential for the OFDM modulation process.

Finally, the dataset's LoS states at each simulation time step are used to compute the blockage state $y_u[\tau + f]$ and link it with the corresponding data sequence $\{X_u[t]\}_{t=\tau-p+1}^{\tau}$, which is then partitioned into training, validation, and test sets to support the training and evaluation of our predictive models.

B. FRAMEWORK TRAINING

Once the necessary data is ready, we start the training, validation, and testing of our framework. The implementation is detailed below. All models were trained using Nvidia Quadro P1000 GPUs.

1) COLLABORATIVE PERCEPTION

The V2X-ViT [13] vision transformer component, is an open-source solution designed for collaborative perception in autonomous driving applications. It is known for its ability to detect the bounding boxes of connected vehicles. It comes with a model that has been pre-trained using the parameters presented in Table 1 and a dataset that was synthesized by the CARLA simulator. We adopt the V2X-ViT solution for our framework with some modifications to its architecture to fine-tune it for our MVX dataset using the same parameters.

The raw LiDAR point clouds are first processed into stacked pillar tensors. These tensors are then converted into 2D pseudo-images and input into the PointPillars backbone, which is configured using the parameters outlined in Table 1. This backbone is responsible for extracting informative feature maps, which are then shared with the participant vehicles and the base station in the simulation to lay the groundwork for further feature processing. The aggregated features obtained from the connected agents are input into the V2X-ViT component of our framework. We utilize the open-source V2X-ViT model proposed by Xu et al. [13], which has been designed to perform iterative inter-agent and intra-agent feature fusion using self-attention mechanisms. It is configured with three encoder layers and window sizes of 4, 8, and 16 in the MSwin module. We adopt the Adam optimizer with an initial learning rate of 10^{-3}, which is reduced every 10 epochs by a decay factor of 0.1 as detailed in Table 1. This stage's output features are integrated with the beam vectors for subsequent prediction tasks.

2) PREDICTION HEAD

We train the GRU and APO separately for the prediction head component. First, the time series GRU is trained, using embedded features from the data sequences, to forecast the future blockage state and the optimal beam direction for subsequent time steps. As for the APO, we use the V2X-ViT solution's detection head, which employs the fused feature maps it received and two 1×1 convolution layers to predict the position, size, and yaw angle of the bounding boxes of the vehicles in the network as well as the confidence score of being an object. The area that contains the predicted bounding boxes is then designated as the target active zone. A Q-learning algorithm is used to interactively optimize the antenna's position. This algorithm uses the coverage map within the target active zone as its state space and the communication rate in bits as the reward signal. The algorithm moves the antenna to different positions in the predefined feasible areas to maximize the reward and thus enhance the V2X network's communication rate.

V. NUMERICAL RESULTS AND EVALUATION

In this study, we incorporate LiDAR and wireless data in a collaborative perception-based network management framework and employ a variety of evaluation metrics to rigorously assess each of the framework's components, compare the framework to multiple alternative methods, and evaluate its overall performance efficiency.

A. EVALUATION METRICS AND METHODS USED FOR COMPARISON

In our investigation, multiple baseline methods are employed to evaluate the importance of the various data streams that are captured by the agents. Two distinct solutions form the basis of comparison: one that utilizes exclusively LiDAR data that is from the base station's viewpoint, and the other that considers solely data from the ego user's perspective. To ensure the comparison is fair, all methods evaluated incorporate PointPillars as their backbone architecture and follow the same data processing protocols.

To evaluate the proposed framework's wireless communication performance, we define three benchmark scenarios to be the baseline methods. The first scenario features a base station that utilizes a system that neither anticipates nor responds to LoS blockages. The second features a base station that employs a static beamforming management system that maintains a fixed beam orientation towards the communication zone without dynamic adaptation. The third scenario features a base station that utilizes a heuristic-based beam management system that strategically selects a beam vector from a predefined codebook to optimize data rate coverage within the target area. When it comes to 3D bounding box detection performance, the fine-tuned V2X-ViT solution's ability to correctly identify bounding boxes is

quantified using average precision (AP) at intersection-over-union (IoU) thresholds of 0.5 and 0.7, to provide insight into its level of precision at different levels of recall. As for blockage prediction performance, the solution's binary decision-making performance is assessed using its accuracy metric and the F1 score derived from the confusion matrix. The F1 score is a harmonic mean of the precision P and the recall R, and offers a balance between the two, defined as

$$F1 = 2 \times \frac{P \times R}{P + R}, \tag{9}$$

where P is the ratio of true positives to the sum of true and false positives, and R is the ratio of true positives to the sum of true positives and false negatives. The solution's beam direction prediction performance is reported as a distance-based accuracy (DbA) score. This score quantifies how much the predicted beam direction deviates from the ground truth optimal direction, providing nuanced insights into the predictions' proximity to the true values. The DbA is computed as

$$\text{DbA} = 1 - \frac{d_{gt}}{d_{max}}, \tag{10}$$

where d_{gt} is the distance to ground truth calculated using the Euclidean distance between the predicted and true beam directions, and d_{max} is the maximum possible distance in the defined feature space. The APO and the overall framework performance are evaluated using the received signal power and the data rate within the ground truth active zone. For the received signal power, the ray tracing simulations that were conducted to compute the propagation paths between all the transmitters and receivers provide a channel coefficient a_i for each path i, and the received power is calculated using the channel coefficient as follows:

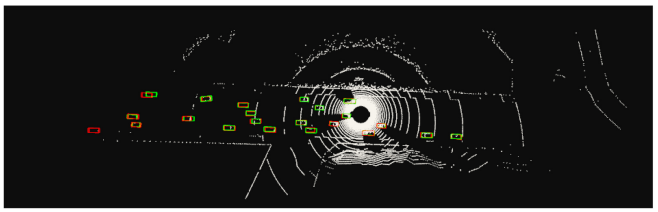
$$P_{\text{rx}}(dB) = 10 \cdot \log_{10}\left(\sum_i |a_i|^2\right). \tag{11}$$

The data rate is assessed using the coverage map, which reflects the signal power's distribution across the active zone.

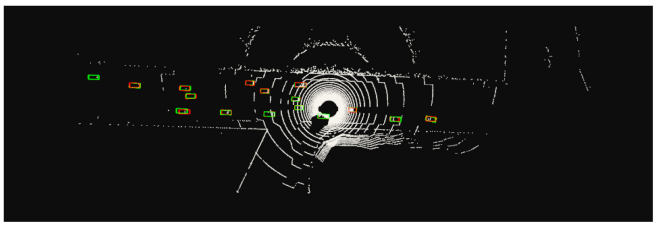
B. BOUNDING BOX DETECTION

How efficiently our framework can detect the bounding boxes of both the ego user and the associated agents in our V2X network is integral to its functionality. It is the foundation upon which various predictive elements of our proposed solution are built.

Despite the pre-trained V2X-ViT solution's proficiency in 3D bounding box detection tasks being proven in autonomous driving scenarios when trained and tested using the V2XSet dataset proposed by [13], we observed performance enhancements upon fine-tuning it on our MVX dataset. This refinement is visually evident in Figure 7, which contrasts the detection capabilities before and after being fine-tuned and tested on our MVX dataset. Optimization yields bounding boxes that are detected with greater precision and closely mirror the ground truth. The fine-tuned model outperforms its non-fine-tuned counterpart, enhancing by 4.1% and 2.9% for IoU thresholds of 0.5 and 0.7, respectively. This improvement is pivotal to the architecture's overall success. Accurate bounding box detection critically informs the predictive modeling of future optimal beam vectors and blockage states as well as the active zone for APO. Hence, the model's 3D detection performance directly influences the robustness and reliability of our entire MVX-ViT framework.



(a) Pre-trained V2X-ViT



(b) Fine-tuned V2X-ViT

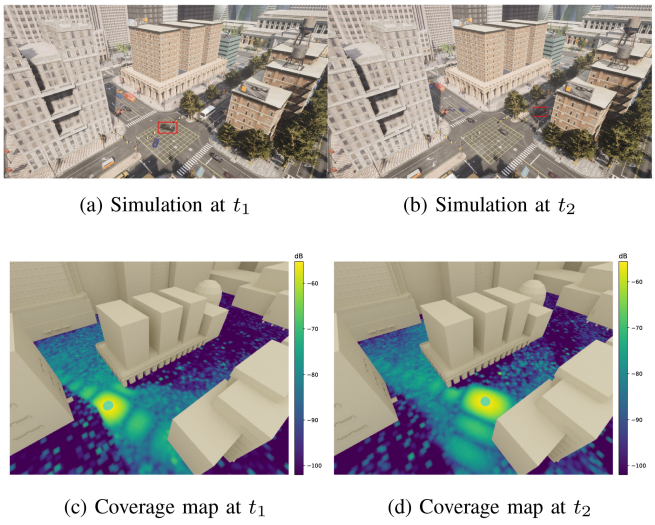
FIGURE 7. Comparison of the bounding boxes detection of (a) the pre-trained model V2X-ViT and (b) the fine-tuned V2X-ViT. Bounding boxes in green depict the ground truth, while those in red indicate the predictions.

TABLE 2. Performance of predictive components for different methods.

Models	Beam Pred.	Blockage Pred.	
	DbA	Accuracy	F1 Score
User Persp.	0.606	0.502	0.413
Base Station Persp.	0.727	0.720	0.611
Collab. Percep. (Ours)	**0.882**	**0.912**	**0.812**

C. MAIN PERFORMANCE COMPARISON

We evaluate the predictive components of our framework by comparing our solutions with different baseline blockage and beam prediction methods. The blockage prediction task utilizes the fused feature map that the V2X-ViT solution learned and the simulated rays to construct data sequences for time series GRU model training. This model utilizes data from the past p time steps to predict the blockage of LoS in the upcoming f time steps. Table 2 compares the performance of the baseline methods considered and our proposed collaborative perception solution. When it comes to blockage prediction, our approach outperforms the user perspective method by a margin of 41% in terms of accuracy and 39.9% in terms of F1 score and the base station perspective method by a margin of 19.2% in terms of accuracy and 20.1% in terms of F1 score over the base station perspective method.



(a) Simulation at t_1 (b) Simulation at t_2

(c) Coverage map at t_1 (d) Coverage map at t_2

FIGURE 8. A comparative example of the dynamic beam direction adaptation for ego vehicle tracking: (a) ego vehicle position at t_1, (b) ego vehicle position at t_2 10 seconds later, (c) coverage map at t_1, and (d) coverage map at t_2.

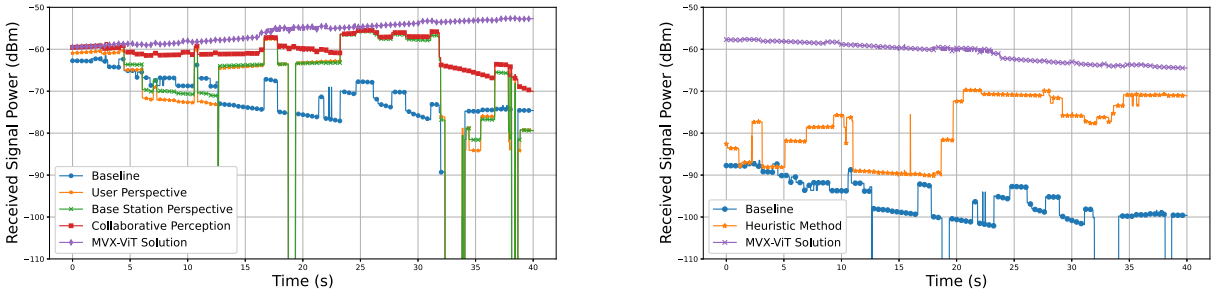
As for the beam direction prediction, Figure 8 illustrates an example using our solution and showcases how the antenna array's beam pattern adapts dynamically to maintain alignment with the movement of the ego vehicle, which is distinctly represented by a red box in the CARLA simulation and a blue point in the Sionna simulation. This adaptation is visualized over 10 seconds, between two specific time steps, t_1 and t_2. The coverage map provided in the figure not only highlights how precisely our solution tracks the vehicle but also illustrates the nuanced changes that occur in the beam's orientation over time, which attests to the system's responsiveness and accuracy when it comes to real-time beam direction adjustment. The results prove that collaborative perception effectively enables the GRU model to accurately predict future blockages. This is attributable to the fact that the aggregated data from the connected agents provides a broader view of the environment and equips the base station with insights outside its immediate field of view. This comprehensive environmental understanding underscores the importance of incorporating collaborative perception in V2X networks.

When it comes to evaluating wireless communication performance, experiments were conducted to observe how the received signal power evolved across various scenarios when our proposed solution and the baseline solutions were used. As for LoS blockage prediction, we simulate a highly dynamic environment with frequent LoS blockages, with N_b operational base stations, one of which is actively communicating with the ego vehicle. All base stations use our future LoS blockage prediction solution to proactively manage handovers, shifting communication from antennas predicted to experience blockages to those predicted to remain blockage-free, thereby maintaining connectivity. When the base station that is actively serving the ego vehicle predicts a potential LoS blockage, it triggers the prediction of the future LoS state of the ego user and all surrounding base stations. If another base station is predicted to maintain an unblocked LoS state, the communication link is handed over to this base station. This proactive approach ensures that the communication link remains unblocked, even in the presence of potential blockages. This dynamic approach's performance is compared to that of three contrasting strategies: 1) a non-adaptive method that does not account for blockages, 2) a user-perspective predictive model, and 3) a base station-perspective predictive model.

We evaluate beam direction prediction using a different scenario in which the ego vehicle is highly dynamic and its movement severely affects the received signal power. In this scenario, our solution enables a base station to adjust its communication beam in real time in response to the ego vehicle's movement. We compare this predictive method to static and heuristic beam management systems. The static system maintains a constant beam orientation, while the heuristic system selects beam vectors from a predefined codebook to enhance the data rate within the target area. In both comparisons, we integrate our MVX-based solution, which employs our models for both LoS blockage and beam direction prediction.

Figure 9 (a) presents how LoS blockages affect the received signal strength when the different methods considered are used. The baseline non-reactive method exhibits severe signal attenuation, with power reductions of more than 99%, due to high frequencies being vulnerable to blockages and potentially experiencing frequent disconnections, which are represented by the vertical lines in the plots. Conversely, the user-perspective and base station-perspective predictive models mitigate some signal degradation with reduced power dips, reaching a nearly 90% and an 88% drop in signal power, respectively, in the detected blockages and the same drop percentages for the undetected blockages. We can interpret the reason for this from the plots: the slight delay in detecting the future blockages that cause the start of the drop in signal before processing the information to prevent the blockages since the amount of information collected at that point is limited and the base stations need more time to collect more data to identify the future blockages. Our collaborative perception-based blockage prediction model excels at detecting and responding to LoS blockages and, in turn, limits power drops to around 70%. This marked improvement underscores the benefit of sharing information between connected agents, which enhances the base stations' situational awareness and accelerates decision-making. Towards the end of the analysis period, all methods experience a decrease in signal strength due to the vehicle's movement away from the beam's focus. Our MVX-ViT solution effectively counteracts this trend by utilizing predicted beam vectors to adjust the beam direction and maintain signal integrity.

Figure 9 (b) indicates how the ego vehicle's movement affects the received signal strength when different prediction methods are used. First, the baseline, which utilizes a static



(a) For different LoS blockage prediction methods

(b) For different beam direction prediction methods

FIGURE 9. Evolution of received signal power over time.

beam vector, experiences a significant decline in signal strength of more than 99%. This sharp decline underscores the challenges static systems face in dynamic vehicular environments. The heuristic alternative, which employs a predefined beam codebook, delivers marginal improvements. In this approach, the coverage area is segmented into distinct sectors, with the beam direction adjusted following the strongest received signal power obtained from user feedback, assuming sector beam vector prediction using the received feedback is perfect. This method does not meet the strict requirements of 6G V2X communications, with an average decline in signal strength of around 60%. The persistence of a fixed beam in each sector leads to continued signal degradation and reflects the method's limited flexibility when there are rapid environmental and behavioral shifts. Conversely, the method that integrates collaborative perception for beam direction prediction substantially mitigates signal fluctuation. Continuously monitoring and predicting the optimal beam direction based on the information shared by the connected agents cancels out the adverse effects of user movement on signal reception. This adaptability is key to maintaining a high-quality communication link and meeting 6G V2X network standards.

D. ABLATION STUDY

In our evaluation of the impact the infrastructure has on our V2X system, we compare two different types of simulations: 1) a V2V configuration in which only vehicles are equipped with sensors and communicate data to the base station, and 2) a V2X setup in which infrastructural elements are also equipped with sensors and contribute to data collection. We also conduct an ablation study to understand how the number of connected agents influences system performance. In this case, agents refer to all the vehicles, except the ego vehicle, that can collect, process, and transmit information to the base station. We base our evaluation on the average data rate, which we calculated by aggregating the data rates observed at each time step over multiple simulation runs and agent counts and then computing the mean value.

As Figure 10 shows, an increase in the number of agents correlates with improved performance for both the V2V and

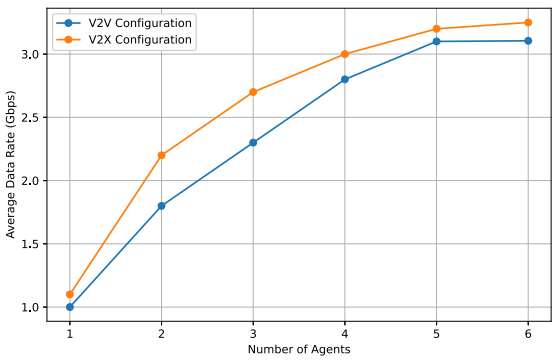
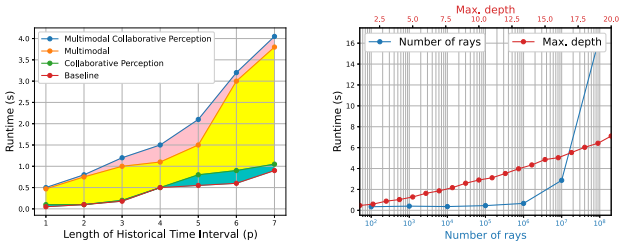


FIGURE 10. Evolution of data rate with the number of agents in 6G V2X environment.

V2X configurations. Notably, the V2X setup outperforms the V2V configuration in terms of average data rate in the considered scenarios. We attribute this enhancement to the infrastructure sensors, which typically experience fewer obstructions and thus provide a broader and less obstructed view. This broader perspective produces a richer dataset, which yields more insightful features that are crucial for accurately interpreting the surrounding environment.

E. COMPUTATION STUDY

We evaluated our framework's performance by demonstrating the level of performance achieved by our proposed approach. However, increasing the volume of data to be processed introduces computational challenges that can lead to communication latency. We address this by conducting a computational evaluation in which we compare our proposed solution with three different approaches. First, we consider a baseline method that involves training a GRU model to predict blockages and beam directions using wireless data from the perspective of the base station. Second, we evaluate a multimodal approach in which a GRU model is trained using multimodal data from the base station's perspective. Third, we assess a collaborative approach that involves a GRU model being trained using collaborative wireless data.



(a) GRU model (b) Ray Tracing simulation

FIGURE 11. Runtime efficiency of two possible approaches.

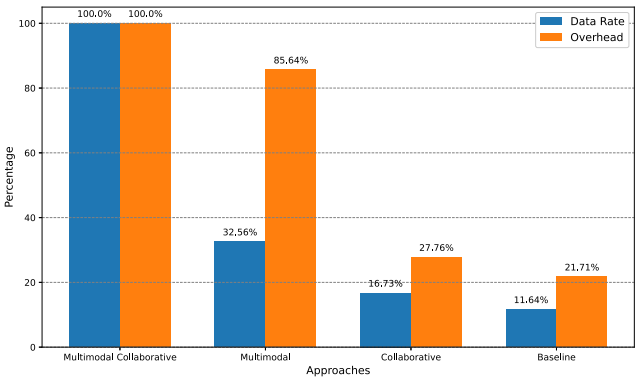


FIGURE 12. Comparison of Data Rate and Overhead by Approach.

Finally, we consider our proposed multimodal collaborative perception approach. Figure 11 (a) presents each GRU model's runtime evolution for different numbers of past time steps p. It highlights the overhead that is introduced by each approach and the impact that incorporating multimodal collaborative perception data has on runtime, which is the amount of time it takes for the GRU model to make a prediction. Despite the fact that GRUs are inherently efficient and simple, the volume of historical data considered significantly influences their performance. The colored area under the curves illustrates the computational overhead that is added by each approach.

To fairly quantify and evaluate the latency introduced by each approach, we further investigate each method's efficiency by comparing the amount of overhead it introduces with its overall performance reported as a percentage. We calculate the area under the runtime curves to determine the amount of overhead that is introduced by each approach and illustrate the percentage of overhead and data rate of each approach compared to the proposed method in Figure 12. The wireless-based collaborative perception approach exhibits a small performance improvement of 5.09% compared to the baseline method due to the irrelevance of using collaborative wireless data to make decisions that are dependent on the ego vehicle's physical surroundings. Conversely, the multimodal aspect introduces the most computational overhead in both the multimodal and multimodal collaborative perception approaches due to

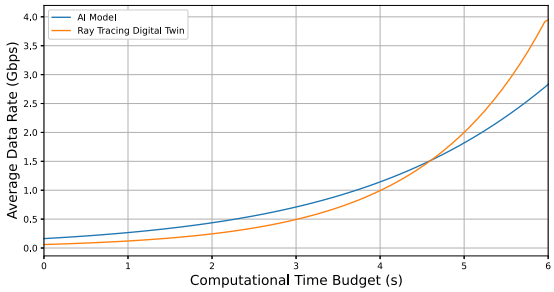


FIGURE 13. Trade-off between system complexity and operational performance.

the size of the LiDAR data and the ViT's processing time. Despite this, the multimodal approach increases the average data rate by only 20.92% compared to the baseline method, while our multimodal collaborative perception approach provides an 88.36% improvement in performance over that of the baseline method with just 14.36% more computational cost than the multimodal solution. This significant improvement is due to the rich information that connected agents contribute to the LiDAR dataset about the physical surroundings of the base station and the ego vehicle.

The availability of computational power allows for an alternative approach: integrating our framework with MVX co-simulation in digital twin-based decision-making. In this scenario, instead of relying on the time series GRU model's predictions, we can deploy an additional convolutional layer to predict the future position of network elements (vehicles) from the fused feature maps. Ray tracing simulation using the future position of the network elements is then employed to determine the future LoS state and the optimal beam direction. The following analysis evaluates both of these methods. In our study, the time series GRU model's runtime efficiency is directly related to the number of past time steps p used to make the prediction, as depicted in Figure 11 (a). This dependency highlights that the model's performance is sensitive to the amount of past data it considers. Conversely, the computational demand for ray tracing simulation, which is illustrated in Figure 11 (b), is related to the maximum depth, which is defined by the number of interactions between a simulated ray and the scene objects to trace the signal paths. This depth factor directly influences the simulation's runtime, with deeper tracing necessitating more computational resources. Balancing system complexity and operational performance is further explored in Figure 13, where we compare an AI model-based approach and a digital twin approach using the communication system's overall quality (average data rate) and different computational time budgets. The average data rate was calculated by combining the data rates recorded at each time step over multiple simulation runs using all 4 maps and then computing the mean value. It provides a comprehensive measure of the system's performance under varying conditions. This analysis reveals that there is a distinct computational time threshold that

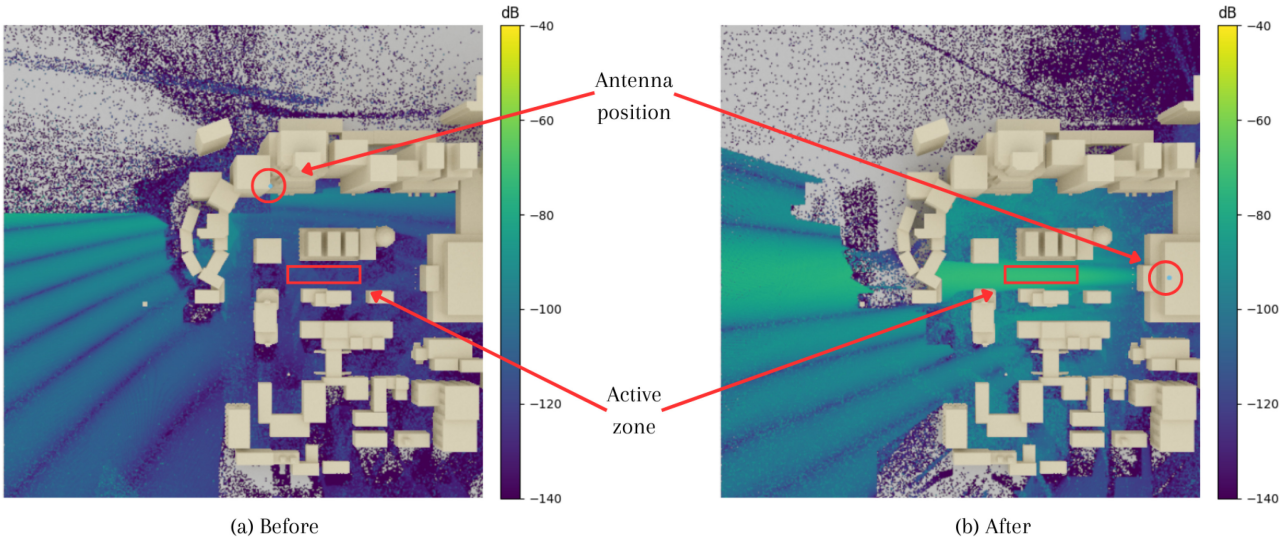


FIGURE 14. Simulation snapshot of the coverage map before and after the application of the APO component.

divides the performance curves into two main parts. Below this threshold, the GRU model performs best and efficiently manages predictions with limited computational resources. However, above this threshold, ray tracing digital twin simulation performs best as the increased computational allowance enables it to deliver more accurate predictions. This highlights a critical trade-off in the computational strategies: the GRU model excels under strict time and computational power constraints and is suitable for scenarios in which rapid decision-making is needed. In contrast, the digital twin approach, with its more intensive computational requirements, is better suited for situations in which the depth and accuracy of the predictions are prioritized over the computational power.

Ultimately, being able to select between these two methods offers a significant advantage. In environments where rapid response and low latency are crucial, such as in highly dynamic V2X networks, the GRU model's speed and efficiency make it the preferred option. Conversely, in contexts where additional computational resources can be afforded and a slight compromise in decision speed is acceptable, particularly in less critical scenarios, the digital twin approach becomes advantageous.

F. APO EVALUATION

The final component of our framework is the APO solution, which exemplifies the practical application of the MVX co-simulation framework and highlights the importance of it being configurable and differentiable. In the following experiment, we compare using our APO component with two different Q-learning-based approaches. The bounding boxes detected for vehicles in multiple simulations are utilized to generate a map that simulates traffic flow. This map is then used to define the targeted active zone, which is denoted by a red rectangle in Figure 14. Afterwards, the APO solution focuses on this active zone, learns to identify a more advantageous position for the antenna, and shifts the antenna to that position. This process maximizes the average received power in the active zone and proposes a new position for the base station, which is indicated by a red circle in Figure 14.

The first Q-learning-based approach considered utilizes a classic Q-learning algorithm in which an agent iteratively explores the environment by choosing actions (antenna movements) either randomly or based on the highest Q-values in a Q-table in accordance with an epsilon-greedy strategy to balance exploration and exploitation. This method enables the agent to learn optimal positions by updating the Q-values based on the rewards received and ultimately converges to the best antenna positions. The second Q-learning-based approach considered enhances the classic Q-learning algorithm by incorporating gradient ascent in the exploration phase. In this case, the agent computes the gradients of the average received power in the active zone with respect to the antenna's position and moves in the direction that maximizes the reward with guidance from a learning rate. Figure 15 demonstrates that the gradient-based method converges to optimal positions more quickly by leveraging the differentiability of the ray tracing simulation and therefore achieves a more efficient and precise learning process than the more stochastic classic Q-learning-based approach does.

VI. CONCLUSION

In this paper, we present a comprehensive co-simulation framework for V2X networks that utilizes two state-of-the-art simulators - CARLA and Sionna. This framework makes it possible to generate datasets with complete control over both the physical and wireless environments' configurations. We introduce novel AI-based predictive models for blockage and beam prediction, along with an innovative antenna position optimization solution. Backed by the multimodal MVX

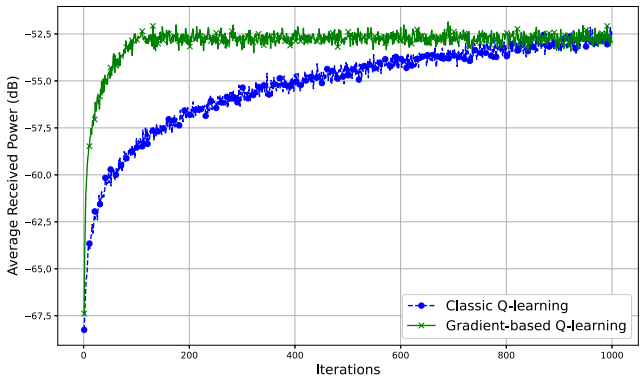


FIGURE 15. Convergence of the APO, comparing the performance of the classic Q-learning and gradient-based Q-learning methods in terms of the average received power in the active zone (reward) over iterations.

dataset, our framework employs collaborative perception to significantly enhance V2X communication. Thorough evaluations demonstrate that our collaborative perception approach outperforms traditional user- and base station-based perspectives. Integrating LiDAR and wireless data that has been processed using our fine-tuned V2X-ViT model significantly enhances bounding box detection and thereby improves the accuracy of beam and blockage prediction. This improvement was not only theoretical but also observable in practice, as our models showed significantly higher data rates and signal strength values in various simulated V2X scenarios. Our ablation evaluation further confirms the importance of incorporating infrastructural elements in V2X systems and reveals the added value that infrastructural data can provide in complex network environments, as well as the number of connected agents that contribute to the collaborative perception of the infrastructure. The results of our computational study underscore our framework's versatility and ability to offer efficient GRU model-based predictions for time-sensitive scenarios and more computationally intensive digital twin simulations for scenarios where precision is paramount. Our framework's APO component brought to light a novel application for our MVX co-simulation environment. It can quickly converge and effectively optimize antenna positions in response to simulated traffic patterns. This advancement opens up new doors for various applications in real-world V2X communication systems to enhance the efficiency of network management.

In conclusion, our work contributes significantly to the field of V2X communication by providing a robust and versatile framework for future research and development. By addressing the challenges associated with dynamic network environments, we leverage the power of AI, collaborative perception, and multimodal data to pave the way for more resilient, efficient, and intelligent V2X network management systems, that are set to meet the demands of the future 6G networks. This work facilitates future research on AI applications for future V2X communication systems.

This work also focuses on LiDAR/wireless-based collaborative network management. Our future work will involve multi-sensor fusion for joint V2X perception and prediction, which is made possible by the configurability of our MVX co-simulation framework. This future direction promises to further enrich the landscape of AI applications in V2X communication systems.

REFERENCES

[1] J. G. Andrews, T. Bai, M. N. Kulkarni, A. Alkhateeb, A. K. Gupta, and R. W. Heath, "Modeling and analyzing millimeter wave cellular systems," *IEEE Trans. Commun.*, vol. 65, no. 1, pp. 403–430, Jan. 2017.

[2] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proc. IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct. 2018.

[3] T. Wild, V. Braun, and H. Viswanathan, "Joint design of communication and sensing for beyond 5G and 6G systems," *IEEE Access*, vol. 9, pp. 30845–30857, 2021.

[4] C. De Lima et al., "Convergent communication, sensing and localization in 6G systems: An overview of technologies, opportunities and challenges," *IEEE Access*, vol. 9, pp. 26902–26925, 2021.

[5] G. Charan and A. Alkhateeb, "Computer vision aided blockage prediction in real-world millimeter wave deployments," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, 2022, pp. 1711–1716.

[6] G. Charan, M. Alrabeiah, T. Osman, and A. Alkhateeb, "Camera based mmWave beam prediction: Towards multi-candidate real-world scenarios," 2023, *arXiv:2308.06868*.

[7] Y. Tian, Q. Zhao, Z. el Abidine Kherroubi, F. Boukhalfa, K. Wu, and F. Bader, "Multimodal transformers for wireless communications: A case study in beam prediction," 2023, *arXiv:2309.11811*.

[8] Z. Zhang and J. Fisac, "Safe occlusion-aware autonomous driving via game-theoretic active perception," in *Proc. 17th Robot., Sci. Syst.*, 2021, pp. 1–13. [Online]. Available: http://dx.doi.org/10.15607/RSS.2021.XVII.066

[9] Z. Lei, S. Ren, Y. Hu, W. Zhang, and S. Chen, "Latency-aware collaborative perception," 2022, *arXiv:2207.08560*.

[10] T.-H. Wang et al., "V2VNet: Vehicle-to-vehicle communication for joint perception and prediction," 2020, *arXiv:2008.07519*.

[11] Q. Chen, "F-Cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3D point clouds," 2019, *arXiv:1909.06459*.

[12] Q. Chen, S. Tang, Q. Yang, and S. Fu, "Cooper: Cooperative perception for connected autonomous vehicles based on 3D point clouds," 2019, *arXiv:1905.05265*.

[13] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, and J. Ma, "V2X-ViT: Vehicle-to-everything cooperative perception with vision transformer," 2022, *arXiv:2203.10638*.

[14] M. Giordani, M. Mezzavilla, S. Rangan, and M. Zorzi, "Multi-connectivity in 5G mmWave cellular networks," in *Proc. Mediterr. Ad Hoc Netw. Workshop (Med-Hoc-Net)*, 2016, pp. 1–7.

[15] M. Gschwandtner, R. Kwitt, A. Uhl, and W. Pree, "BlenSor: Blender sensor simulation toolbox," in *Proc. Int. Symp. Vis. Comput.*, 2011, pp. 199–208.

[16] D. Krajzewicz, J. Erdmann, M. Behrisch, and L. Bieker, "Recent development and applications of SUMO—Simulation of urban MObility," *Int. J. Adv. Syst. Meas.*, vol. 5, nos. 3–4, pp. 1–11, 2012.

[17] "Wireless InSite." Remcom. Accessed: 1997. [Online]. Available: https://www.remcom.com/wireless-insite-em-propagation-software

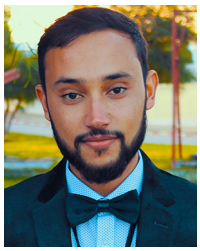
[18] A. Klautau, N. González-Prelcic, and R. W. Heath, "LIDAR data for deep learning-based mmWave beam-selection," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 909–912, Jun. 2019.

[19] M. Alrabeiah, A. Hredzak, Z. Liu, and A. Alkhateeb, "ViWi: A deep learning dataset framework for vision-aided wireless communications," in *Proc. IEEE 91st Veh. Technol. Conf.*, 2020, pp. 1–5.

[20] A. Alkhateeb et al., "DeepSense 6G: A large-scale real-world multi-modal sensing and communication dataset," *IEEE Commun. Mag.*, vol. 61, no. 9, pp. 122–128, Sep. 2023.

[21] G. Charan, M. Alrabeiah, and A. Alkhateeb, "Vision-aided 6G wireless communications: Blockage prediction and proactive handoff," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 10193–10208, Oct. 2021.

[22] S. Wu, C. Chakrabarti, and A. Alkhateeb, "Proactively predicting dynamic 6G link blockages using LiDAR and in-band signatures," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 392–412, 2023.
[23] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. 1st Annu. Conf. Robot Learn.*, 2017, pp. 1–16.
[24] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. Ma, "OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, 2022, pp. 2583–2589.
[25] Y. Li et al., "V2X-Sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 10914–10921, Oct. 2022.
[26] H. Yu et al., "DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 21329–21338.
[27] R. Mao, J. Guo, Y. Jia, Y. Sun, S. Zhou, and Z. Niu, "DOLPHINS: Dataset for collaborative perception enabled harmonious and interconnected self-driving," in *Proc. Asian Conf. Comput. Vis.*, 2023, pp. 495–511. [Online]. Available: https://doi.org/10.1007
[28] J. Hoydis et al., "Sionna: An open-source library for next-generation physical layer research," Mar. 2022, *arXiv:2203.11854*.
[29] J. Hoydis et al., "Sionna RT: Differentiable ray tracing for radio propagation modeling," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, 2023, pp. 317–321.
[30] A. Ali, N. González-Prelcic, and R. W. Heath, "Millimeter wave beam-selection using out-of-band spatial information," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1038–1052, Feb. 2018.
[31] M. Alrabeiah and A. Alkhateeb, "Deep learning for mmWave beam and blockage prediction using sub-6 GHz channels," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5504–5518, Sep. 2020.
[32] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016.
[33] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)* 2019, pp. 12689–12697.
[34] A. Alkhateeb, I. Beltagy, and S. Alex, "Machine learning for reliable mmWave systems: Blockage prediction and proactive handoff," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, 2018, pp. 1055–1059.
[35] R. S. Kalyanshetti, V. Kulkarni, M. Jagannath, A. P. Yashaswi, S. Srikanth, and S. Gurugopinath, "Deep learning-based blockage prediction for intelligent reflecting surfaces-aided V2X networks," in *Proc. Int. Conf. Wireless Commun. Signal Process. Netw. (WiSPNET)*, 2024, pp. 1–5.
[36] F. S. Woon and C. Y. Leow, "Intelligent reflecting surfaces aided millimetre wave blockage prediction for vehicular communication," in *Proc. IEEE 6th Int. Symp. Telecommun. Technol. (ISTT)*, 2022, pp. 11–15.
[37] P. S. R. Henrique and R. Prasad, *6G The Road to the Future Wireless Technologies 2030*. Aalborg, Denmark: River Publ., 2021.
[38] H. Tataria, M. Shafi, A. F. Molisch, M. Dohler, H. Sjöland, and F. Tufvesson, "6G wireless systems: Vision, requirements, challenges, insights, and opportunities," *Proc. IEEE*, vol. 109, no. 7, pp. 1166–1199, Jul. 2021.
[39] M. Z. Chowdhury, M. Shahjalal, S. Ahmed, and Y. M. Jang, "6G wireless communication systems: Applications, requirements, technologies, challenges, and research directions," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 957–975, 2020.
[40] M. Noor-A-Rahim et al., "6G for vehicle-to-everything (V2X) communications: Enabling technologies, challenges, and opportunities," *Proc. IEEE*, vol. 110, no. 6, pp. 712–734, Jun. 2022.
[41] F. Tang, Y. Kawamoto, N. Kato, and J. Liu, "Future intelligent and secure vehicular network toward 6G: Machine-learning approaches," *Proc. IEEE*, vol. 108, no. 2, pp. 292–307, Feb. 2020.
[42] V. Petrov, T. Kurner, and I. Hosako, "IEEE 802.15. 3D: First standardization efforts for sub-terahertz band communications toward 6G," *IEEE Commun. Mag.*, vol. 58, no. 11, pp. 28–33, Nov. 2020.



GHAZI GHARSALLAH (Student Member, IEEE) received the B.S. degree in engineering from École Polytechnique de Tunisie, Tunisia, in 2017. He is currently pursuing the Ph.D. degree in electrical engineering at École de Technologie Supérieure, Université du Québec, Montreal, Canada. His research interests include AI-based solutions for network management in 6G V2X networks.

GEORGES KADDOUM (Senior Member, IEEE) received the bachelor's degree in electrical engineering from École Nationale Supérieure de Techniques Avancées, Brest, France, the M.S. degree in telecommunications and signal processing (circuits, systems, and signal processing) from Université de Bretagne Occidentale and Telecom Bretagne, Brest, in 2005, and the Ph.D. degree (Hons.) in signal processing and telecommunications from the National Institute of Applied Sciences, University of Toulouse, Toulouse, France, in 2009. He is currently a Professor and the Research Director of the Resilient Machine Learning Institute, and the Tier 2 Canada Research Chair of École de Technologie Supérieure (ÉTS), Université du Québec, Montreal, Canada. He has published more than 300 journal articles, conference papers, and two chapters in books, and has eight pending patents. His research interests include wireless communication networks, tactical communications, resource allocations, and network security. He received the Best Papers Award from the 2014 IEEE International Conference on Wireless and Mobile Computing, Networking, Communications, the 2017 IEEE International Symposium on Personal Indoor and Mobile Radio Communications, and the 2023 IEEE International Wireless Communications and Mobile Computing Conference. He received the IEEE Transactions on Communications Exemplary Reviewer Award in 2015, 2017, and 2019. He received the Research Excellence Award from Université du Québec in 2018. In 2019, he received the Research Excellence Award from ÉTS in recognition of his outstanding research outcomes. He also won the 2022 IEEE Technical Committee on Scalable Computing Award for Excellence (Middle Career Researcher). He has received the prestigious 2023 MITACS Award for Exceptional Leadership. He served as an Associate Editor for IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY and IEEE COMMUNICATIONS LETTERS. He is also serving as an Area Editor for IEEE TRANSACTIONS ON MACHINE LEARNING IN COMMUNICATIONS AND NETWORKING and an Editor for IEEE TRANSACTIONS ON COMMUNICATIONS.