IEEE ComSoc | IEEE COMPUTER SOCIETY | IEEE Signal Processing Society | VTS | IEEE Transactions on Machine Learning in Communications and Networking

Received 7 December 2024; revised 24 March 2025 and 1 June 2025; accepted 4 June 2025.
Date of publication 11 June 2025; date of current version 19 June 2025.

The associate editor coordinating the review of this article and approving it for publication was M. Mozaffari.

Digital Object Identifier 10.1109/TMLCN.2025.3578577

Multimodal Collaborative Perception for Dynamic Channel Prediction in 6G V2X Networks

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, Quebec, QC G1K 9H6, 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 (ghazi.gharsallah.1@ens.etsmtl.ca)

This work was supported by Canada Research Chair in Unlocking the Power of the Internet of Things (IoT) 6G-Networks.

ABSTRACT The evolution toward sixth-generation (6G) wireless communications introduces unprecedented demands for ultra-reliable low-latency communication (URLLC) in vehicle-to-everything (V2X) networks, where fast-moving vehicles and the use of high-frequency bands make it challenging to acquire the channel state information to maintain high-quality connectivity. Traditional methods for estimating channel coefficients rely on pilot symbols transmitted during each coherence interval; however, the combination of high mobility and high frequencies significantly reduces the coherence times, necessitating substantial bandwidth for pilot transmission. Consequently, these conventional approaches are becoming inadequate, potentially causing inefficient channel estimation and degraded throughput in such dynamic environments. This paper presents a novel multimodal collaborative perception framework for dynamic channel prediction in 6G V2X networks, integrating LiDAR data to enhance the accuracy and robustness of channel predictions. Our approach synergizes information from connected agents and infrastructure, enabling a more comprehensive understanding of the dynamic vehicular environment. A key innovation in our framework is the prediction horizon optimization (PHO) component, which dynamically adjusts the prediction interval based on real-time evaluations of channel conditions, ensuring that predictions remain relevant and accurate. Extensive simulations using the MVX (Multimodal V2X) high-fidelity co-simulation framework demonstrate the effectiveness of our solution. Compared to baseline methods—namely, a classical LS-LMMSE approach and a wireless-based model that solely relies on channel measurements—our framework achieves up to a $\mathbf{30.82\%}$ reduction in mean squared error (MSE) and a $\mathbf{32.76\%}$ increase in goodput. These gains underscore the efficiency of the PHO component in reducing prediction errors, maintaining low bit error rates, and meeting the stringent requirements of 6G V2X communications. Consequently, our framework establishes a new benchmark for AI-driven channel prediction in next-generation wireless networks, particularly in challenging urban and rural scenarios.

INDEX TERMS 6G, V2X, collaborative perception, channel prediction.

I. INTRODUCTION

A. MOTIVATION

THE evolution toward sixth-generation (6G) wireless communications represents a transformative leap in the telecommunications domain, promising to deliver unprecedented data rates, ultra-low latency, and pervasive connectivity [1], [2], [3], [4], [5], [6], [7]. Building upon the foundation of fifth-generation (5G) networks—which introduced enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable low-latency communications (URLLC)—6G aims to not only further enhance these capabilities by integrating artificial intelligence (AI) into network architectures and using new spectral resources such as high frequencies [1], [2], [3],

but also to bring additional functionalities such as global connectivity, integrated sensing and communication, and AI-driven network optimization. These advancements are essential to enable truly ubiquitous and adaptive network services that would go beyond the traditional 5G paradigm, and are crucial to support next-generation applications like autonomous vehicular networks, holographic communications, and immersive extended reality (XR), all of which require real-time data processing and adaptive communication strategies.

Vehicle-to-everything (V2X) communications represent a critical application for 6G networks, expected to fundamentally transform the way vehicles interact with each other and their environment. By leveraging the advanced capabilities of 6G, V2X communications aim to meet the stringent Quality of Service (QoS) requirements of modern transportation systems. This includes enabling features such as real-time traffic management, high-precision navigation, and instantaneous safety alerts [6], [7]. To support these demanding applications, 6G is anticipated to utilize an extensive range of frequency bands—including sub-6 GHz, millimeter-wave (mmWave), terahertz (THz), and optical wireless bands—to provide the necessary bandwidth and reliability.

Despite the promising potential of 6G V2X communications, their deployment introduces complex challenges, especially in highly dynamic vehicular environments. These settings—characterized by fast-moving vehicles and constantly changing urban landscapes—demand a communication network capable of maintaining high-speed connectivity and high reliability [4], [8]. Furthermore, the use of high-frequency bands in such environments decreases the coherence time duration, necessitating a greater number of pilot signals to maintain acceptable channel estimation accuracy. This increase in pilot signals significantly raises overhead and communication latency, which is detrimental to time-sensitive vehicular applications and may lead to dangerous consequences. Consequently, there is a critical need for advanced techniques that provide faster and more accurate estimates of the channel state to address these challenges effectively.

Accurate channel prediction plays a fundamental role in V2X communications. It enables the system to actively adjust communication parameters to suit rapidly changing conditions. In 6G V2X communication systems, the ability to dynamically modify the transmission power, modulation scheme, and coding rates by forecasting channel conditions with high precision is crucial. This optimization of the transmission parameters is what enables the maintenance of high QoS in these dynamic environments. When a channel prediction model anticipates a deterioration in channel quality, it can proactively instruct the communication system to switch to a more robust modulation scheme or to increase power levels. Similarly, when an improvement in channel conditions is predicted, the system can reduce the transmission power or switch to a higher-order modulation scheme to increase the data rate without risking packet loss. Furthermore, accurate channel predictions enable the system to apply adaptive coding strategies. The system can effectively balance throughput and robustness by adjusting the error correction levels according to the predicted channel conditions. In robust channel conditions, less redundancy is required, allowing for higher throughput. Conversely, in poor channel conditions, increased redundancy ensures that even if data packets are corrupted during transmission, the information can still be successfully recovered. This adaptability reduces the likelihood of data retransmissions, which are costly in latency and bandwidth, making it particularly beneficial for safety-critical communications in V2X contexts. Effective error handling supported by accurate channel predictions is crucial for maintaining uninterrupted communication flows. It is essential for the real-time functionalities of autonomous and connected vehicles and vital for operational safety in intelligent transportation systems.

One of the primary advantages of AI-based channel prediction models is their ability to learn the channel's time and frequency correlations. This was proven by several AI-based channel prediction methods that have been proposed to reduce the number of required pilots by leveraging historical channel data to predict future channel states [9], [10], [11], [12]. These methods have shown promise in decreasing the latency associated with traditional channel estimation by exploiting the inherent time and frequency correlation in channel evolution. However, in 6G V2X networks, these correlations are expected to be lower due to vehicles' high mobility and the use of higher frequency bands. This dynamic nature of 6G V2X environments demands even more sophisticated AI solutions to ensure reliable and efficient channel predictions. Given these challenges, there is an increasing need to use additional information from other data sources, other than historical channel data, to enhance prediction accuracy. By integrating data from various sensors available in connected vehicles, such as LiDAR, cameras, and radars, we can achieve a more comprehensive understanding of the vehicular environment. This multimodal collaborative perception approach allows the system to capture diverse and complementary information, paving the way for more robust and precise channel predictions in highly dynamic 6G V2X networks.

In addition, connected vehicles in 6G V2X networks are expected to be equipped with a diverse array of sensors [13], including cameras, LiDARs, radars, and ultrasonic sensors, each providing different modalities of environmental data. The availability of multiple connected agents, each with its unique perspective and sensory inputs, enables the development of collaborative multimodal solutions. By synergizing data from these varied sources, it is possible to construct a more comprehensive and accurate representation of the vehicular environment [14]. Such collaborative approaches can significantly enhance the precision of channel predictions by incorporating real-time, multi-perspective insights into the dynamic conditions affecting signal propagation. This holistic understanding is crucial for optimizing communication

strategies and ensuring robust, reliable connections across the rapidly changing landscapes of modern transportation networks.

The complexity and high dimensionality of multimodal data streams in 6G V2X environments pose a significant challenge in extracting pertinent features from such diverse inputs. Vision transformers have proven their performance in collaborative perception tasks for autonomous driving [14]. These models are particularly adept at handling data sequences with varying lengths, allowing them to capture nuanced interdependencies between different types of sensor inputs. By aligning and focusing on the most relevant features across these modalities, transformers can extract a rich set of features that comprehensively represent the dynamic vehicular environment. Once these features are extracted, Deep Learning (DL) time series models can navigate the extracted information in dynamic vehicular environments. DL models, known for their ability to process and analyze large volumes of data, are effective in environments where patterns and relationships evolve over time. Particularly, gated recurrent unit (GRU) models excel in accurate and fast predictions of wireless channels over other recurrent neural network (RNN) variants [15]. GRUs' relatively easy architecture and efficient performance give them a distinct advantage over other RNN variants, enabling them to be trained faster while being less computationally complex. These models leverage layered architectures to learn temporal patterns and dependencies critical for predicting changes in the channel state influenced by varying vehicular speeds, obstacles, and other environmental factors.

B. RELATED WORK

The research community has made significant advancements in leveraging AI for channel prediction, demonstrating the potential of various machine learning models to enhance the reliability and accuracy of 6G V2X communications. Early contributions by [16] provided a channel parameter prediction solution using feed-forward neural networks (FNN) and radial basis function neural networks (RBF-NN), and showed promising results in modeling mmWave massive MIMO channels in indoor 5G environments. Recent advancements have introduced generative adversarial networks (GANs) as a powerful tool for channel modeling. A novel GAN framework proposed by [17] offered a practical approach to model wireless channels autonomously.

Subsequent studies expanded on this concept, with ChannelGAN [18], [19] and other GAN-based architectures [20] synthesizing channel data that accurately mirrored real-world conditions. Long short-term memory (LSTM)-based models have attracted attention for their ability to manage sequence-related data effectively. For instance, LSTM was used in [21] to predict the received power and other channel parameters for beyond 5G networks. Integrating LSTM with GANs in [22] was shown to enhance channel prediction accuracy, enriching channel data and effectively predicting future channel states in indoor scenarios. Authors in [11] proposed a mobility-induced channel prediction (MICP) method that uses mobility parameters to predict channel frequency responses, demonstrating the importance of mobility information in channel prediction. Similarly, graph attention networks (GAT) and GRU networks were employed by [23] to model the scatterer density in MmWave MIMO channels, showing substantial improvements in predictive accuracy. Further enriching the landscape of channel prediction, recent studies have proposed integrated solutions that combine multiple AI techniques. A notable example in [9] includes integrating a Kalman filter and LSTM for channel tracking and prediction in intelligent reflecting surface-aided wireless communication systems, reducing channel training overhead. Similarly, the space-time joint predictive channel model proposed by [24] using a space-time generative adversarial network (STGAN) and GRU framework demonstrated its capability to reconstruct lost data and accurately predict unknown channel conditions in both Line-of-Sight (LoS) and Non-Line-of-Sight (NLoS) indoor scenarios, with stationary transmitters and receivers. These studies collectively emphasize the effectiveness of AI models in accurately modeling and predicting future channel states by learning channel time, frequency, and space correlations. Among the diverse AI models used for channel prediction, GRUs have been recognized as particularly effective when handling time series data, as highlighted in a comprehensive survey comparing various DL time series models used for channel prediction [15].

Integrating multimodal data sources has also proven beneficial in channel prediction tasks. A deep multimodal learning architecture was proposed by [25] and [26], designed for massive MIMO systems using various data modalities to enhance channel prediction accuracy, where the authors introduced the use of the user's location as an additional modality for channel prediction. However, the integration of other types of data that could significantly enrich environmental understanding, such as LiDAR data, has been limited due to the lack of available datasets.

Beyond channel prediction, the integration of multimodal data has shown proficiency in predicting line-of-sight blockage and beam direction.

For instance, a deep learning-based approach was proposed in [27] for blockage prediction in 6G V2X using images from the perspective of the base station, and [28] showed the efficacy of proactively predicting dynamic 6G link blockages using LiDAR and in-band signatures, illustrating how high-resolution environmental sensing can detect and mitigate obstructions in real time. Similarly, in [29], the authors presented a LiDAR-aided beam prediction approach in real-world mmWave Vehicle-to-infrastructure (V2I) communications, demonstrating how LiDAR can significantly improve beam management under fast-changing mobility conditions. In addition, in [30], the authors proposed a LiDAR-aided channel model for vehicular intelligent sensing-communication integration by developing a geometry-based stochastic modeling approach that uses

LiDAR point clouds to characterize static and dynamic scatterers under varying traffic conditions. This work focused on channel modeling rather than direct channel prediction. In another relevant study [31], the authors introduced an AI-based sensor attack detection and classification framework for autonomous vehicles in a 6G-V2X environment, using GPS and LiDAR to detect and classify sensor anomalies for enhanced cybersecurity. Additionally, [32] introduced MVX (Multimodal V2X), which is a configurable co-simulation framework that integrates CARLA (Car Learning to Act) and Sionna simulators to generate a multimodal multi-view V2X dataset, with an AI-based solution for future LoS blockages and optimal beam direction prediction, outperforming traditional methods in prediction accuracy and stability. The findings briefly reviewed above highlight that LiDAR-based perception is emerging as a powerful tool to address the rapid channel variations in future high-frequency vehicular networks.

These studies demonstrate the potential of multimodal data and collaborative perception in enriching the information base for AI models, thereby enhancing their prediction capabilities. Collaborative perception has emerged as a powerful strategy in V2X communications, particularly in autonomous driving applications. Authors in [14] introduced a cooperative perception framework utilizing a vision Transformer (V2X-ViT) to fuse information across on-road agents' LiDAR data. This framework addresses common V2X challenges such as asynchronous information sharing and pose errors, setting new benchmarks for 3D object detection using simulated datasets generated using CARLA [33] and OpenCDA [34] simulators.

C. CONTRIBUTION

Despite the advancements highlighted in the related work, several key limitations remain, which we address in our proposed solution.

First, while existing AI models have shown promise in channel prediction, they primarily rely on single-modal data, mainly the historical channel estimates, which limits the robustness of the predictions in highly dynamic 6G V2X environments. Our solution addresses this gap by integrating LiDAR data with historical channel estimates, offering a more comprehensive and real-time understanding of the environment, thereby enhancing prediction accuracy. Second, most prior works focus on data collected from individual agents, which can lead to incomplete or biased channel predictions, especially in complex urban settings with multiple obstacles. By leveraging MVX [32], the collaborative perception co-simulation framework, our approach synergizes data from multiple connected agents and infrastructure elements, leading to a richer and more reliable representation of the environment. This collaborative approach is crucial for capturing the diverse and rapidly changing conditions that impact channel behavior in 6G V2X networks. Finally, the prediction horizon in existing channel prediction solutions is typically fixed and adapted to the mobility of the network, although its precise value is often unspecified [15]. Fixed prediction horizons can lead to frequent, unnecessary predictions, leading to communication latency, or outdated channel states, which can cause high channel prediction errors, which degrade communication performance. Our solution introduces a dynamic prediction horizon optimization (PHO) component that adjusts the prediction interval in real time based on the current channel conditions and anticipated variability leveraging the multimodal collaborative perception data. This ensures that the predictions remain accurate and relevant, thereby improving overall communication reliability and efficiency.

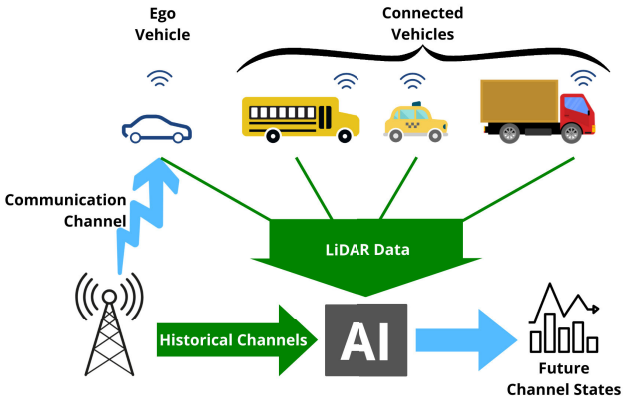
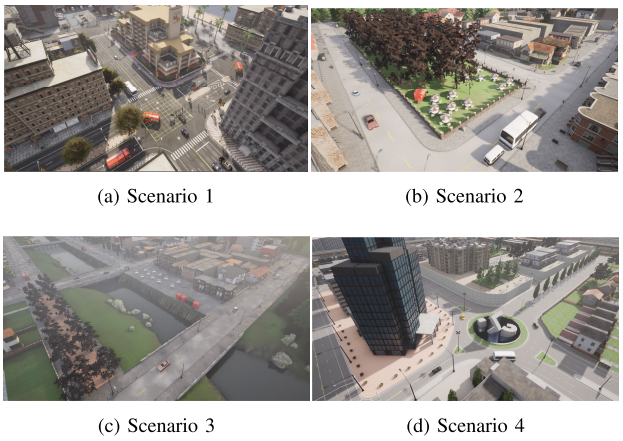


FIGURE 1. The framework of online channel prediction in 6G V2X networks using multimodal collaborative perception data.

In this paper, we address these limitations by presenting the first multimodal collaborative perception AI solution for 6G V2X channel prediction, as illustrated in Figure 1. Our approach utilizes vision transformers (ViTs) and convolutional neural networks (CNNs), integrated with a dynamic PHO component, and leverages the MVX configurable co-simulation framework. This framework simulates the 6G V2X environment with remarkable realism, enabling the generation of high-fidelity multimodal datasets for the training and testing of the proposed framework. Using this high-fidelity simulation, we thoroughly evaluate the performance of our multimodal collaborative perception solution and the PHO component throughout multiple simulations and ablation studies in various 6G V2X scenarios. The results demonstrate the excellent performance of our proposed channel prediction framework, significantly reducing prediction errors, and showcasing the effectiveness of the PHO component in dynamically adjusting the prediction interval, thereby ensuring communication reliability and efficiency in highly dynamic 6G V2X environments. All source code and data used in our experiments will be available at [35].

II. SYSTEM AND CHANNEL MODELS

In this section, we outline the system model's key components and architecture, which integrates the MVX co-simulation environment with a multimodal data-driven



(a) Scenario 1 (b) Scenario 2

(c) Scenario 3 (d) Scenario 4

FIGURE 2. Considered scenarios in MVX co-simulation.

approach for 6G V2X collaborative perception channel prediction. This comprehensive framework enables an accurate simulation of vehicular dynamics and wireless channel propagation, facilitating the generation of high-fidelity data necessary for training and evaluating advanced AI models.

A. SYSTEM MODEL

To replicate a highly realistic physical environment using the CARLA simulator, we employ the MVX framework [32], to simulate traffic scenarios—including vehicular and pedestrian movements—within a virtual urban landscape. The MVX co-simulation framework is a highly configurable and scalable platform that integrates the CARLA simulator for detailed physical and sensor modeling with the Sionna differentiable ray tracing simulator for wireless channel simulations. It generates a comprehensive multimodal dataset by capturing high-resolution LiDAR and camera data, along with accurate ground truth annotations and wireless channel responses. This simulation can create complex road networks, accurately representing real-world 6G V2X environments. Moreover, it also allows vehicles to be equipped with a variety of sensors, including LiDAR, cameras, GPS positioning, and speed information, thereby ensuring that a wide range of data is available for advanced V2X applications.

We select four different representative autonomous driving scenarios, illustrated in Figure 2, with distinct weather conditions. The diversity of the maps used is essential to validate the stability and robustness of the AI solution. Each scenario involves multiple Roadside Units (RSUs) and vehicles equipped with LiDAR sensors. The vehicles are initialized in specific locations at the beginning of each simulation round, with one vehicle selected as the target user communicating with the RSU and the other vehicles acting as connected agents. In this context, equipping the RSUs and vehicles with LiDAR sensors enhances the amount of 3D information about the environment collected from different perspectives, facilitating the development of collaborative perception solutions.

Benefiting from the help of other agents and RSUs, interconnected autonomous vehicles can expand their perceptual fields to gather information about invisible areas. The number of connected agents is set randomly, leading to diverse interactions and data collection scenarios. This setup allows for comprehensive testing and validation of our proposed AI solution for 6G V2X channel prediction, which uses multimodal collaborative perception—the process of integrating information from multiple sensing modalities and from different networked agents to form a richer, more robust representation of the dynamic vehicular environment.

B. CHANNEL MODEL

We use the Sionna simulator, which is integrated within the MVX co-simulation framework, to simulate the propagation of radio signals in 6G networks using ray tracing. This simulation accurately models the transmission and reception of signals between base stations, vehicles, and connected agents, providing both time and frequency domains channel impulse responses (CIRs) that are used for link-level simulations for LoS and NLoS conditions. MVX offers a comprehensive simulation environment that generates high-quality datasets for training and testing predictive models.

We consider an OFDM MIMO wireless communication system with S sub-carriers, N_t transmit antennas, and N_r receive antennas. Let $x(t)$ be the transmit signal at time instant t and $h(t)$ denote the temporal evolution of the channel. The received signal $y(t)$ including additive noise $n(t)$ is given by:

$$y(t) = h(t) * x(t) + n(t). \tag{1}$$

The Sionna ray tracing simulator calculates LoS and NLoS propagation paths between all transmitters and receivers. The simulation resolution is controlled by setting the maximum number of interactions a ray can have with scene objects. Diffuse reflections or scattered paths are determined by randomly sampling directions after each interaction with a scene object. The Paths object contains comprehensive information about each identified path, facilitating the generation of channel impulse response and the application of Doppler shifts to simulate time evolution. The channel frequency response $H(f)$ for each path i at carrier frequency f is computed as the sum of the M propagation paths generated by ray tracing, as follows:

$$H(f) = \sum_{i=1}^{M} a_i e^{-\mathrm{j}2\pi f \tau_i}, \tag{2}$$

where $a_i \in \mathbb{C}$ and τ_i are the complex coefficient and delay describing the ith path, respectively. Moreover, a_i is defined as follows:

$$a_i = \frac{\lambda}{4\pi} \mathbf{C}_{\mathrm{R}} \left(\theta_i^{\mathrm{R}}, \varphi_i^{\mathrm{R}}\right)^{\mathrm{H}} \mathbf{T}_i \mathbf{C}_{\mathrm{T}} \left(\theta_i^{\mathrm{T}}, \varphi_i^{\mathrm{T}}\right), \tag{3}$$

where λ is the wavelength, $\left(\theta_i^{\mathrm{T}}, \varphi_i^{\mathrm{T}}\right)$ and $\left(\theta_i^{\mathrm{R}}, \varphi_i^{\mathrm{R}}\right)$ denote the angles of departure and arrival of the i-th path, respectively, $\mathbf{T}_i : \mathbb{C}^2 \mapsto \mathbb{C}^2$ represents the transfer function for the i-th path., and $\mathbf{C}_{\mathrm{T}}(\theta, \varphi) \in \mathbb{C}^2$ and $\mathbf{C}_{\mathrm{R}}(\theta, \varphi) \in \mathbb{C}^2$ denote

the radiation patterns of the transmit and receive antennas, respectively. This formulation is adapted from [36]. Given the dynamic nature of the V2X network, all moving vehicles are assigned a velocity vector, $\mathbf{v} \in \mathbb{R}^3$ (collected from CARLA simulation). Based on this, we compute the path-specific Doppler shifts f_Δ and apply them to all paths, resulting in time-dependent path coefficients as follows:

$$a(t) = ae^{j2\pi f_\Delta t}. \tag{4}$$

In real-world V2X networks, it is not feasible to access the channel's historical evolution to predict future channel states. Therefore, for a more realistic simulation, we consider imperfect Channel State Information (CSI) where, instead of training the model on perfect channel information, we use pilot-based channel estimations to predict the future channel state. The channel estimation process in this work consists of two main steps: least-squares (LS) estimation at pilot-carrying resource elements, followed by interpolation for data-carrying resource elements using linear minimum mean square error (LMMSE). The LS estimation model is described as follows:

$$\mathbf{y} = \mathbf{h} \odot \mathbf{p} + \mathbf{n}, \tag{5}$$

where $\mathbf{y} \in \mathbb{C}^M$ is the received signal vector, $\mathbf{p} \in \mathbb{C}^M$ is the vector of pilot symbols, $\mathbf{h} \in \mathbb{C}^M$ is the channel vector to be estimated, and $\mathbf{n} \in \mathbb{C}^M$ is a zero-mean noise vector with variance N_0. The operator $\odot$ denotes element-wise multiplication. The LS channel estimate $\hat{\mathbf{h}}$ is computed as:

$$\hat{\mathbf{h}}_{\text{LS}} = \mathbf{y} \odot \frac{\mathbf{p}^*}{|\mathbf{p}|^2} = \mathbf{h} + \hat{\mathbf{n}}, \tag{6}$$

where $\hat{\mathbf{n}}$ represents the noise component in the estimate. The LMMSE interpolation method necessitates the estimation of the time, frequency, and spatial covariance matrices of the channel. These covariance matrices are derived using Monte Carlo sampling. To estimate the frequency covariance matrix, we generate multiple samples of the channel model to create a set of frequency-domain channel realizations $\{\mathbf{h}_k\}$, $1 \leq k \leq K$, where K is the total number of samples and $\mathbf{h}_k \in \mathbb{C}^N$ represents complex-valued channel frequency response samples. The frequency covariance matrix $\mathbf{R}(f) \in \mathbb{C}^{N\times N}$ is then computed as follows:

$$\mathbf{R}(f) \approx \frac{1}{K}\sum_{k=1}^{K} \mathbf{h}_k \mathbf{h}_k^H. \tag{7}$$

The same methodology is applied to estimate the time and spatial covariance matrices. In practical applications, interpolation is first performed across the time, followed by frequency, and then spatial dimensions. The LMMSE interpolation is then expressed as:

$$\hat{\mathbf{H}}_{\text{LMMSE}} = \mathbf{R}_{\text{combined}} \left(\mathbf{R}_{\text{combined}} + \Sigma_{\text{combined}}\right)^{-1} \hat{\mathbf{H}}_{\text{LS}}, \tag{8}$$

where $\mathbf{R}_{\text{combined}}$ is the combined covariance matrix incorporating the frequency, time, and spatial covariance matrices of the channel, and Σ_{combined} is the combined diagonal matrix of channel estimation error variances. By implementing this detailed channel model, we ensure a realistic simulation and prediction of the channel state, which is crucial for optimizing communication in 6G V2X networks.

III. PROBLEM FORMULATION

In this section, we define the optimization problem of predicting future channel states in 6G V2X networks using historical channel states and collaborative LiDAR data points. The goal is to maximize the channel prediction accuracy while minimizing the pilot overhead and communication latency.

Let $\mathbf{H}_t$ represent the channel matrix at time t, $\hat{\mathbf{H}}_t$ the estimated channel obtained through pilot-based channel estimation, and $\tilde{\mathbf{H}}_t$ the predicted channel at time t. In the context of multimodal collaborative data, let N_a represent the number of connected agents at time t. Each agent j provides a set of LiDAR data points $\mathbf{L}_t^j$. The overall collected LiDAR data points $\mathbf{L}_t$ from the connected agents at a time step t can be represented as:

$$\mathbf{L}_t = \{\mathbf{L}_t^1, \mathbf{L}_t^2, \ldots, \mathbf{L}_t^{N_a}\}. \tag{9}$$

We define two sets: $\hat{\mathcal{H}} = \{\hat{\mathbf{H}}_{t_i}\}_{i=0}^{N-1}$ being the history of estimated channel states and $\mathcal{L} = \{\mathbf{L}_{t_i}\}_{i=0}^{N-1}$ denoting the history of collected LiDAR data points, where t_i denotes the time indices where we performed channel estimation and LiDAR data collection, and N is the set size. Our objective is to predict the future channel matrix $\{\mathbf{H}_{t+k}\}_{k=1}^{\tau}$, leveraging both the last p elements from these historical sets, denoted as $\hat{\mathcal{H}}_{\text{last }p}[t] = \{\hat{\mathbf{H}}_{t_{N-1-i}}\}_{i=0}^{p-1}$ and $\mathcal{L}_{\text{last }p}[t] = \{\mathbf{L}_{t_{N-1-i}}\}_{i=0}^{p-1}$.

The optimization problem can be formulated as a time series problem where the last p estimated channel states and LiDAR data points are used to predict the future τ channel states. In order to improve the robustness of our channel prediction model, we adopt a relative prediction error (normalized error) approach. This method minimizes the prediction error relative to the magnitude of the actual channel states, thereby making the model more resilient to variations in the channel state magnitudes. Mathematically, this can be represented as follows:

$$\begin{aligned} &\underset{\theta}{\text{minimize}}\ \mathbb{E}\left[\frac{\sum_{k=1}^{\tau}\left\|\mathbf{H}_{t+k} - \tilde{\mathbf{H}}_{t+k}\right\|^2}{\sum_{k=1}^{\tau}\|\mathbf{H}_{t+k}\|^2}\right], \\ &\text{s.t.}\ \left(\tilde{\mathbf{H}}_{t+1}, \cdots, \tilde{\mathbf{H}}_{t+\tau}\right) = f_\theta\left(\hat{\mathcal{H}}_{\text{last }p}[t], \mathcal{L}_{\text{last }p}[t]\right), \end{aligned} \tag{10}$$

where f_θ is a function, parameterized by θ, that maps the historical sequence of the channel states and the LiDAR data points to the predicted future channel states. The objective function minimizes the cumulative mean squared error (MSE) over the prediction horizon τ between the actual future channel states $\mathbf{H}_{t+k}$ and the predicted channel states $\tilde{\mathbf{H}}_{t+k}$.

This hybrid approach, involving both channel estimation and prediction, mitigates the latency introduced by the need to frequently transmit pilot signals. The essential advantage is not the complete elimination of this latency but

rather the optimization of the frequency of these estimations. By integrating channel prediction, the system can extend the intervals between necessary estimations in comparison to classic channel estimation approaches. This extension, known as the prediction horizon, which is the future time interval over which a model forecasts channel conditions based on current environmental data, allows predictive models to anticipate future channel states. In existing channel prediction solutions, the prediction horizon is typically adjusted based on the mobility of the network, although its precise value is often unspecified [15]. It is important to note that if the prediction horizon is shorter than the pilot interval of the classic pilot-based channel estimation method, the system's predictive capabilities could be compromised, as the frequency of channel estimations would be equal to or greater than that of the classic pilot-based channel estimation approach, introducing high communication latency. On the other hand, in highly dynamic and rapidly changing channel conditions, an excessively large prediction horizon could lead to inaccurate channel predictions, potentially compromising communication quality.

Therefore, given its critical importance, we consider optimizing the prediction horizon by dynamically adjusting it based on the current channel conditions. This adjustment aims to optimize the accuracy-latency trade-off, enhancing the overall efficiency of the channel prediction component. To this end, we introduce a dynamic prediction horizon approach that adapts the prediction interval based on the current, historical, and predicted evolution of the channel state and collected LiDAR data points. This component is trained to predict the optimal prediction horizon τ_t^* at a time step t, defined as the highest value of τ that keeps the MSE of the channel prediction lower than a predefined maximum threshold. Mathematically, this can be represented as follows:

$$\tau_t^* = \max_{\tau} \tau \text{ subject to } \mathbb{E}\left[\frac{\sum_{k=1}^{\tau} \left\| \mathbf{H}_{t+k} - \tilde{\mathbf{H}}_{t+k} \right\|^2}{\sum_{k=1}^{\tau} \left\| \mathbf{H}_{t+k} \right\|^2}\right] \leq \epsilon, \tag{11}$$

where ϵ is defined based on the maximum acceptable channel prediction error in 6G V2X communications. The PHO component is trained to predict τ_t^*, mathematically, this can be represented as follows:

$$\begin{aligned} &\underset{\phi}{\text{minimize}} \mathbb{E}\left[(\tau_t^* - \tilde{\tau}_t(\phi))^2\right], \\ &\text{s.t. } \tilde{\tau}_t = g_\phi\left(\hat{\mathcal{H}}_{\text{last } p}[t], \mathcal{L}_{\text{last } p}[t], \left(\tilde{\mathbf{H}}_{t+1}, \cdots, \tilde{\mathbf{H}}_{t+\tau}\right)\right), \end{aligned} \tag{12}$$

where g_ϕ is a function parameterized by ϕ that maps the historical and predicted evolution of the channel state, along with the collected LiDAR data points to the predicted prediction horizon. After the computation of $\tilde{\tau}_t$ at a time step t, we only consider the predicted channels for the future $\tilde{\tau}_t$ time steps $\left(\tilde{\mathbf{H}}_{t+1}, \cdots, \tilde{\mathbf{H}}_{t+\tilde{\tau}_t}\right)$.

IV. PROPOSED SOLUTION

In this paper, we introduce an AI-based multimodal collaborative perception solution for channel prediction using the MVX co-simulation environment, which provides ray-traced wireless channels and LiDAR data from multiple perspectives, including vehicles and infrastructure perspectives. Our approach conceptualizes V2X perception as a heterogeneous multi-agent perception system, where various agents, such as base stations and vehicles, simultaneously perceive their surroundings and communicate with each other. This collaborative approach aims to provide a more comprehensive understanding of the environment, facilitating more accurate predictions.

The overall architecture of our proposed framework is illustrated in Figure 3. The framework comprises four key components. The first component is data sequence preparation, which involves generating and organizing data sequences from the MVX co-simulation environment, including both wireless channel information and LiDAR point clouds. The second component, feature extraction and sharing, focuses on extracting relevant features from the collected data and sharing them among different agents to ensure a holistic perception of the environment. The third component of our framework, V2X-ViT, is a dedicated vision transformer specifically designed for V2X collaborative perception [14]. This component processes the shared features to extract useful information, which enhances the system's overall perception capabilities. The final stage of the framework involves the prediction head component, which includes the main GRU model responsible for predicting future channel states. Following a comprehensive comparison in [15], which identified GRUs as the most effective time-series model for wireless channel prediction, in this study, we adopt a GRU architecture for all baseline and proposed evaluations. Although alternative architectures may yield different absolute performance levels, the present study focuses on isolating and quantifying the benefits of multimodal data fusion and prediction horizon optimization. Additionally, the PHO component dynamically determines the optimal prediction horizon to be considered, ensuring stable and accurate predictions.

The following subsections detail the proposed framework, along with the constraints and objectives that guide our efforts to enhance communication efficiency and reliability in complex urban environments.

A. FEATURES EXTRACTION

Traditional channel prediction approaches often rely on statistically simulated channels from the perspective of a single network element, typically a base station [15]. This is understandable as it is less costly and time-consuming than collecting real-world data. However, these mathematical models are not realistic because they often fail to capture the intricate and dynamic interactions between radio waves and the complex environments in which they propagate.

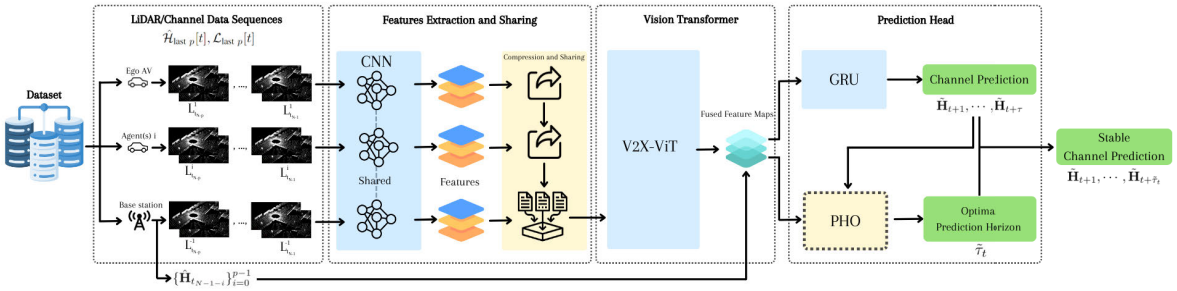


FIGURE 3. **Overview of our proposed V2X multimodal collaborative perception-based channel prediction framework.**

Specifically, statistical models may oversimplify factors such as multipath reflections, scattering from various objects, mobility of both transmitters and receivers, and environmental changes over time.

To address these limitations, we employ a ray tracing-based simulation of the wireless channels, which provides a realistic and deterministic approach for wireless communication simulations. In a real-world implementation of channel prediction, estimated channel states are used to conduct the predictions. In this context, the channel measurement and estimation process is inevitably affected by noise [15]. The survey in [15] investigated the impact of noise and demonstrated that it can significantly influence the performance evaluation of prediction methods. To replicate a realistic scenario and account for uncertainties from channel estimation, we distort the channel with a Gaussian noise. In our simulations, the variance of the additive Gaussian noise is set to $\sigma_n^2 = 0.01$. This approach ensures that the model is trained with realistic noisy channel data.

Our approach to collaborative perception in dynamic V2X networks emphasizes the synergistic interactions among vehicles (agents) and base stations. Each vehicle in the network actively participates in data sharing by transmitting a rich set of features extracted from LiDAR data points to the base station. The comprehensive data collection from multiple agents ensures the base station has a holistic view of the network environment. To clarify the mapping process from LiDAR data to the wireless channel, our framework starts by extracting spatial features from the high-resolution LiDAR point clouds. These point clouds capture detailed environmental information, including obstacles, terrain variations, and the dynamic positions of surrounding objects. By fusing these LiDAR-derived features with conventional wireless measurements, the proposed model effectively captures complex propagation phenomena, including blockages and scattering effects, thus enhancing the accuracy and robustness of channel state predictions in dynamic 6G V2X scenarios. Feature extraction from the collected data is a crucial component of our system. We adopt the PointPillars method [37], an anchor-based technique known for efficiently handling point cloud data, reducing computational demand, optimizing memory [38], and transforming raw LiDAR point clouds into a structured format. The structured data, resembling a 2D pseudo-image, is processed through the PointPillars backbone to yield informative feature maps. These feature maps, represented as $\{F_i[t]\}_{i=1}^{N_a}$ at a time step t for agent i ($i = -1$ for the base station), contain essential spatial information that is then transmitted to the base station. At the core of our collaborative perception framework is the V2X-ViT solution [14] for V2X collaborative perception, which simultaneously covers V2V and V2I communication. The V2X-ViT solution introduces an innovative heterogeneous multi-agent self-attention (HMSA) module 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, 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, effectively addressing feature misalignments that may arise due to localization errors and time delays. Additionally, the HMSA and MSwin modules facilitate the capture of inter- and intra-agent interactions. This results in an enriched, aggregated fused feature map $I[t] = \{I_i[t]\}_{i=1}^{N_a}$.

The final stage in our perception pipeline is the prediction head component, which receives the fused feature maps and applies the GRU time series and PHO models to predict the future channel state for the optimized future time interval.

B. PREDICTION HEAD

Channel prediction aims to design the mapping function f_θ, as defined in (10). This paper proposes a GRU-based model to construct this mapping function and a PHO component to determine the optimal prediction horizon. The scheme of the proposed GRU-based channel prediction framework consists of two phases: offline training, presented in Algorithm 1, and online prediction in Algorithm 2.

1) GRU Model

For the offline training phase, a supervised learning algorithm is applied to train the model to optimize (10). At each time step t, the actual channel labels for the next τ time steps $\{\mathbf{H}_{t+i}\}_{i=1}^{\tau}$ and the past p τ-step historical channels $\{\mathbf{H}_{t-(i\tau)}\}_{i=0}^{p-1}$ are extracted from the ray-traced channels to

Algorithm 1 Offline Training

Initialize: $\theta, \phi, \tau, p, \epsilon, \sigma$
Input: Ray-traced channels and LiDAR dataset

TRAIN GRU:
while GRU not converged **do**
  for each input data sequence **do**
    Compute $\{\hat{\mathbf{H}}_{t-(i\tau)}\}_{i=0}^{p-1}$ using (8)
    Extract fused feature map $I[t]$ from $\{\mathbf{L}_{t-(i\tau)}\}_{i=0}^{p-1}$
    $\{\tilde{\mathbf{H}}_{t+i}\}_{i=1}^{\tau} \leftarrow f_\theta\left(\{\hat{\mathbf{H}}_{t-(i\tau)}\}_{i=0}^{p-1}, I[t]\right)$
    Compute the loss $\mathcal{L}_{\text{GRU}}$ using (10)
    Backpropagate the gradients $\nabla_\theta \mathcal{L}_{\text{GRU}}$
    Update θ
  end for
end while

TRAIN PHO:
while PHO-GRU not converged **do**
  for each input data sequence **do**
    Compute τ_t^* using (11)
    $\tilde{\tau}_t \leftarrow g_\phi\left(\{\hat{\mathbf{H}}_{t-(i\tau)}\}_{i=0}^{p-1}, I[t], \{\tilde{\mathbf{H}}_{t+i}\}_{i=1}^{\tau}\right)$
    Compute the loss $\mathcal{L}_{\text{PHO}}$ using (12)
    Backpropagate the gradients $\nabla_\phi \mathcal{L}_{\text{PHO}}$
    Update ϕ
  end for
end while
Return: θ and ϕ

Algorithm 2 Online Prediction (at a Time Step t)

Input: Ray-traced channel $\mathbf{H}_t$, LiDAR data $\mathbf{L}_t$
for each time step t **do**
  Compute $\hat{\mathbf{H}}_t$ using (8)
  Update historical set of channel estimations $\hat{\mathcal{H}}$ with $\hat{\mathbf{H}}_t$
  Update historical set of LiDAR data points $\mathcal{L}$ with $\mathbf{L}_t$
  Extract last p elements $\hat{\mathcal{H}}_{\text{last }p}[t]$ and $\mathcal{L}_{\text{last }p}[t]$
  $\{\tilde{\mathbf{H}}_{t+i}\}_{i=1}^{\tau} \leftarrow f_\theta\left(\hat{\mathcal{H}}_{\text{last }p}[t], \mathcal{L}_{\text{last }p}[t]\right)$
  $\tilde{\tau}_t \leftarrow g_\phi\left(\hat{\mathcal{H}}_{\text{last }p}[t], \mathcal{L}_{\text{last }p}[t], \{\tilde{\mathbf{H}}_{t+i}\}_{i=1}^{\tau}\right)$
end for
Return: $\{\tilde{\mathbf{H}}_{t+i}\}_{i=1}^{\tilde{\tau}_t}$

form a training sample. Next, we add Gaussian noise and apply the LS-LMMSE channel estimation model defined in (8) to obtain the input channel state estimations for the last p τ-step channel estimations $\{\hat{\mathbf{H}}_{t-(i\tau)}\}_{i=0}^{p-1}$. Using these channel state estimations, along with the aggregated fused feature map $I[t]$ extracted from the collected LiDAR data points $\{\mathbf{L}_{t-(i\tau)}\}_{i=0}^{p-1}$, we train the GRU model to predict the future τ channel states $\{\tilde{\mathbf{H}}_{t+i}\}_{i=1}^{\tau}$, as detailed in Algorithm 1.

After training the transformer-based feature extraction architecture and the GRU channel prediction model, we deploy the trained framework for online prediction. During the online prediction phase, at each time step t, as presented in Algorithm 2, we compute the noisy LS-LMMSE channel estimation $\hat{\mathbf{H}}_t$ using the ray-traced channel $\mathbf{H}_t$. This channel estimation is added to the historical set of channel estimations $\hat{\mathcal{H}}$, and the collected LiDAR data points from the connected agents are added to the historical set of LiDAR data points $\mathcal{L}$. The last p elements of the historical set of channel state estimations and collected LiDAR data points, $\hat{\mathcal{H}}_{\text{last }p}[t]$ and $\mathcal{L}_{\text{last }p}[t]$, are then used as input to our framework to predict the future τ channel states $\{\tilde{\mathbf{H}}_{t+i}\}_{i=1}^{\tau}$.

At this point, the PHO component filters the predicted future channel states to adjust the prediction horizon dynamically, ensuring optimal performance and reliability in the V2X network environment.

2) PHO

In our proposed channel prediction framework, we introduce an intelligent and adaptive PHO component. The primary function of PHO is to dynamically adjust the prediction horizon τ based on the anticipated variability of channel conditions. This dynamic adjustment ensures an optimal balance between prediction accuracy and communication latency, which is essential for the demanding environments characteristic of 6G V2X communications.

The PHO component operates by analyzing three critical inputs: historical CSI estimations, collected LiDAR data points, and the initial channel predictions for the next τ time steps generated by the main GRU model. By processing these inputs, PHO employs a dedicated GRU model to assess the current channel state in terms of its stability and fluctuation patterns. This assessment allows PHO to predict the optimal prediction horizon τ^*. When the PHO component predicts high variability or instability in the channel conditions, it reduces the prediction horizon to enhance the accuracy and reliability of the predictions. A shorter prediction horizon enables more frequent updates, thereby enhancing the accuracy and reliability of the predictions in rapidly changing environments. Conversely, if the channel is expected to remain stable, PHO keeps the prediction horizon extended. This decreases the frequency of channel estimations, thereby reducing both computational load and communication latency associated with pilot-based channel measurements. This adaptive approach offers several key advantages. By tailoring the prediction horizon to future channel conditions, PHO ensures that the system maintains high levels of prediction accuracy without experiencing unnecessary computational or communication overhead. In highly dynamic vehicular environments, where channel conditions can change unpredictably, the ability to adjust τ in real-time is crucial for sustaining reliable and efficient communication links.

During offline learning, as detailed in Algorithm 1, we compute the optimal prediction horizon τ^* using (11) for each time sequence. Subsequently, the PHO GRU model is trained to provide the predicted prediction horizon $\tilde{\tau}_t$ using the past p τ-step channel estimations $\{\hat{\mathbf{H}}_{t-(i\tau)}\}_{i=0}^{p-1}$, the

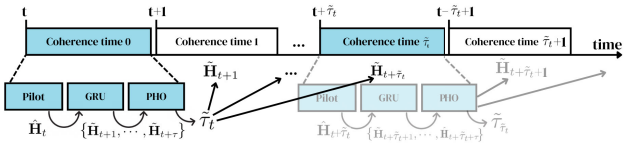


FIGURE 4. Channel prediction timeline at time step t.

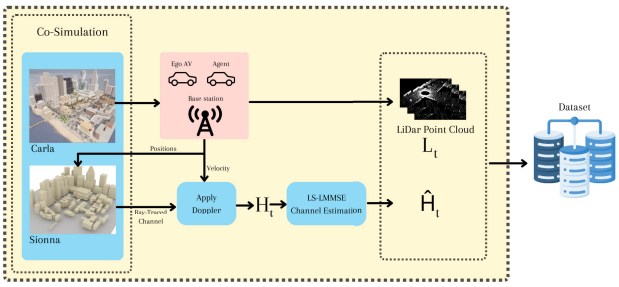


FIGURE 5. Multimodal data generation process using MVX co-simulation framework.

aggregated fused feature map $I[t]$, and the predicted future τ channel states $\{\tilde{\mathbf{H}}_{t+i}\}_{i=1}^{\tau}$, according to (12). After training the PHO component, during online prediction presented in Algorithm 2 at a time step t, the GRU model performs the prediction of the future τ channel states $\{\tilde{\mathbf{H}}_{t+i}\}_{i=1}^{\tau}$. The PHO component then filters these predictions and provides only the future $\tilde{\tau}_t$ predicted channel states. The process is then repeated at time step $t + \tilde{\tau}_t$, as depicted in Figure 4.

In the following sections, we elaborate on the implementation of the proposed framework.

V. IMPLEMENTATION

In this section, we detail the implementation of the co-simulation framework used for data generation and the proposed multimodal collaborative perception framework for channel prediction.

A. SIMULATION SCENARIO AND DATA GENERATION

To train our framework, we employ a high-fidelity simulation environment that integrates both physical and wireless simulations provided by the MVX framework. The Carla simulator creates realistic scenarios and generates high-resolution LiDAR data from multiple vehicles and infrastructure perspectives. This is combined with the Sionna simulator, which offers differentiable ray tracing simulations for wireless channels, as depicted in Figure 5.

1) PHYSICAL ENVIRONMENT CONFIGURATION

The dataset used in our study is generated through co-simulations conducted across 60 unique scenes on four distinct maps using the Carla simulator in random weather conditions. Each scene lasts exactly 50 seconds and includes two base stations and a random number of intelligent agents.

TABLE 1. Implementation parameters.

Parameter	Value
LiDAR Sensors Parameters [39]	
Channels	64
Range	120m
Rotation Frequency	20
Points Generated per Second	10^6
Antennas Parameters [40] [4] [41]	
Number of rows and columns	8x8
Carrier Frequency	5.9 GHz
Wavelength λ	50.8 mm
Vertical and Horizontal Antenna Spacing	25.4 mm
Antenna Pattern	3GPP TR 38.901 model
Type of Polarization	dual polarization "VH"
Requirements of 6G V2X [42]	
Latency	$\leq$ 10 ms
Reliability	$\geq$ 99%
Data Rate	$\geq$ 100 Mbps
Max. Latency	$\leq$ 20 ms
Transmission Rate	$\geq$ 10 messages/sec
V2X-ViT Architecture Parameters [14]	
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

In each map, illustrated in Figure 2, the ego vehicle follows a predefined route and collects sensor data synchronously at each time step. Upon completion of a scenario, the simulator is reset to start a new round. A random number of smart agents simultaneously collect data, while other traffic participants, such as cars and pedestrians, are initially positioned randomly and then controlled using CARLA's AI mechanisms. LiDAR data is acquired using sensors with a 120-meter range, configured according to the specifications in Table 1. These sensors are mounted on each vehicle and on the top of each RSU, using ray-casting to simulate rotating LiDAR. Due to the small displacement of vehicles over tens of milliseconds, we assume that the LiDAR data points remain unchanged for 100 ms [14]. Consequently, we record LiDAR point clouds at a frequency of 10 Hz. Additionally, we collect ground truth position and velocity vectors of the vehicles participating in the simulation.

2) WIRELESS ENVIRONMENT CONFIGURATION

The previously mentioned physical world scenarios from CARLA are coupled with equivalent maps in the Sionna simulator using MVX co-simulation framework, as shown in the example presented in Figure 6, to run ray tracing simulations for outdoor sub-6 GHz communication environments.

The ray tracing simulation quality is tuned by adjusting the maximum number of interactions (or bounces) that rays can have with objects in the scene. Due to the stochastic nature of the shoot-and-bounce ray tracing algorithm, multiple runs can yield different path calculations. To ensure reproducibility, we fixed the random seed in TensorFlow.

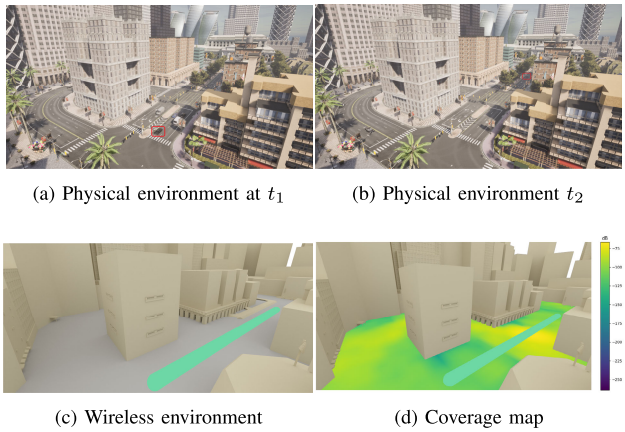
Following the 3rd Generation Partnership Project (3GPP) standards, we configure our ray tracing simulation with the planar array antenna parameters specified in Table 1 to generate the required channel dataset for our framework. The duration of each time step δt is critical to the stability and accuracy of channel predictions. Typically, papers assume that the simulation time step is much less than the coherence time, which is not always evident in highly dynamic 6G V2X networks operating at high frequencies. To ensure that the predicted channel states remain valid and do not become outdated, we set the time step duration to the minimum possible coherence time. The coherence time T_c is defined as the duration over which the channel impulse response is considered invariant. It is inversely proportional to the Doppler spread f_Δ of the channel, which is influenced by the relative velocity of the transmitter and receiver and the carrier frequency f_c. The coherence time can be approximated by:

$$T_c \approx \frac{1}{f_\Delta} \tag{13}$$

Given that $f_\Delta = \frac{vf_c}{c}$, where v is the relative velocity, f_c is the carrier frequency, and c is the speed of light, we set the time step to the minimum possible coherence time calculated based on the maximum expected relative velocity of users and carrier frequency in our scenarios. This ensures that the channel state remains relatively stable within each time step, allowing for accurate predictions. For urban scenarios, we consider a maximum vehicle speed of 60 km/h. For rural scenarios, we consider a maximum speed of 150 km/h.

According to recent V2X research [4], the sub-6 GHz band offers robust and reliable links essential for safety-critical applications and cooperative sensing. Modern vehicles are increasingly equipped with various sensors, and the sub-6 GHz spectrum remains integral for ensuring stable connectivity, particularly in scenarios with high mobility and in obstructed environments. Our choice of 5.9 GHz as a baseline frequency is underpinned by this evidence and aligns with existing V2X testbeds. While 6G is expected to harness mmWave or even sub-THz frequencies for ultra-high throughput, our proposed multimodal collaborative perception framework can be extended to higher bands by adapting the underlying channel models. Assuming a carrier frequency f_c of 5.9 GHz, the minimum coherence time duration is approximately 3.05 ms for urban scenarios and 1.22 ms for rural scenarios. Thus, we set the time step duration δt to 1 ms. This value ensures that the channel state remains relatively stable within each time step, enabling accurate predictions.

URLLC requirements in emerging 6G scenarios frequently specify latencies below 1ms and reliability approaching 99.999% for mission-critical services. However, many V2X applications, such as cooperative sensing, can tolerate higher latencies and lower reliability targets [42]. Accordingly, the adopted 10ms latency and 99% reliability reflect a broader class of V2X use cases. Yet, our proposed framework could, in principle, support stricter URLLC targets; however, this would also demand end-to-end optimization of prediction-horizon, coding, and queueing to meet sub-1 ms / $>$ 99.999% requirements, which we leave for future investigation.



(a) Physical environment at t_1 (b) Physical environment t_2

(c) Wireless environment (d) Coverage map

FIGURE 6. One of the MVX co-simulation environments used for the data generation and framework training, showing the movement of the ego vehicle bounded in a red box during a time interval starting at t_1 and finishing at t_2.

Finally, the dataset at each simulation time step is linked with the corresponding LiDAR data from the Carla simulator. The combined dataset 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 prepared, we begin the training, validation, and testing of our framework. The implementation details are as follows. All models are trained using Nvidia Quadro P1000 GPUs.

1) FEATURES EXTRACTION

The V2X-ViT [14] vision transformer is an open-source tool designed for LiDAR collaborative perception in autonomous driving, recognized for its effectiveness in detecting 3D bounding boxes of connected vehicles. This tool includes a pre-trained model with the parameters listed in Table 1. For our framework, we utilize a fine-tuned version of the V2X-ViT model. Initially, raw LiDAR point clouds are processed into stacked pillar tensors, which are then converted into 2D pseudo-images. These pseudo-images are fed into the PointPillars backbone, which is configured according to the parameters in Table 1. This backbone extracts informative feature maps, which are then shared among participant vehicles and the base station to facilitate further feature processing. The aggregated features from the connected agents

are fed into the V2X-ViT model, which was proposed by Xu et al. [14]. The V2X-ViT model uses self-attention mechanisms to perform iterative inter-agent and intra-agent feature fusion. It is configured with three encoder layers and window sizes of 4, 8, and 16 in the MSwin module. We employ 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 shown in Table 1. The output features from this stage, embedded with the history of estimated channel states, are then passed to the next component of our framework, i.e. the prediction head.

2) PREDICTION HEAD

We train the GRU and PHO components separately within the prediction head. We begin by training the GRU time series model using embedded features from the data sequences to forecast the future τ channel states. The number of layers of the model and the value of τ are set after experimental testing, where we investigate the performance of the GRU model using different values of τ and different number of layers. We set τ to the maximum possible value that ensures model stability, which can then be dynamically adjusted by the PHO component. This approach allows us to dynamically shorten the prediction horizon based on real-time evaluations by the PHO component, ensuring the flexibility and adaptability of our framework to varying network conditions. To train the PHO component, we compute the optimal prediction horizon τ^* based on the parameter ϵ, which represents the minimum acceptable MSE. This parameter is determined in accordance with the stringent requirements of 6G V2X networks. We set the maximum acceptable MSE ϵ to 0.01, based on the required signal-to-noise ratio (SNR) and bit error rate (BER) for reliable communication. By setting ϵ to 0.01, we ensure that the predicted channel states maintain the necessary accuracy, supporting the high reliability and low latency demands of 6G V2X communications.

With these implementation parameters defined, we now proceed to the next section to present the experimental results and evaluation of our multimodal collaborative perception framework for channel prediction in 6G V2X networks.

VI. RESULTS AND ANALYSIS

In this section, we present our experimental results and evaluate the performance of the proposed framework under diverse vehicular scenarios. For the sake of comparison, we benchmark our solution against two baseline methods: a classical LS-LMMSE approach and a channel prediction model that relies solely on conventional wireless measurements (without LiDAR data), thus providing a comprehensive performance assessment under common vehicular channel prediction settings.

A. FEATURE EXTRACTION EVALUATION

The effectiveness of the feature extraction component and ViT model is pivotal to the overall accuracy of our proposed framework. To evaluate the performance of the ViT model, we assess its capability to accurately detect and localize the bounding boxes of both the ego vehicle and connected agents within our V2X network simulation. This evaluation leverages the merged collaborative perception LiDAR data points generated by the feature extraction component. We employ the Intersection-over-Union (IoU) metric to quantify detection accuracy, where IoU is defined as the ratio of the area of overlap to the area of union between the predicted bounding boxes and the ground truth bounding boxes. The ViT model utilized in our framework is initially pre-trained on the V2XSet dataset [14] for the task of 3D bounding box detection. To tailor the model to the specific characteristics of our V2X network simulation environment, we fine-tune the pre-trained ViT model using our custom dataset. This fine-tuning process allows the model to adapt to the unique spatial and environmental features present in MVX simulation scenarios.

TABLE 2. Detection performance of ViT model before and after Fine-tuning.

Model	IoU (%)	Precision (%)	Recall (%)
Pre-trained ViT	72.1	74.5	70.3
Fine-tuned ViT	83.6	85.7	78.9

Our evaluation results demonstrate a significant improvement in detection accuracy post-fine-tuning. As shown in Table 2, the IoU score of the ViT model increases by 11.5%, achieving an IoU of 83.6%. This enhancement indicates that the fine-tuned model exhibits a superior ability to accurately detect and localize both the ego vehicle and connected agents within the V2X network. The increased IoU score reflects better alignment between predicted and actual bounding boxes, thereby validating the effectiveness of our feature extraction and model fine-tuning processes.

The improvement in detection accuracy is crucial for the success of our overall architecture. Accurate bounding box detection ensures that the collaborative perception system has reliable information about the positions and movements of vehicles and agents within the network. This, in turn, enhances the quality of the estimated CSI used for channel prediction, leading to more reliable communication links in dynamic vehicular environments. Furthermore, the robust performance of the ViT model in detecting bounding boxes under varying conditions within the simulation underscores the resilience and adaptability of our feature extraction framework. By accurately capturing the spatial dynamics and environmental context through LiDAR data, our framework effectively supports the intricate requirements of 6G V2X communications.

B. CHANNEL PREDICTION PERFORMANCE

We start by evaluating the performance of the GRU channel prediction model. The evaluation focuses on determining the optimal number of GRU layers and analyzing the potential overfitting of the model. To identify the optimal number of GRU layers, we set the values of p and τ to 20 and

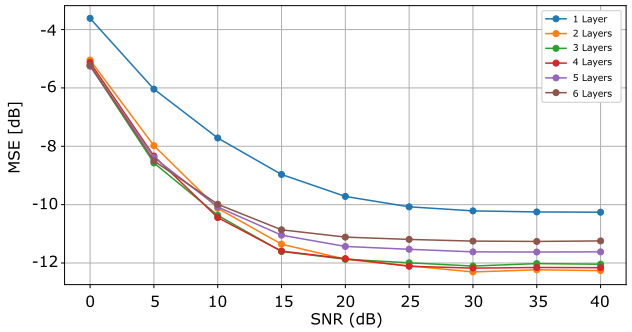


FIGURE 7. MSE of GRU model with varying number of layers for different SNR values.

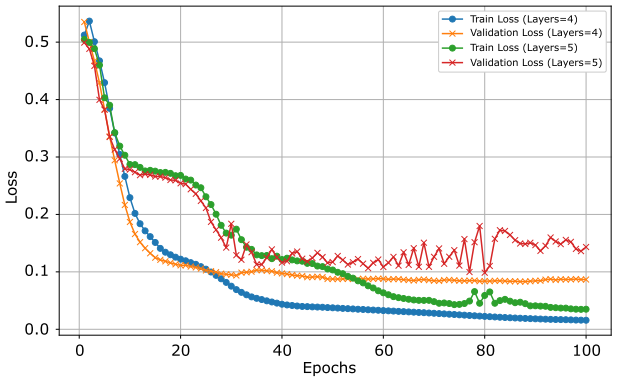


FIGURE 8. Training and validation loss evolution over epochs for the 4-layer and 5-layer models.

10-time steps, respectively, and conducted extensive experiments with different layer configurations, varying from a 1-layer model to a 6-layer model. The MSE for each configuration was evaluated across different SNR values. Of note, while the simulation scenario involves multi-user communication, the SNR values presented in Figure 7 correspond specifically to the ego vehicle.

The results in Figure 7 indicate that increasing the number of GRU layers initially leads to a reduction in overall MSE. However, beyond 4 layers, the MSE starts to increase. This increase is caused by model overfitting, where the model learns to perform well on the training data but fails to generalize to unseen data. Overfitting occurs when the model is too complex and starts capturing noise in the training data as if it were a true signal.

To confirm overfitting, we analyzed the loss performance evolution for both the training and validation datasets over multiple epochs, comparing the 4-layer and 5-layer models as depicted in Figure 8.

In Figure 8, both the training and validation loss of the 4-layer model decrease over the epochs, indicating that the model is effectively learning and not overfitting. Conversely, while the training loss of the 5-layer model continues to decrease, the validation loss starts to increase after a certain number of epochs. This confirms that the 5-layer model is overfitting, learning to perform well on the training data but

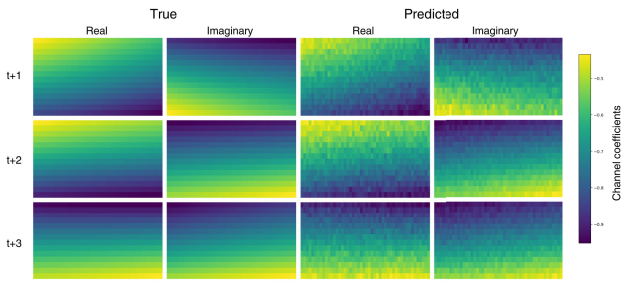


FIGURE 9. Visualization of an example of true and predicted future channels.

failing to generalize. Therefore, for the remainder of the evaluation experiments, we will consider a 4-layer architecture for the GRU model.

Further analysis of our model's performance involves visualizing and comparing the past p channels, the true future τ channels, and the predicted future τ channels. Each channel is visualized as a 2D image, showing the real and imaginary parts of the channel frequency response for each transmit antenna and subcarrier frequency, SNR is fixed at approximately $20dB$, which is achieved by setting the transmission power for each user to $30dBm$.

The visual comparison in Figure 9 demonstrates that the GRU model effectively captures the temporal, spatial, and frequency dynamics of the channel state. The predicted channels closely align with the true future channels, highlighting the model's ability to accurately forecast future channel states. This capability is crucial for optimizing communications in 6G V2X networks, ensuring high reliability and low latency. In summary, the 4-layer configuration was found to be optimal, offering a balance between model complexity and predictive performance.

C. PREDICTION HORIZON EVALUATION

To assess the impact of the prediction horizon on the performance of our channel prediction solution, in Figure 10, we conduct extensive simulations to compare the MSE of channel predictions across different values of τ for 1000 simulations in both urban and rural environments.

As shown in Figure 10, the MSE increases steadily with the prediction horizon. This trend can be attributed to the weakening temporal correlation of the channel as the prediction horizon extends, making accurate predictions increasingly challenging. When the prediction horizon approaches 10 ms, the performance of the GRU model reaches a plateau, indicating that the coherence time of the channel is around 10 ms. The rate at which the channel decorrelates is influenced by the mobility of the vehicles. This explains the earlier saturation of MSE in the rural scenario compared to the urban scenario. In rural environments, higher vehicle speeds result in increased Doppler shifts, leading to shorter coherence times and thus lower prediction performance for the GRU model.

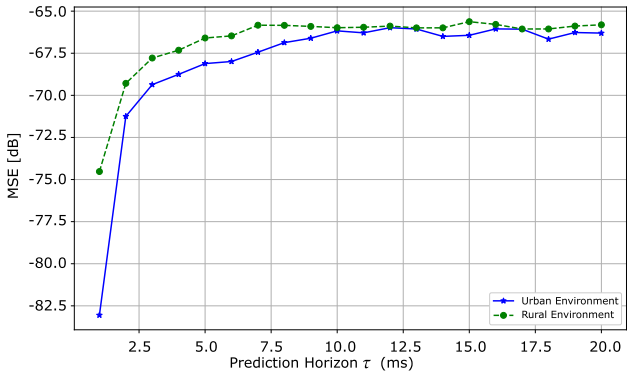


FIGURE 10. MSE versus prediction horizons τ in urban and rural environments.

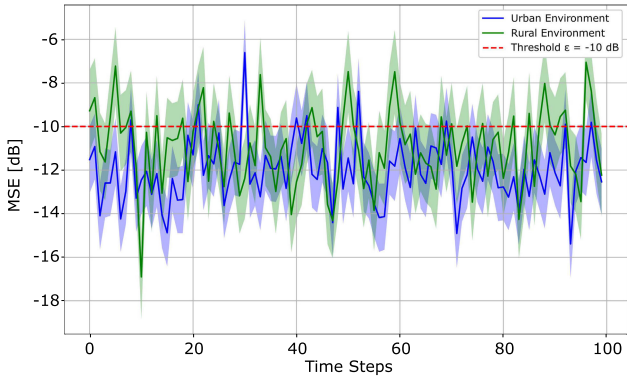


FIGURE 11. Temporal evolution of the channel prediction MSE in rural and urban environments.

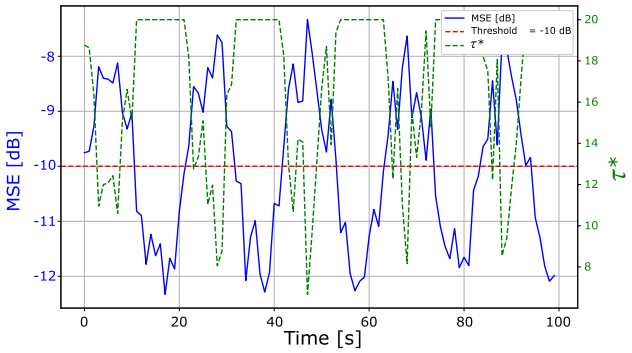


FIGURE 12. Evolution of the optimal prediction horizon (τ*) and the MSE.

To further examine the effect of the prediction horizon on the channel prediction performance, we analyze the temporal evolution of the MSE for both rural and urban simulations. Figure 11 shows these results, where the plotted line represents the mean MSE value and the shaded area indicates the observed variance across the simulations.

Fluctuations in MSE caused by the dynamic nature of the channel are observed. High MSE spikes correspond to highly dynamic channel conditions where the prediction horizon is too long. To address this, we set a maximum acceptable MSE (ϵ), compute the optimal prediction horizon τ^* using (11), and train the GRU model of the PHO component to dynamically adjust the prediction horizon to ensure that the MSE does not exceed this limit.

Figure 12 illustrates the evolution of both the optimal prediction horizon (τ^*) and the MSE. The figure shows a negative correlation between these variables: lower MSE values correspond to higher τ^* values (close to the maximum prediction horizon τ), indicating accurate channel predictions, while higher MSE values result in lower τ^* values, confirming that the inaccuracies occur at the end of the prediction horizon.

In highly dynamic regions, the channel behavior becomes chaotic and unpredictable, necessitating more frequent predictions to maintain. This explains the sharp spikes in the MSE. The PHO component learns to predict optimal prediction horizons, enabling the system to adapt to these sensitive areas and limit the prediction head to ensure reliable communication.

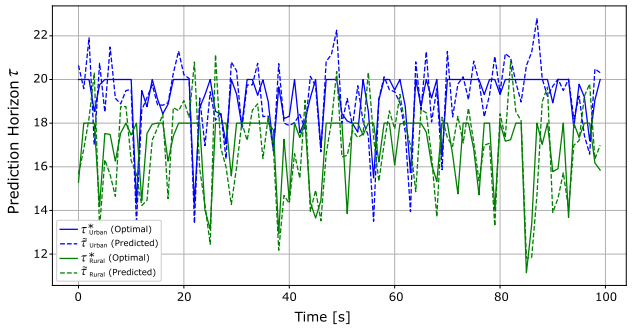
Figure 13 (a) shows the evolution of the optimal and predicted prediction horizons τ^* and $\tilde{\tau}$ for both rural and urban scenarios. The results demonstrate that the GRU model of the PHO component effectively predicts the optimal prediction horizon, with the predicted values closely matching the optimal ones. The PHO component successfully anticipates most of the sudden drops in the optimal prediction horizon with minimal error, proving its utility in preventing high prediction errors in rapidly changing channel conditions.

Figure 13 (b) compares the performance of our framework with and without the PHO component. The results indicate that the PHO component effectively predicts regions where the MSE exceeds the predefined limit (ϵ) and helps to improve channel prediction performance by reducing the spikes in the MSE, keeping them at or below the acceptable threshold. This demonstrates the significant added value of the PHO component in enhancing the accuracy and reliability of our channel prediction framework for 6G V2X networks.

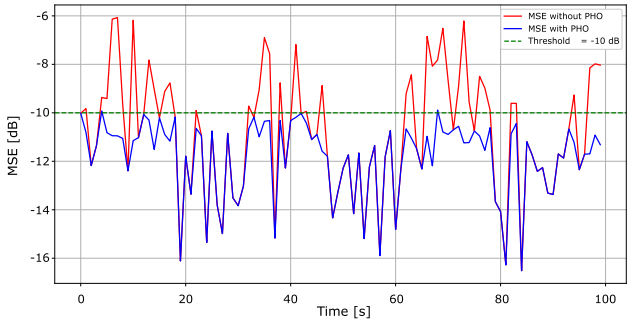
D. LINK-LEVEL COMMUNICATION PERFORMANCE EVALUATION

For a practical and realistic evaluation of the proposed channel prediction framework, we define a site-specific link-level communication scenario between the ego vehicle, multiple connected agents, and the base station. Using ray-traced CIRs from our simulations, we compare different MIMO detection algorithms and perform a BER evaluation for multi-user MIMO OFDM transmission in the uplink direction.

In this experiment, we adopt 16-QAM for the uplink modulation scheme. We consider the K-Best MIMO detection algorithm [43], using various channel estimation and prediction methods. First, we implement standard pilot-based channel estimation using LS-LMMSE interpolation, where the estimated channel is assumed to remain fixed for a duration equivalent to the channel's coherence time. Second,



(a) Evolution of τ^* and $\tilde{\tau}$



(b) Impact of the use of the PHO on the MSE evolution.

FIGURE 13. (a) Evolution of the optimal and predicted prediction horizons for rural and urban scenarios. (b) Performance comparison of the framework with and without the PHO component.

we evaluate a naive channel prediction method that considers the estimated channel as the predicted channel for the next τ time steps. Third, we consider a perfect CSI approach, where the K-Best detector has perfect knowledge of the channel state at each time step. Finally, we evaluate our channel prediction framework with and without the PHO component, denoted as VGG and VGG-τ, respectively. VGG follows the arrangement defined in Figure 4, where channel estimation and prediction are performed at each time step, and the PHO determines the number of future time steps during which data transmission is performed using the GRU-predicted channels. In contrast, VGG-τ automatically uses the predicted τ channels for the next τ time steps.

Figure 14 illustrates the BER evolution for each of the considered methods across different SNR values.

As expected, the naive method exhibits the worst BER performance, maintaining almost uniformly high BER values across all SNR ranges. This confirms the dynamic nature of the channel in such environments and the inadequacy of simplistic prediction methods for high-frequency communications.

Conversely, our predictive solution, both with and without the PHO component, significantly outperforms the standard channel estimation approach. Notably, the predictive solution with the PHO component achieves remarkably lower BER values at low SNR levels. This improvement is attributed

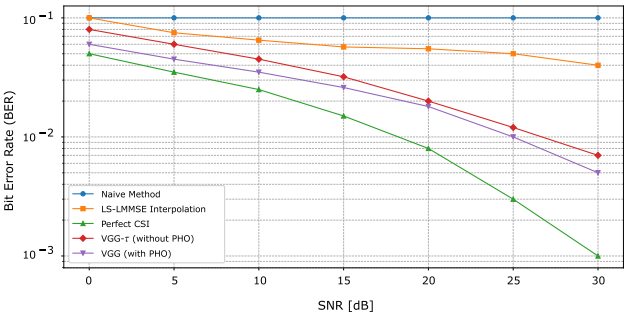


FIGURE 14. BER performance for different SNR values using various channel estimation and prediction methods.

to the PHO's ability to dynamically adjust the prediction horizon, thus mitigating the channel prediction inaccuracies at the end of the prediction horizon. In low SNR conditions, the predicted channels towards the end of the prediction horizon tend to be less accurate. The PHO component in the VGG framework effectively adjusts the prediction horizon to prevent BER performance degradation. Consequently, VGG consistently achieves lower BER values compared to VGG-τ, highlighting the efficacy of the PHO component in enhancing communication quality under challenging conditions.

The analysis of these results indicates that our proposed predictive framework, particularly when equipped with the PHO component, provides substantial benefits in terms of communication reliability and quality. By dynamically adjusting the prediction horizon based on real-time evaluations, the PHO component ensures that the channel state predictions remain accurate, thereby maintaining low BER levels and supporting the stringent requirements of 6G V2X networks. This adaptability is crucial for ensuring robust and efficient communications in highly dynamic and high-frequency vehicular environments.

E. ABLATION STUDY

In the proposed framework, we employ a multimodal collaborative perception solution for channel prediction. To rigorously evaluate the contributions of these key components, we conduct the following ablation studies.

1) COLLABORATIVE PERCEPTION

The proposed collaborative perception approach leverages perspectives from both the infrastructure and the connected agents. To assess the added value of the information collected from each perspective, we compare two different configurations in both urban and rural environments.

First, we consider a Vehicle-to-Vehicle (V2V) configuration in which only vehicles are equipped with sensors. Channel predictions in this setup are performed solely based on the data collected from the vehicles' perspectives. Second, we evaluate a V2X configuration where infrastructural elements, such as base stations, are also equipped with sensors and contribute to data collection. For each of these

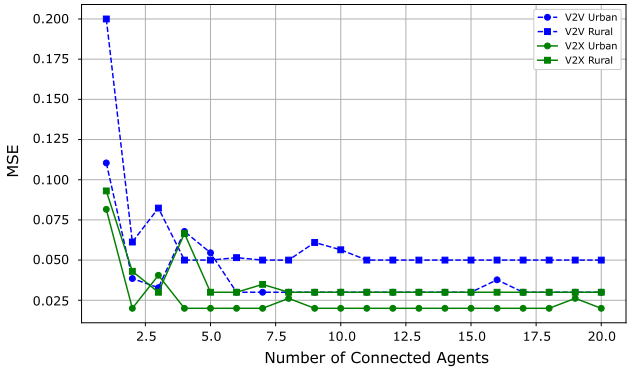


FIGURE 15. Impact of the number of connected agents on the channel prediction MSE in V2V and V2X configurations.

configurations, we measure the channel prediction MSE across different numbers of connected agents.

As illustrated in Figure 15, an increase in the number of agents improves the performance of both V2V and V2X configurations. Notably, the V2X setup consistently outperforms the V2V configuration in terms of channel prediction accuracy. This enhancement can be attributed to the infrastructure's perspective, which typically experiences fewer obstructions and provides a broader and less occluded view of the environment. The richer amount of 3D information about the surroundings, obtained from this broader perspective, results in more insightful features which are crucial for accurate predictions. However, the higher quality of channel predictions achieved with the V2X setup comes with an increased computational cost. In the subsequent evaluation, we further investigate the trade-offs between prediction accuracy and computational efficiency to ensure the proposed framework's practical applicability in real-world 6G V2X networks.

2) MULTIMODAL DATA

Integrating LiDAR data has proven to be a significant enhancement in our AI-based multimodal collaborative perception channel prediction framework. To evaluate the impact of LiDAR data, we conduct a series of experiments analyzing the performance of channel prediction in diverse vehicular environments.

First, we investigate the role of the LiDAR data in providing spatial awareness and environmental context. We set up experiments in an urban environment with numerous static and dynamic obstacles, comparing two GRU models: one using only wireless data and one utilizing both wireless and LiDAR data. The performance of both models, alongside the pilot-based channel estimation approach, is presented in Figure 16.

The results indicate that the wireless-based model exhibits the highest MSE values, while the multimodal model outperforms both, demonstrating superior performance in terms of MSE. This can be attributed to the type and amount of information each model leverages. The wireless-based

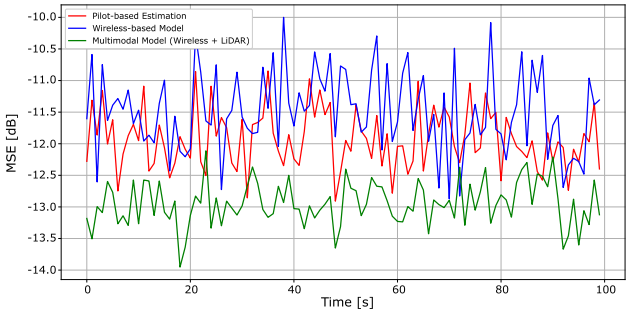


FIGURE 16. Performance comparison of wireless-based and multimodal models in terms of MSE.

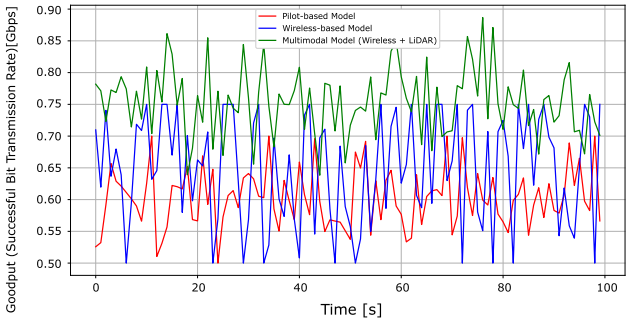


FIGURE 17. Goodput comparison over time for wireless-based, pilot-based, and multimodal models.

TABLE 3. Performance comparison of the proposed solution against baseline methods.

Method	MSE [dB]	Goodput [Gbps]
LS-LMMSE	−11.91	0.58
Wireless-based Model	−11.43	0.63
Proposed	−13.03	0.77

model relies solely on past estimated channels to predict future channel states, inherently limiting its accuracy to that of the channel estimation approach. In contrast, the multimodal model incorporates LiDAR data, providing three-dimensional spatial information that complements the wireless data. This spatial awareness helps identify obstacles and reflectors that influence signal propagation, reflection, diffraction, and scattering, thereby refining path loss predictions and modeling multi-path components with greater accuracy.

To provide a more comprehensive evaluation, we perform experiments assessing both the accuracy and latency of the two models by analyzing the evolution of the goodput over time, defined as the rate of successful bit transmissions. As shown in Figure 17, the wireless model achieves higher goodput rates than the channel estimation model in some periods. Despite its lower channel prediction accuracy, the wireless model benefits from reduced communication latency due to the anticipation of channel states and the use of fewer pilots.

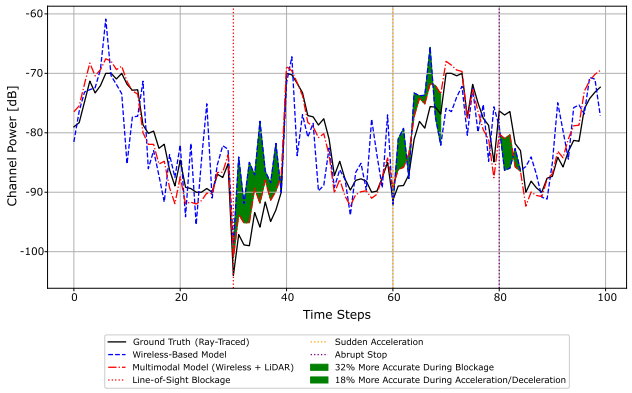


FIGURE 18. Evolution of channel power for the ground truth, wireless-based, and multimodal models.

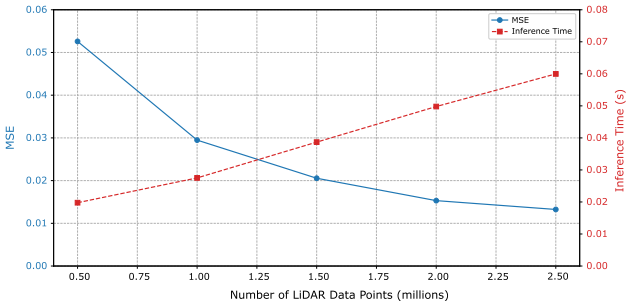


FIGURE 19. Effect of varying the number of LiDAR data points per second on the MSE and inference time.

The multimodal model, however, consistently outperforms both approaches. By incorporating LiDAR data, it achieves higher goodput rates through accurate anticipation of channel state fluctuations, effectively minimizing both communication latency and BER. The results presented in Table 3 indicate that, in terms of MSE, the proposed solution achieves a 22.73% linear-scale improvement compared to the LS-LMMSE baseline and a 30.82% improvement over the wireless-based model. Likewise, it demonstrates a 32.76% boost in goodput relative to LS-LMMSE and a 22.22% increase compared to the wireless-based baseline.

To further understand the superior performance of the multimodal model compared to the wireless approach, we analyzed the temporal dynamics captured by sequential LiDAR scans. We simulated a busy urban intersection, observing how the movement of vehicles and obstacles influences the predicted channels. Figure 18 shows the evolution of the ground truth ray-traced channel power, the predicted channel power using a wireless-based model, and the predicted channel power using a multimodal model that combines wireless and LiDAR data.

During the simulation, we apply three actions to test the models' responses: blocking the line of sight between the transmitter antenna and the ego user using a connected agent, applying a sudden high acceleration to the ego user, and executing an abrupt stop of the ego user's movement. Figure 18 shows that by integrating temporal LiDAR data, our model achieves 18% greater overall accuracy in predicting channel variations and is 32% more accurate during line-of-sight blockage events. This enhanced accuracy is attributed to the LiDAR's ability to track object movements, providing real-time information about rapid changes that affect channel characteristics. These experiments demonstrate that LiDAR data significantly enhances the channel prediction model by providing detailed spatial and temporal information, enriching the feature sets for better situational awareness. The integration of LiDAR data thus proves indispensable for achieving accurate and reliable channel predictions in dynamic vehicular environments.

3) COMPUTATION STUDY

To investigate the trade-off between LiDAR-based channel prediction accuracy and computational complexity, we conducted an experiment where we varied the number of LiDAR data points generated from 0.5M to 2.5M per second. As shown in Figure 19, while a higher LiDAR point density yields a lower Mean Squared Error (MSE), it also increases inference time. Specifically, with an increase in the point count, MSE decreases from 0.2 to 0.07, indicating more accurate environmental modeling and, consequently, improved channel predictions. However, this improvement comes at the cost of higher computational overhead, as inference time rises from 0.3s to 1.5s.

This highlights the inherent trade-off when determining LiDAR sampling density in resource-constrained V2X systems. In scenarios where real-time performance is paramount, a moderate reduction in LiDAR points can substantially decrease processing overhead, with only a negligible loss in prediction accuracy. Conversely, applications demanding the highest possible fidelity may opt for denser point clouds, accepting thus a higher computational burden. Overall, the results of our experiment underscore the flexibility of the proposed multimodal perception framework, which can be configured to effectively balance accuracy and efficiency based on specific system requirements.

In terms of real-world deployment, the proposed multimodal collaborative perception framework shows promising performance improvements in simulation. However, several challenges remain when scaling to larger networks or more diverse environments. Processing high-density LiDAR data requires substantial computational resources, which may require techniques such as downsampling or edge-cloud processing to ensure real-time performance in large-scale V2X networks. Moreover, communication overhead associated with exchanging multimodal sensor data can be significant, so efficient data compression and scheduling strategies are necessary. The need for adaptive learning methods and robust system design is further underscored by variations in traffic density and environmental conditions. Future work should focus on addressing these challenges, ensuring that the proposed framework not only delivers superior performance in controlled simulations, but also effectively scales in diverse

real-world vehicular networks. It is also important to emphasize that the MVX co-simulation stack couples CARLA's physics-grade vehicle and sensor engine with Sionna's differentiable ray-tracing, providing centimetre-level LiDAR accuracy, 3GPP-compliant MIMO antenna patterns, and Doppler-aware channel impulse responses. While this digital twin captures the dominant spatial and temporal effects of sub-6 GHz V2X links, several real-world factors remain abstracted: (i) sensor artifacts under rain/fog and LiDAR occlusion by dirty lenses, (ii) RF front-end impairments and beam-tracking latency, and (iii) distribution shift when the trained model is deployed on road geometries unseen in the simulated scenes.

VII. CONCLUSION

In this paper, we introduced a groundbreaking multimodal collaborative perception framework for channel prediction in 6G V2X networks, integrating ViTs and CNNs to harness diverse data sources. Our approach has demonstrated significant advancements in prediction accuracy and communication reliability—most notably, reducing mean squared error by up to 30.82% and increasing goodput by up to 32.76% as compared to conventional baselines—thereby addressing the unique challenges posed by highly dynamic and high-frequency vehicular environments inherent in 6G V2X systems. The impact of our solution on 6G V2X networks is profound. By incorporating multimodal data, particularly LiDAR, our framework provides a richer, more comprehensive understanding of the environment. This enhanced situational awareness allows for more accurate prediction of channel variations, which is critical to achieving URLLC required by 6G V2X applications. In scenarios where milliseconds can determine the success of safety-critical communications, the ability to predict and adapt to channel conditions with high precision is indispensable.

The inclusion of the PHO component in our framework further amplifies its impact by dynamically adjusting the prediction horizon based on real-time evaluations of channel conditions and the environment understanding. This adaptability ensures that the system remains robust even in the face of rapidly changing environments, optimizing the trade-off between prediction accuracy and communication latency. As a result, our solution not only enhances the reliability of V2X communications but also contributes to the overall efficiency and safety of autonomous driving and intelligent transportation systems. Moreover, our extensive experimental analysis, including rigorous ablation studies, validated the contributions of our collaborative perception approach. The results confirmed that leveraging infrastructural perspectives and intelligently utilizing multimodal data are crucial for achieving superior channel prediction in 6G V2X networks. Our framework sets a new standard for future research, demonstrating that advanced AI-driven solutions can effectively address the complexities of next-generation vehicular communications.

In conclusion, the multimodal collaborative perception framework we proposed has the potential to significantly influence the development of 6G V2X networks, ensuring that these systems can meet the demanding requirements of future vehicular applications. This work paves the way for further innovations in AI-powered wireless communications, driving progress towards more reliable, efficient, and intelligent transportation networks.

REFERENCES

[1] P. Porambage, G. Gür, D. P. M. Osorio, M. Liyanage, A. Gurtov, and M. Ylianttila, "The roadmap to 6G security and privacy," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 1094–1122, 2021.

[2] W. Jiang, B. Han, M. A. Habibi, and H. D. Schotten, "The road towards 6G: A comprehensive survey," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 334–366, 2021.

[3] P. S. R. Henrique and R. Prasad, *6G The Road to the Future Wireless Technologies 2030*. River Publishers, 2021, pp. 1–26.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] C.-X. Wang, J. Huang, H. Wang, X. Gao, X. You, and Y. Hao, "6G wireless channel measurements and models: Trends and challenges," *IEEE Veh. Technol. Mag.*, vol. 15, no. 4, pp. 22–32, Dec. 2020.

[9] Y. Wei, M.-M. Zhao, A. Liu, and M.-J. Zhao, "Channel tracking and prediction for IRS-aided wireless communications," *IEEE Trans. Wireless Commun.*, vol. 22, no. 1, pp. 563–579, Jan. 2023.

[10] H. Kim, J. Choi, and D. J. Love, "Massive MIMO channel prediction via meta-learning and deep denoising: Is a small dataset enough?" *IEEE Trans. Wireless Commun.*, vol. 22, no. 12, pp. 9278–9290, Dec. 2023.

[11] F. Peng, S. Zhang, Z. Jiang, X. Wang, and W. Chen, "A novel mobility induced channel prediction mechanism for vehicular communications," *IEEE Trans. Wireless Commun.*, vol. 22, no. 5, pp. 3488–3502, May 2023.

[12] C. Luo, J. Ji, Q. Wang, X. Chen, and P. Li, "Channel state information prediction for 5G wireless communications: A deep learning approach," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 1, pp. 227–236, Mar. 2020.

[13] D. K. Pin Tan et al., "Integrated sensing and communication in 6G: Motivations, use cases, requirements, challenges and future directions," in *Proc. 1st IEEE Int. Online Symp. Joint Commun. Sens.*, Feb. 2021, pp. 1–6.

[14] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, and J. Ma, "V2X-ViT: Vehicle-to-everything cooperative perception with vision transformer," in *Computer Vision—ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., Cham, Switzerland: Springer, 2022, pp. 107–124.

[15] O. Stenhammar, G. Fodor, and C. Fischione, "A comparison of neural networks for wireless channel prediction," *IEEE Wireless Commun.*, vol. 31, no. 3, pp. 235–241, Jun. 2024.

[16] J. Huang, C. X. Wang, L. Bai, J. Sun, and Y. Yang, "A big data enabled channel model for 5G wireless communication systems," *IEEE Trans. Big Data*, vol. 6, no. 2, pp. 211–222, Jun. 2020.

[17] Y. Yang, Y. Li, W. Zhang, F. Qin, P. Zhu, and C.-X. Wang, "Generative-adversarial-network-based wireless channel modeling: Challenges and opportunities," *IEEE Commun. Mag.*, vol. 57, no. 3, pp. 22–27, Mar. 2019.

[18] H. Xiao, W. Tian, W. Liu, and J. Shen, "ChannelGAN: Deep learning-based channel modeling and generating," *IEEE Wireless Commun. Lett.*, vol. 11, no. 3, pp. 650–654, Mar. 2022.

[19] Z. Liu, Z. Teng, Y. Song, X. Ye, and Y. Ouyang, "Channel modeling and generation: Train generative networks and generate 6G channel data," in *Proc. IEEE 8th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2022, pp. 72–78.

[20] K. Mao et al., "Machine-learning-based 3-D channel modeling for U2V mmWave communications," *IEEE Internet Things J.*, vol. 9, no. 18, pp. 17592–17607, Sep. 2022.
[21] L. Bai, Q. Xu, S. Wu, S. Ventouras, and G. Goussetis, "A novel atmosphere-informed data-driven predictive channel modeling for B5G/6G satellite-terrestrial wireless communication systems at Q-band," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 14225–14237, Dec. 2020.
[22] Z. Li, C.-X. Wang, J. Huang, W. Zhou, and C. Huang, "A GAN-LSTM based AI framework for 6G wireless channel prediction," in *Proc. IEEE 95th Veh. Technol. Conf. (VTC-Spring)*, Jun. 2022, pp. 1–5.
[23] Z. Li, C.-X. Wang, C. Huang, L. Yu, J. Li, and Z. Qian, "A novel scatterer density-based predictive channel model for 6G wireless communications," in *Proc. IEEE 97th Veh. Technol. Conf. (VTC-Spring)*, Jun. 2023, pp. 1–5.
[24] Z. Li et al., "A GAN-GRU based space-time predictive channel model for 6G wireless communications," *IEEE Trans. Veh. Technol.*, vol. 73, no. 7, pp. 9370–9386, Jul. 2024.
[25] Y. Yang, F. Gao, C. Xing, J. An, and A. Alkhateeb, "Deep multimodal learning: Merging sensory data for massive MIMO channel prediction," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 1885–1898, Jul. 2021.
[26] Y. Yang, F. Gao, C. Xing, J. An, and A. Alkhateeb, "Sensory data assisted downlink channel prediction for massive MIMO," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2021, pp. 1–6.
[27] G. Gharsallah and G. Kaddoum, "ViT LoS V2X: Vision transformers for environment-aware LoS blockage prediction for 6G vehicular networks," *IEEE Access*, vol. 12, pp. 133569–133583, 2024.
[28] 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.
[29] S. Jiang, G. Charan, and A. Alkhateeb, "LiDAR aided future beam prediction in real-world millimeter wave V2I communications," *IEEE Wireless Commun. Lett.*, vol. 12, no. 2, pp. 212–216, Feb. 2023.
[30] Z. Huang, L. Bai, M. Sun, and X. Cheng, "A LiDAR-aided channel model for vehicular intelligent sensing-communication integration," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 12, pp. 20105–20119, Dec. 2024.
[31] M. Begum, G. Raja, and M. Guizani, "AI-based sensor attack detection and classification for autonomous vehicles in 6G-V2X environment," *IEEE Trans. Veh. Technol.*, vol. 73, no. 4, pp. 5054–5063, Apr. 2024.
[32] G. Gharsallah and G. Kaddoum, "MVX-ViT: Multimodal collaborative perception for 6G V2X network management decisions using vision transformer," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 5619–5634, 2024.
[33] 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.
[34] R. Xu, Y. Guo, X. Han, X. Xia, H. Xiang, and J. Ma, "OpenCDA: An open cooperative driving automation framework integrated with co-simulation," in *Proc. IEEE Int. Intell. Transp. Syst. Conf. (ITSC)*, Sep. 2021, pp. 1155–1162.
[35] MVX. (2025). *Multimodal Collaborative Perception co-Simulation for 6G V2X*. [Online]. Available: https://ghazigh.github.io/MVX
[36] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[37] 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)*, Jun. 2019, pp. 12689–12697.
[38] 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)*, May 2022, pp. 2583–2589.
[39] 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. (ACCV)*, Jan. 2022, pp. 4361–4377.
[40] *Study on Evaluation Methodology of New Vehicle-to-Everything V2X Use Cases for LTE and NR; (Release 15)*, document TR 37.885, 3GPP, Sophia Antipolis, France, 2018.
[41] S. B. Prathiba, G. Raja, and N. Kumar, "Intelligent cooperative collision avoidance at overtaking and lane changing maneuver in 6G-V2X communications," *IEEE Trans. Veh. Technol.*, vol. 71, no. 1, pp. 112–122, Jan. 2022.
[42] P. Rajalakshmi, "Towards 6G V2X sidelink: Survey of resource allocation—Mathematical formulations, challenges, and proposed solutions," *IEEE Open J. Veh. Technol.*, vol. 5, pp. 344–383, 2024.
[43] W. Fu and J. S. Thompson, "Performance analysis of K-best detection with adaptive modulation," in *Proc. Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2015, pp. 306–310.

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

GEORGES KADDOUM (Senior Member, IEEE) received the bachelor's degree in electrical engineering from the École Nationale Supérieure de Techniques Avancées (ENSTA Bretagne), Brest, France, the joint M.S. degree in telecommunications and signal processing (circuits, systems, and signal processing) from the Université de Bretagne Occidentale and Telecom Bretagne (ENSTB), Brest, in 2005, and the Ph.D. degree (Hons.) in signal processing and telecommunications from the National Institute of Applied Sciences (INSA), University of Toulouse, Toulouse, France, in 2009. He is currently a Professor and the Research Director of the Resilient Machine Learning Institute (ReMI) and the Tier 2 Canada Research Chair of École de Technologie Supérieure (TS), 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 (WIMOB), the 2017 IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), and the 2023 IEEE International Wireless Communications and Mobile Computing Conference (IWCMC). 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 TS in recognition of his outstanding research outcomes. He also won the 2022 IEEE Technical Committee on Scalable Computing (TCSC) Award for Excellence (Middle Career Researcher). He 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 serving as an Area Editor for IEEE TRANSACTIONS ON MACHINE LEARNING IN COMMUNICATIONS AND NETWORKING and an Editor for IEEE TRANSACTIONS ON COMMUNICATIONS.