SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

July 2025, Vol. 68, Iss. 7, 170305:1–170305:16
https://doi.org/10.1007/s11432-024-4423-4

Special Topic: Integration of Large AI Model and 6G

Personalizing rate-splitting in vehicular communication via large multi-modal model†

Shengyu ZHANG[1], Shiyao ZHANG[2], Weijie YUAN[3*], Jia SHI[4], Zan LI[4] & Tony Q.S. QUEK[1,5]

[1]*Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore 487372, Singapore*
[2]*School of Advanced Engineering, Great Bay University, and Great Bay Institute for Advanced Study (GBIAS), Dongguan 523000, China*
[3]*School of System Design and Intelligent Manufacturing, Southern University of Science and Technology, Shenzhen 518055, China*
[4]*State Key Laboratory of Integrated Service Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, China*
[5]*Department of Electronic Engineering, Kyung Hee University, Yongin 17104, Republic of Korea*

Received 31 October 2024/Revised 18 February 2025/Accepted 28 April 2025/Published online 20 June 2025

Abstract Achieving perfect channel state information at the transmitter (CSIT) in vehicle-to-everything (V2X) communication is often impractical due to the dynamic nature of vehicular environments and the inherent feedback and processing delays. To address this challenge, we propose a novel roadway-geometry-aware predictive beamforming design for rate-splitting multiple access (RSMA)-enabled V2X. Our approach designs an RSMA-based large multi-modal model (LMM) to enhance the fairness and robustness in complex V2X-enabled traffic systems. This innovative methodology employs spatial-temporal traffic data to significantly enhance the quality and performance of beamforming in V2X communications. Specifically, we first represent the complicated road-geometry data by a connected lane graph, and then extract the spatial features via a graph convolutional network (GCN) module. Meanwhile, we extract the temporal dependencies between the historical CSIT and the desired beamformer by leveraging a Transformer encoder module. Moreover, we incorporate road semantic information as an additional input to the decoder module, enabling the generation of a more context-aware beamforming design. Simulation results demonstrate that our proposed LMM outperforms the conventional deep learning approach and optimization approach. It significantly improves the effectiveness and robustness of V2X communication in diverse traffic conditions.

Keywords large multi-modal model, rate-splitting multiple access, predictive beamforming, road semantic, vehicle-to-everything

Citation Zhang S Y, Zhang S Y, Yuan W J, et al. Personalizing rate-splitting in vehicular communication via large multi-modal model. Sci China Inf Sci, 2025, 68(7): 170305, https://doi.org/10.1007/s11432-024-4423-4

1 Introduction

With the development of advanced applications and services within intelligent transportation systems (ITS), transportation systems have been revolutionized through the integration of advanced communication technologies, sensors, and data analytics, significantly enhancing efficiency, safety, and sustainability [1–3]. A key advancement within ITS is vehicle-to-everything (V2X) communication, which facilitates real-time data exchange among vehicles, infrastructure, and other entities in the environment [4,5]. The increasing data intensity of modern vehicles, driven by factors such as sensor data sharing and video streaming capabilities, poses significant challenges for conventional communication frameworks. Consequently, there is a pressing need for ultra-reliable low-latency V2X communications that can handle the massive volumes of data generated by these advanced vehicular technologies. To meet these demands, technologies such as massive multiple-input and multiple-output (MIMO) have emerged as fundamental enablers of next-generation V2X communications [6–8]. MIMO technology has revolutionized wireless

* Corresponding author (email: yuanwj@sustech.edu.cn)
† Special Topic: The 27th Annual Meeting of the China Association for Science and Technology

communication by enabling the simultaneous transmission of multiple data streams over the same frequency band, relying on sophisticated techniques to manage interference channels effectively [9,10]. This capability commonly necessitates accurate channel state information at the transmitter (CSIT) to optimize performance [11]. In spatial division multiple access (SDMA), multiple users share the same frequency band by utilizing spatial separation techniques [12–14], where precise CSIT becomes crucial for effective alignment of spatial beams, thereby minimizing interference between users and maximizing overall data throughput. Similarly, non-orthogonal multiple access (NOMA) employs a different strategy but also relies heavily on accurate CSIT; here, users share the same time and frequency resources, but power levels are allocated based on individual channel conditions [15]. In both SDMA and NOMA frameworks, inaccurate CSIT can be detrimental: in SDMA, imprecise CSIT may result in suboptimal beam alignment and increased interference, while in NOMA, it can lead to inefficient power distribution, resulting in resource wastage and compromised communication fairness.

Unfortunately, in the context of V2X communications, the acquisition of perfect CSIT presents a significant challenge, due to the rapid movement of vehicles and the inherent delays in feedback and processing [16–18]. The complexity of connected vehicle (CV) participation in traffic operations within V2X communication scenarios introduces additional challenges. The varying distances between the base station (BS) and the planned trajectories of the CVs lead to significant fluctuations in signal propagation characteristics. This variability further complicates the task of accurately predicting or estimating the channel state in real time, necessitating advanced methodologies to ensure reliable communication and operational efficiency. Traditional beamforming techniques, such as SDMA and NOMA schemes, which rely heavily on precise CSIT, often fail to meet performance expectations under such dynamic and complex conditions. Recognizing these limitations, given these limitations, rate-splitting multiple access (RSMA) has emerged as a more flexible and efficient approach for addressing the unique challenges posed by V2X communication systems [9, 10, 19, 20]. RSMA offers an adaptive framework for managing interference in highly dynamic V2X scenarios, where rapid changes in vehicle positions create challenges for maintaining accurate CSIT. By splitting each user's message into a common and a private part, RSMA allows the common part to be decoded and removed as interference, while treating the private parts of other users as manageable noise. This flexible message-splitting strategy enhances resilience to CSI variability, enabling more reliable and efficient interference management in V2X networks.

Despite the promising potential of RSMA in scenarios with imperfect CSIT, its implementation poses significant challenges. One of the primary difficulties lies in the design and optimization of beamforming strategies under conditions of partial or imperfect CSIT. One of the earliest and foundational studies in this area was presented by Mishra et al. [21], who formulated the precoder optimization problem for RSMA-enabled MIMO systems under imperfect CSIT. Following this, Krishnamoorthy et al. [22] introduced a novel approach based on null-space precoding, where linear combinations of the null-space basis vectors of the successively augmented MIMO channel matrices were employed as precoding vectors. Amor et al. [23] explored statistical precoding techniques for RSMA in frequency-division-duplex (FDD) MIMO systems. Similarly, Dizdar et al. [24] proposed a low-complexity precoding scheme specifically designed for RSMA in overloaded MIMO systems. Despite these significant advances, most of these studies have primarily focused on fixed users in relatively static environments, which overlook the inherent mobility of users. Addressing this issue, closed-form solutions have been proposed for optimal power allocation in RSMA-based systems. For example, in [25, 26], a closed-form solution was derived to maximize a lower bound on the system's performance. Zhu et al. [27] made further contributions by deriving a closed-form expression for the ergodic sum-rate in finite blocklength RSMA systems under imperfect CSIT. While these closed-form expressions are of great theoretical importance, their practical applicability to V2X scenarios remains limited due to their static nature and the simplifying assumptions made about the mobility and variability of the communication environment. In an effort to bridge the gap between theory and real-world V2X applications, Zhang et al. [5, 28, 29] introduced an RSMA-based V2X system designed specifically for interference management by leveraging incentive-driven methods to encourage vehicles to participate in interference management. However, while their method demonstrated potential in managing interference, the iterative optimization process required was computationally intensive and time-consuming. This limitation hinders its practical application in real-time V2X communications, especially given that traditional state-of-the-art methods often rely on assumptions that convert inherent randomness into deterministic scenarios. In contrast, deep learning (DL)-based approaches are better suited to address the stochastic nature of V2X environments, enabling more effective and timely solutions.

Recently, predictive beamforming, leveraging the power of DL, has emerged as a promising solution

for beamforming design in V2X [30–39]. This cutting-edge approach introduces an innovative design paradigm, wherein a single-step predictor is integrated into the beamforming pipeline, effectively simplifying the conventional two-step process: the estimation or prediction of CSI followed by precoder generation. Predictive beamforming eliminates the need for this separation by directly deriving beamforming parameters from historical or partial CSIT using a learned predictive model. This direct approach to beamforming offers substantial advantages. By embedding a single-step predictor, the system can learn and exploit potential non-linear relationships and complex temporal dependencies inherent in the CSI, which are crucial in dynamic V2X scenarios. This enables the model to better capture and adapt to the rapid variations and interdependencies in the wireless channel, ultimately improving beamforming accuracy and robustness in high-mobility environments. Several studies have demonstrated the efficacy of predictive beamforming in various communication environments. For instance, Yuan et al. [30] introduced a learning-based predictive beamforming framework specifically designed to address the problem of beam misalignment in unmanned aerial vehicle (UAV) communications. Building on this foundational work, Liu et al. [31] extended the concept of predictive beamforming to V2X networks, proposing a DL-based approach for directly predicting the beamforming matrix. This method demonstrated the significant potential of DL-based predictive models in improving the reliability and robustness of V2X communications. Furthermore, the applicability of this approach was later expanded to UAV communication scenarios by Liu et al. [32]. In addition to these contributions, Wang et al. [33] introduced a novel predictive beamforming solution relying on Transformer. Despite the impressive advancements in predictive beamforming, most of the existing approaches are limited in their ability to handle the complexities introduced by real-world V2X environments. The majority of predictive beamforming models primarily rely on historical CSIT data derived from simplified parameter models or idealized road conditions. This assumption limits their practical applicability in environments where road geometries introduce significant variations in channel conditions. Complex urban environments with varying road topologies can cause rapid and unpredictable changes in channel states, making it difficult for traditional predictive models to maintain accurate beamforming predictions [40].

Integrating traffic data from various sources enhances the decision-maker's understanding of the spatial-temporal dependencies present in historical CSIT. However, in real-world implementations, traffic data are inherently complex, originating from a variety of sources, including Internet of Things (IoT) devices and camera-based monitoring systems. This diversity leads to different interpretations and features that describe the spatial dependencies among CVs. Effectively harnessing this multimodal data requires sophisticated analytical approaches. One promising solution is the large multimodal model (LMM), which exhibits powerful reasoning capabilities for understanding the intricate relationships within large volumes of multimodal data. Drawing inspiration from large language models (LLMs), LMMs are specifically designed to process and integrate information from diverse modalities, thereby distinguishing themselves from traditional LLMs that primarily focus on sequence-to-sequence language inputs. Numerous applications have been developed based on the conventional LLM framework, including robotics [41], autonomous driving [42], and smart cities [43]. Moreover, the utility of LLMs has been widely recognized in the field of wireless communication, where they have been successfully employed to enhance system performance through optimized resource allocation [44] and improved signal processing [45].

The superiority of LLM-aided predictive beamforming has been demonstrated in [46, 47]. However, it primarily considers a fixed precoder with single-modal input, which limits its applicability in V2X scenarios characterized by complex traffic and communication data. In contrast, the integration of multimodal data within LMMs leverages the foundational principles of LLMs while addressing the complexities of intersecting traffic and communication data. By incorporating multiple types of input, LMMs enable a deeper understanding and analysis of the intricate interactions between various traffic data sources, leading to more informed and effective decision-making in V2X communication systems.

In this study, we propose a novel multimodal data processing framework that employs an LMM to seamlessly integrate road geometries, semantic road information, and historical CSIT for optimal beamforming design in RSMA-enabled V2X systems. This multimodal approach allows the model to better capture the spatial and temporal dependencies inherent in dynamic V2X environments, resulting in beamforming solutions that are not only more accurate but also more adaptive to real-time changes. By combining the strengths of LLMs with multimodal data processing, our work significantly enhances the beamforming design process, providing robust solutions that can scale to the complexities of modern V2X systems. Our main contributions can be summarized as follows.

- We conduct an in-depth analysis of RSMA performance in V2X communication systems, particularly

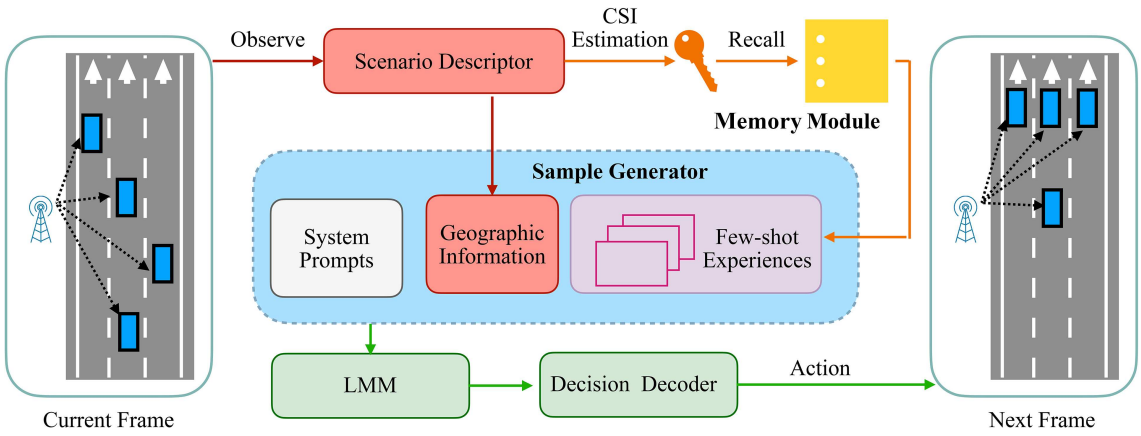

Figure 1 (Color online) Proposed LMM framework in V2X.

in environments characterized by complex road geometries. Specifically, we formulate the maximization of the minimum fairness (MMF) problem for beamforming design in RSMA-enabled V2X communications. We assess RSMA performance through its minimum ergodic user rate under complex road geometries, demonstrating that RSMA offers significantly higher reliability compared to conventional SDMA.

- Given the large-scale and multimodal nature of traffic data and communication data, which includes diverse and interwoven information from different systems, we design an LMM framework for V2X. Unlike existing approaches that only analyze historical CSIT data, our framework integrates and leverages complex, cross-system data, enhancing the adaptability and precision of the predictive beamforming model.
- We propose an innovative LMM, referred to as GTR. The GTR model employs a graph convolutional network (GCN) to effectively capture the complex topology of road networks, enabling a comprehensive understanding of CSIT in the presence of intricate road geometries. To further enhance its analysis, GTR leverages a Transformer encoder that processes these complex features, facilitating deeper insight into the temporal and spatial dependencies within the V2X environment. Subsequently, a Transformer decoder module refines this analysis, while road semantic information enriches the model's contextual comprehension of the surrounding environment. Based on these multi-modal inputs, GTR ultimately generates highly accurate beamforming designs for RSMA in V2X scenarios, significantly improving the robustness of the beamforming process.
- Simulation results demonstrate that incorporating multi-modal inputs, including geometrically related traffic information, helps the LMM better understand the temporal-spatial dependencies inherent in complex and information-rich CSIT data. This leads to more robust and effective beamforming designs in V2X communication systems, while maintaining a milliseconds level processing time.

The rest of the paper is organized as follows. Section 2 provides a detailed overview of the proposed LMM framework, while Section 3 elaborates on the GTR, used in LMM. The performance evaluation and analysis of LMM are presented in Section 4. Finally, Section 5 summarizes our key findings and conclusion.

2 LLM framework for V2X

Drawing inspiration from the knowledge-driven paradigm that underpins 6G communications [48], we propose an innovative multi-modal learning framework designed to enhance V2X communications by effectively integrating road geometry and semantic information. This LMM framework comprises four essential modules: descriptor, memory module, sample generator, and decision decoder, as illustrated in Figure 1. The process begins with the descriptor, which observes and analyzes the current road geometry, generating detailed descriptions while processing feedback from CVs. This information is then utilized by the sample generator, which synthesizes the scenario description with few-shot experiences retrieved from the memory module. These enriched samples are subsequently input into a robust LMM. Finally, the decision decoder interprets the model's outputs to determine optimal actions. A comprehensive discussion of the implementation details for the descriptor, memory module, sample generator, and decision decoder will be provided in the following subsections.

2.1 Descriptor

To facilitate the LMM in understanding the interference channel, the descriptor module is designed to encode various scenarios as multi-modal inputs. This module plays a critical role in collecting and synthesizing a comprehensive description of the current driving environment from multiple perspectives. Specifically, it gathers data on road geometry and CSI feedback obtained from CVs. The road geometry encompasses vital parameters such as the direction of the road, lane width, and information regarding intersections or crossroads. These factors are essential for interpreting the dynamics of the interference channel, as they provide insights that significantly enhance the understanding of CSI fluctuations. By combining road geometry with the evolving communication conditions, the descriptor improves the predictive capabilities of the system, allowing for more accurate assessments of how environmental changes may affect signal propagation and reception.

Simultaneously, the CSI feedback serves to update the estimated CSI during the most recent communication round. In this framework, we operate under the assumption that the CSI remains relatively constant over short durations, enabling a stable basis for decision-making processes. The descriptions generated by the descriptor module consist of both static road attributes and dynamic CSI information. All these data points are projected into a learnable data matrix, facilitating further processing. Furthermore, the CSI data matrix acts as a key to retrieve relevant historical experiences stored in the memory module. By creating this comprehensive and structured representation of the driving environment, the descriptor enables LMM to make informed and contextually aware decisions, ultimately enhancing the performance and reliability of V2X communications.

2.2 Memory module

The memory module is designed to store and organize the rich historical data, including scene descriptions that capture the essential characteristics of previous communication situations. These scene descriptions, which compress both static and dynamic elements of the road and the vehicular environment, act as unique keys to the memory module. When faced with a new scenario, the current scene description is matched with relevant past experiences, allowing the system to retrieve pertinent memories. This retrieval process provides the LMM with essential knowledge, improving its ability to make informed decisions regarding beamforming strategies. Moreover, the current scene descriptions are also stored in the memory module for future use, contributing to a continuously growing knowledge base that reflects a wide range of communication conditions. This ongoing accumulation of data enables the system to adapt to various road types, traffic patterns, and CV behaviors over time, thereby improving its decision-making accuracy and flexibility.

2.3 Sample generator

As illustrated in the blue dashed box in Figure 1, each sample generated for input into the large AI model consists of three key components: system prompts, geographic information obtained directly from the scenario descriptor, and historical CSIT retrieved or synthesized from the memory module. These components jointly form a comprehensive input framework that enhances the decision-making capabilities of the AI model. The system prompts play a crucial role in providing a contextual overview of the road geometry observed in the current scenario. These prompts are derived from semantic information related to road names and other contextual indicators, which convey important details about the environment in which the CVs operate. By embedding this semantic context into the system prompts, the sample generator enables the AI model to understand the channel conditions better and the historical behavior patterns of CVs on that particular road segment.

In each decision frame, the sample generator constructs samples based on the real-time driving scenario, taking into account not only the immediate environmental factors but also the historical context provided by previous experiences stored in the memory module. This approach ensures that the samples are dynamically generated to reflect the current state of the driving environment while incorporating relevant information from past scenarios. The synthesis of these inputs allows the sample generator to create a comprehensive representation of the driving context, which is essential for making informed decisions.

The synthesized samples, including the historical CSIT samples, road map, and the sematic information (i.e., traffic condition and road name), are subsequently input into the LMM. The model then engages in a sophisticated analysis of the sample, utilizing its learned representations to determine the most

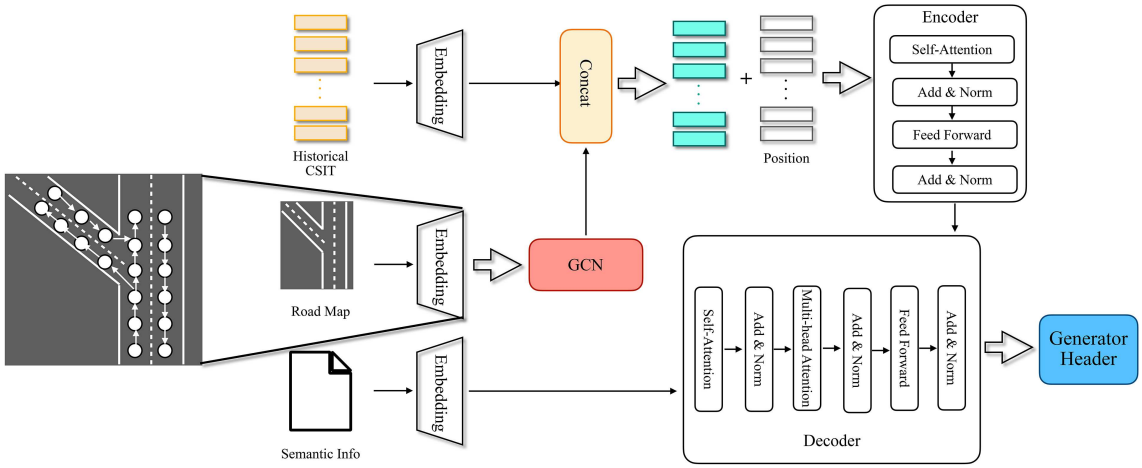

Figure 2 (Color online) Structure of the proposed GTR.

appropriate actions for the current frame. This decision-making process is informed by the interplay between the system prompts, geographic information, and historical CSIT, enabling the AI model to consider both immediate and contextual factors when formulating its output.

2.4 Decision decoder

Upon feeding the generated sample into the LMM, the subsequent step involves decoding the final decision through the action decoder. This crucial component translates the outcomes produced by the AI model into actionable strategies for V2X communications. The action decoder serves as the interface between the AI's predictive capabilities and the real-world operational requirements, ensuring that the decisions made by the AI model are effectively implemented in a manner that optimally supports V2X interactions.

The system continuously refines its decisions by repeating the aforementioned procedures, thereby establishing a dynamic closed-loop decision-making system. This feedback mechanism enables the action decoder to learn from previous outcomes and adapt future actions, ultimately driving improvements in performance and responsiveness in real-time applications.

3 Large multi-modal model

3.1 Work flow

In this section, we present the design and architecture of the proposed Transformer-like LMM, named GTR. GTR is designed to handle the multi-modal inputs from the sample generator descriptor efficiently and effectively. The primary objective of GTR is to facilitate the integration of complex road geometries, historical channel state information, and semantic data, ensuring optimal performance in V2X communication systems. As illustrated in Figure 2, the GTR model is composed of four key modules: the GCN, the encoder module, the decoder module, and the generator header. Each of these modules plays a critical role in ensuring the model's ability to process multi-modal data and make informed decisions for beamforming.

Upon receiving the data samples generated by the sample generator, GTR first embeds the original road map into a structured lane graph. This graph represents the road geometry, including key features such as lane width and intersection points, which are critical for understanding the spatial context of the communication environment. The lane graph is then processed by the GCN module, which is specifically designed to extract complex spatial relationships between various road elements. The GCN effectively transforms the raw geometric data into rich spatial features, enabling the model to capture intricate details of the road geometry that would be otherwise missed by traditional approaches. Simultaneously, the historical CSIT is embedded into a fixed-dimensional feature space, ensuring consistency and compatibility for further processing. This is particularly important as it allows the model to handle diverse data modalities within a unified framework. Once the road geometry and historical CSIT features are

obtained, they are concatenated into a unified time-series sequence. To account for the temporal dependencies within this data, positional encoding is applied, embedding time-step information directly into the sequence. This ensures that the temporal aspect of the vehicular communication dynamics is not lost during the feature extraction process.

The concatenated sequence is then fed into the Transformer encoder, which is responsible for capturing both spatial and temporal dependencies within the input data. The self-attention mechanism, a unique feature of the Transformer architecture, allows the model to selectively focus on the most relevant parts of the input sequence, enhancing its ability to discern critical patterns and relationships. This enables GTR to effectively handle the dynamic and complex nature of V2X scenarios, where both spatial and temporal features play crucial roles in determining the quality of beamforming decisions. Next, the features extracted by the Transformer encoder are passed to the Transformer decoder, which utilizes road semantic information as an additional input to guide the generation of decisions. This incorporation of semantic data further enhances the model's ability to capture the intricate relationships between road geometry and communication needs, ultimately improving the beamforming decisions. The decoder's ability to leverage both spatial-temporal data and semantic information ensures that the model produces highly context-aware actions.

Finally, the generator header utilizes the outputs of the Transformer decoder to generate control actions for the beamforming process. It outputs the optimal precoder design and common rate allocation parameters necessary for RSMA-enabled V2X communication. These control actions are designed to leverage the comprehensive information extracted throughout the pipeline, ensuring that the beamforming strategy is optimized for both spatial resolution and communication robustness.

In a nutshell, the integration of the RSMA scheme within the LMM framework is designed to enhance beamforming decisions by leveraging the reasoning capabilities of the LMM. Specifically, the LMM processes multi-modal inputs including historical CSIT, road topology, and traffic conditions to generate beamforming strategies that are optimized for future system states. Within this process, RSMA is incorporated by structuring the beamformer outputs to allocate power and precoding weights in a manner that dynamically adapts to user channel conditions and interference levels. This integration enables a more flexible and adaptive multiple access strategy that improves spectral efficiency and user fairness, allowing the system to better handle the fluctuating channel conditions in a V2X environment.

3.2 Structure of the GTR model

3.2.1 *GCN module*

In our specific application, we model the road infrastructure through a lane graph and process it by GCN, which serves as a robust mechanism for capturing spatial dependencies within structured data, particularly when this data can be effectively represented as a graph. As depicted in Figure 2, the lane segments are designated as a set of nodes, denoted by $\boldsymbol{V}$, while the connections or transitions between these lane segments are represented as edges, denoted by $\boldsymbol{E}$. This graph-based approach distinguishes GCNs from traditional convolutional neural networks (CNNs), which primarily operate on grid-like structures such as images. GCNs excel at extracting complex spatial relationships that exist between various lanes, intersections, and the dynamics of CVs within the lane graph. This capability significantly enhances predictive beamforming performance in the context of V2X communications.

The lane graph is formally represented as $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$, where each node $v_i \in \boldsymbol{V}$ corresponds to a specific lane segment of the centerline. The position of each lane node is computed as the average coordinates of its two endpoint locations. By analyzing the connections between lane centerlines, we categorize the lane nodes based on four common connectivity types: predecessor, successor, left neighbor, and right neighbor. Specifically, a lane node's predecessor and successor are defined as neighboring lane nodes that can either travel to or from the given lane node, respectively. It is important to note that, particularly in complex scenarios such as crossroads, a lane node may have multiple predecessor or successor nodes. The left and right neighbors are defined as the closest lane nodes situated to the left and right of the current lane, respectively.

We represent the connectivity of the lane graph using the adjacency matrix $\boldsymbol{A}$. The individual elements of this matrix are defined as $\boldsymbol{A}_{jk}$, representing the relationship between node j and node k. Specifically, we define $\boldsymbol{A}_{jk} = 1$ if node k is a neighbor of node j, indicating a direct connection between the two nodes. The operation performed by a single GCN layer on the lane graph $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$ is mathematically

expressed as

$$\boldsymbol{X}^{(l+1)} = \sigma\left(\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{X}^{(l)}\boldsymbol{W}^{(l)}\right), \tag{1}$$

where $\boldsymbol{X}^{(l)}$ denotes the matrix of node features at layer l, $\boldsymbol{D}$ is the degree matrix derived from $\boldsymbol{A}$, $\boldsymbol{W}^{(l)}$ signifies the learnable weight matrix for layer l, and σ is a nonlinear activation function. By stacking multiple GCN layers, the model acquires progressively abstract representations of the lane segments, which can subsequently enhance the encoder's understanding of road geometry.

3.2.2 *Encoder module*

The encoder in our architecture is designed to process and extract high-dimension feature representations from multimodal inputs. Following the concatenation of the output from the GCN with the features derived from historical CSIT, the encoder effectively embeds this unified sequence, defined as $\boldsymbol{X}_1$, incorporating positional information that enriches the contextual understanding of the data. Subsequently, the embedded sequence $\boldsymbol{X}_1$ is processed through a Transformer encoder, which is renowned for its capability to model complex dependencies within sequential data. The Transformer encoder capitalizes on a unique self-attention mechanism, which facilitates the model's ability to focus on different parts of the input sequence, adjusting its attention based on the relevance of each component. This self-attention mechanism is mathematically represented as follows:

$$\boldsymbol{X}_2 = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{\iota_1}}\right)\boldsymbol{V}, \tag{2}$$

where $\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_Q$ represents the query matrix, $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_K$ denotes the key matrix, and $\boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_V$ signifies the value matrix, all derived from the input data. The matrices $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, and $\boldsymbol{W}_V$ are learnable linear transformations that facilitate the conversion of the input data into these respective matrices. The application of the softmax function ensures that the attention weights are normalized, allowing the model to assign varying levels of importance to different features present in the input sequence. The utilization of the self-attention mechanism ensures that both the CSI and environmental features are emphasized, enabling the model to adapt its decisions based on the most relevant characteristics for V2X. This adaptability is crucial for effective communication and interaction in dynamic environments.

Following the self-attention operation, the output is subjected to an add & norm layer, which plays a critical role in stabilizing the training process. The normalization process is computed as follows:

$$\boldsymbol{X}_3 = \text{Norm}(\boldsymbol{X}_2 + \boldsymbol{X}_1). \tag{3}$$

This equation indicates that the output from the self-attention layer $\boldsymbol{X}_2$ is combined with the original embedded input sequence $\boldsymbol{X}_1$ and then normalized. This addition helps preserve the information from the original input while stabilizing the training.

In the subsequent stage, the model employs two fully connected layers to construct the feed-forward layer, which serves to further process the normalized output. This layer is represented mathematically as

$$\boldsymbol{X}_4 = \max(\boldsymbol{A}_1\boldsymbol{X}_3 + \boldsymbol{b}_1, 0)\boldsymbol{A}_2 + \boldsymbol{b}_2, \tag{4}$$

where $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ represent the weight matrices for the two fully connected layers, while $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are the corresponding bias vectors. The application of the ReLU activation function, represented by $\max(\cdot, 0)$, introduces non-linearity into the model, allowing it to capture more complex patterns within the data. This structured approach ensures that the encoder effectively enhances the understanding of the relationships between the features, ultimately improving the performance of the V2X communication system.

3.2.3 *Decoder module*

In the decoder component of our proposed model, the primary input consists of the semantic road information, which enables the model to adjust its beamforming predictions according to contextual changes in the environment. In particular, the semantic information is input to the decoder because it serves as the conditional context that directly influences the final decision-making process. In contrast, the road map and CSIT data provide raw information that captures the spatial dependencies of the

environment. By feeding the semantic information into the decoder, we enable the model to better adjust its predictions based on dynamic, real-time conditions. Each layer of the decoder is designed to enhance the model's capability to comprehend and utilize this rich semantic information effectively. At each decoding layer, we employ a multi-head attention mechanism, which serves to integrate the output generated by the encoder with the accompanying semantic road information. This multi-head attention mechanism is particularly important as it enables the model to focus on multiple segments of the input data simultaneously. By doing so, it effectively captures a wide range of contextual relationships and interactions that exist within the road network. This attention to detail facilitates the generation of outputs that are sensitive to the complex relationships inherent in the data, such as lane changes, vehicle dynamics, and environmental conditions. The architecture of the decoder ensures that the model not only synthesizes information from the encoder but also leverages the rich semantic features of the road network. This dual-input approach is instrumental in enhancing the decoder's capacity to produce high-quality decisions, making it particularly effective for applications in V2X communications.

3.2.4 *Generator header*

Finally, the output generated from the decoder layer is projected by the generator header into functional actions, specifically the beamforming parameters essential for effective communication strategies. In the context of RSMA, this includes critical components such as common rate allocation $\tilde{\boldsymbol{C}}$ and precoder design $\tilde{\boldsymbol{P}}$. The generation module plays a pivotal role in ensuring that the GTR remains flexible and responsive to the real-time demands of the network. This responsiveness is crucial for adapting to dynamic environmental conditions, varying traffic patterns, and fluctuating communication requirements. By translating the high-level outputs from the decoder into actionable parameters, the generator header facilitates a seamless transition from data processing to practical application, thereby enhancing the overall performance of the V2X communication system.

3.3 Loss function design

We consider the MMF problem for RSMA-enabled V2X communications. Specifically, we consider a BS equipped with N_t antennas, tasked with serving K single-antenna CVs. The transmitted signal at any given time slot t can be expressed as

$$\boldsymbol{x}(t) = \boldsymbol{P}(t)\boldsymbol{s}(t) = \boldsymbol{p}_c(t)s_c(t) + \sum_{k\in\mathcal{K}} \boldsymbol{p}_k(t)s_k(t), \tag{5}$$

where $\boldsymbol{P}(t) \triangleq \{\boldsymbol{p}_c(t), \boldsymbol{p}_1(t), \ldots, \boldsymbol{p}_K(t)\} \in \mathbb{C}^{(K+1)\times N_t}$ represents the precoder matrix designed to facilitate the transmission of the data streams $\boldsymbol{s}(t) \triangleq \{s_c(t), s_1(t), \ldots, s_K(t)\}$. At the CV side, the received signal at CV k can be formulated as

$$y_k(t) = \boldsymbol{h}_k(t)\boldsymbol{x}(t) + n_k, \tag{6}$$

where $n_k \sim \mathcal{CN}(0, \sigma_k^2)$ represents the additive white Gaussian noise (AWGN) component for CV k, and $\boldsymbol{h}_k(t) \in \mathbb{C}^{N_t\times 1}$ is the channel coefficient between the BS and CV k at time slot t. In this context, the instantaneous signal-to-interference-plus-noise ratios (SINRs) for the common and private streams at time slot t can be expressed as follows:

$$\gamma_k^c(t) = \frac{|\boldsymbol{h}_k^{\mathrm{H}}(t)\boldsymbol{p}_c(t)|^2}{\sum_{i\in\mathcal{K}} |\boldsymbol{h}_k^{\mathrm{H}}(t)\boldsymbol{p}_i(t)|^2 + \sigma_k^2}, \tag{7}$$

and

$$\gamma_k^p(t) = \frac{|\boldsymbol{h}_k^{\mathrm{H}}(t)\boldsymbol{p}_k(t)|^2}{\sum_{i\in\mathcal{K}\setminus k} |\boldsymbol{h}_k^{\mathrm{H}}(t)\boldsymbol{p}_i(t)|^2 + \sigma_k^2}. \tag{8}$$

The instantaneous achievable rates for the common and private streams can thus be expressed as

$$R_k^c(t) = B\log_2(1+\gamma_k^c(t)), \tag{9}$$

$$R_k^p(t) = B\log_2(1+\gamma_k^p(t)), \tag{10}$$

where B denotes the downlink bandwidth. To ensure that the common message is accurately decoded by each CV, we impose the following constraint:

$$R^c(t) = \min_{k\in\mathcal{K}}\{R_k^c(t)\}. \tag{11}$$

Since all CVs share the same common message, we define the overall common rate as $R^c(t) = \sum_{k\in\mathcal{K}} C_k(t)$, where $\boldsymbol{C}(t) = \{C_k(t)|k \in \mathcal{K}\}$ represents the portions of the common rate. Consequently, the total instantaneous rate for transmitting the message to CV k is given by

$$R_k^{\text{tot}}(t) = C_k(t) + R_k^p(t). \tag{12}$$

Given the channel uncertainty in V2X, we optimize the overall user experience by maximizing the minimum achievable ergodic rate among all connected vehicles, which can be expressed as follows:

$$\begin{aligned}
\text{P0}: \quad & \max_{\hat{\boldsymbol{C}}(t),\boldsymbol{P}(t)} \quad \min_{k\in\mathcal{K}} \hat{R}_k^{\text{tot}} && \text{(13a)}\\
& \text{s.t.} \quad \hat{R}_k^p \geqslant R_{\text{th}}^p, \forall k \in \mathcal{K}, && \text{(13b)}\\
& \qquad \text{tr}(\boldsymbol{P}^{\text{H}}(t)\boldsymbol{P}(t)) \leqslant P_t, && \text{(13c)}\\
& \qquad \boldsymbol{C}(t) \geqslant 0, && \text{(13d)}
\end{aligned}$$

where $\hat{R}$ is the ergodic user rates. Based on the output generated by the predictor, we design the loss function for the MMF optimization problem defined as P0. The first step in this process involves transforming the predictor's output into meaningful beamforming parameters that satisfy the established quality of service (QoS) constraints defined in P0. Specifically, we reshape the precoder design, defined as $\tilde{\boldsymbol{P}} \in \mathbb{R}^{K\times 2N_t}$, into a new format $\tilde{\boldsymbol{P}} \in \mathbb{C}^{K\times N_t}$ by effectively separating and subsequently combining its real and imaginary components. Next, to ensure the power budget, we verify whether the total transmission power exceeds the available power budget P_t. If the calculated transmission power does exceed this budget, we normalize the power levels to match P_t. Conversely, if the calculated power is within the permissible limits, we retain the computed power levels. This procedure can be mathematically represented as

$$\boldsymbol{P} = \begin{cases} \tilde{\boldsymbol{P}}, & \text{tr}(\boldsymbol{P}^{\text{H}}\boldsymbol{P}) \leqslant P_t, \\ \tilde{\boldsymbol{P}}\sqrt{\dfrac{P_t}{|\text{tr}(\tilde{\boldsymbol{P}}^{\text{H}}\tilde{\boldsymbol{P}})|}}, & \text{tr}(\boldsymbol{P}^{\text{H}}\boldsymbol{P}) > P_t. \end{cases} \tag{14}$$

This normalization process guarantees that the power constraint expressed in (13c) is always satisfied.

Following the successful determination of the precoder design, we proceed to calculate the transmission rates, defined as $\tilde{R}_k^p$ and $\tilde{R}_k^c$. Given the constraint defined in (13b), we establish the following condition for the transmission rate:

$$R_k^p = \begin{cases} \tilde{R}_k^p, & \tilde{R}_k^p \geqslant R_{\text{th}}^p, \\ 0, & \tilde{R}_k^p < R_{\text{th}}^p. \end{cases} \tag{15}$$

This formulation effectively captures the occurrence of outage events, wherein a user terminal is unable to successfully decode the transmitted signal due to insufficient power or other constraints.

Then, we normalize the common rate allocation by

$$\boldsymbol{C} = \frac{|\tilde{\boldsymbol{C}}| \min_{k\in\mathcal{K}}\{\bar{R}_k^c\}}{\sum_{k\in\mathcal{K}} |\tilde{C}_k|}. \tag{16}$$

This normalization ensures that the common rate for each user terminal is appropriately represented, while also maintaining the condition that $\sum_{k\in\mathcal{K}} \boldsymbol{C} \leqslant \bar{R}_k^c$.

To approximate the ergodic rates during the training phase, we implement a Monte Carlo method as described in [49], which leverages the training set to yield an estimated total rate

$$\hat{R}_k^{\text{tot}} = \frac{1}{\Psi}\sum_{i=0}^{\Psi} \left(R_k^p + C_k\right). \tag{17}$$

By substituting the equation from (17) into (13a), the loss function can be expressed as

$$\mathcal{L} = -\min_{k\in\mathcal{K}} \hat{R}_k^{\text{tot}}. \tag{18}$$

This loss function is carefully designed to minimize the negative value of the minimum ergodic user rate, which directly aligns with the objective of MMF rates, as formulated in P0. By optimizing the MMF ergodic rates based on historical CSIT, our approach effectively enhances the quality of beamforming design while proactively adapting to channel variations. This, in turn, maximizes the expected future system throughput in a dynamic V2X environment, ensuring reliable and efficient communication for all users.

4 Numerical results

4.1 Simulation setup

The simulation framework is designed to reflect real-world V2X communication environments, incorporating complex road geometry and dynamic vehicular movement. The underlying road map is randomly generated using Google Maps[1)], capturing a diverse range of urban layouts, including intersections, straight road segments, and multilane configurations. The data samples are synthesized according to the well-known next generation simulation (NGSIM) dataset [50] and the channel model presented in [17]. This dataset includes real-world vehicle trajectories, ensuring that the data align with real-world performance in a V2X environment. The lane graph is constructed as described in Subsection 3.2, where each segment of the road is represented by nodes and edges to model the intricate road geometry. These samples provide the necessary temporal information for training the predictive beamforming model. Following [51], the transmission bandwidth B is set to 5 MHz. To simplify the analysis and focus on the beamforming performance, we normalize the channel noise variance to 1, allowing us to define the transmission power, P_t, directly in terms of signal-to-noise ratio (SNR). Unless otherwise specified, the baseline transmission power is set to $P_t = 30$ dB, reflecting a common-SNR regime suitable for reliable V2X communication. The BS is equipped with $N_t = 4$ antennas, and it serves $N = 4$ uniformly distributed CVs within its coverage area. The historical CSIT duration is set to $T_0 = 1$ s, and the CSIT is sampled at a frequency of 10 Hz, which provides sufficient temporal granularity for learning the underlying spatial-temporal dynamics of the V2X channels. The CVs are assumed to move at an average velocity of $v = 40$ km/h, which is representative of typical urban driving conditions. Moreover, semantic information about the road, such as the road name and lane configurations, is incorporated as an input to the decoder module within GTR.

To ensure a robust and unbiased evaluation of the LMM framework, the dataset is randomly split into three distinct sets: training, validation, and testing. The dataset is partitioned in a (70% : 10% : 20%) ratio, where 70% of the data is used for training the model, 10% is reserved for validation, and the remaining 20% is set aside for testing. This division ensures that the model is not overfitted to the training data and that its generalization capabilities can be rigorously tested on unseen data. The learning process is optimized using the Adam optimizer, a widely used optimization algorithm known for its computational efficiency and fast convergence properties. The learning rate is set to 0.001. The model is trained with a batch size of 512 samples. This relatively large batch size allows for stable updates to the model parameters while maintaining computational efficiency. The training process is carried out for 1000 iterations to ensure that the model converges to an optimal solution. Furthermore, the hidden size of the feature representations is set to 512. This choice of hidden size reflects a trade-off between model complexity and computational cost, ensuring that the model can effectively capture the rich spatial-temporal dependencies in the V2X channels while remaining computationally feasible for real-time deployment.

4.2 Baseline algorithms

In this simulation framework, we conduct a comprehensive evaluation of the proposed LMM by comparing its performance against two well-established benchmarks: (1) the conventional weighted minimum mean squared error (WMMSE) algorithm [11], and (2) a singlemodal Transformer-based approach that does

1) Google Maps. 2024. Accessed: 2024-10-11.

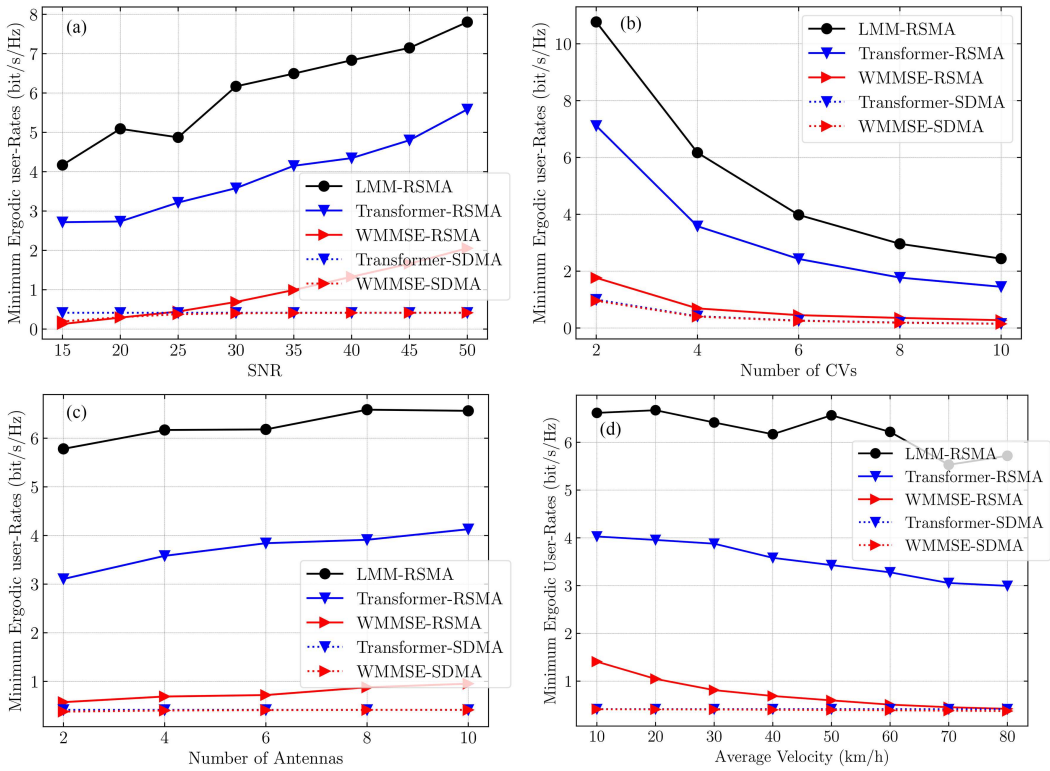

Figure 3 (Color online) MMFR performance of LMM. (a) MMFR vs. SNR; (b) MMFR vs. number of CVs; (c) MMFR vs. number of antennas; (d) MMFR vs. average velocity.

not incorporate road geometry or semantic information. Furthermore, to thoroughly evaluate the efficacy of RSMA, we benchmark its performance against the conventional SDMA approach. The evaluation framework, therefore, encompasses a total of five distinct beamforming schemes, allowing for a holistic examination of the interaction between access methods (RSMA vs. SDMA) and the predictive models employed (WMMSE, Transformer, and LMM).

For the RSMA-based approaches, we consider the following key variations.

- **WMMSE-RSMA:** The traditional WMMSE algorithm applied to the RSMA framework.
- **Transformer-RSMA:** A learning-based RSMA scheme utilizing a Transformer model trained solely on historical CSIT, without the inclusion of road geometry or semantic data.
- **LMM-RSMA:** The proposed LMM framework applied to RSMA, which integrates multi-modal inputs such as road geometry and traffic-related semantic information into the beamforming decision-making process, as described in Section 2.

In parallel, for the conventional SDMA approach, we evaluate the following.

- **WMMSE-SDMA:** The conventional WMMSE algorithm applied to SDMA, serving as a baseline for comparison against the more advanced RSMA strategies.
- **Transformer-SDMA:** A learning-based SDMA scheme utilizing a Transformer model trained solely on historical CSIT, similar to the Transformer-RSMA approach but applied within the SDMA scheme.

4.3 Performance analysis

In Figure 3(a), we present a comprehensive analysis of the MMF user rate (MMFR) performance of LMM under varying SNR in V2X. The results unequivocally demonstrate a consistent enhancement in MMFR as the SNR increases. A notable trend in the analysis is the superior performance of the RSMA scheme compared to the conventional SDMA scheme, observed across both the traditional WMMSE approach and the advanced predictive beamforming methods based on Transformer models. This consistent outperformance of RSMA can be attributed to its inherently robust design, which offers a more flexible and effective strategy for managing interference, particularly under the imperfect CSIT in V2X. In particular, complex road geometry introduces unpredictable channel variations, making it increasingly difficult to

maintain accurate CSI predictions over time. This phenomenon is particularly detrimental to CSIT-sensitive techniques like SDMA, which rely heavily on precise channel information for efficient spatial separation and interference management. Therefore, the SDMA's performance degrades significantly, as it is unable to cope with the rapidly changing channel conditions. This limitation results in a persistently low MMFR performance for SDMA across all SNR levels. On the other hand, RSMA demonstrates resilience in the face of such challenges, owing to its unique approach to interference management. RSMA treats part of the signal from other users as interference, while simultaneously considering a portion of it as noise. This hybrid approach allows RSMA to effectively manage interference even under imperfect CSIT conditions, particularly when faced with complex road geometries that induce significant channel variability. By balancing between interference treatment and noise recognition, RSMA can dynamically adjust to the communication environment, resulting in more reliable and efficient beamforming decisions.

Of particular interest in the results is the remarkable performance of LMM, which not only surpasses the MMFR achieved by the conventional WMMSE-based RSMA approach but also exhibits a substantial improvement over the advanced predictive beamforming technique using a singlemodal Transformer model. Specifically, LMM delivers over a 400% increase in MMFR compared to the conventional WMMSE scheme and a 50% improvement relative to the singlemodal Transformer-based beamforming method. These results highlight the effectiveness of LMM's multi-modal learning framework. The substantial gains in MMFR observed with LMM can be attributed to its ability to capture and process multi-modal inputs, which significantly enhances its capacity to predict and adapt to channel variations. By incorporating information about road geometry, LMM can anticipate how the physical layout of the road will influence CSI over time, allowing it to make more informed decisions regarding beamforming and resource allocation. Furthermore, LMM's integration of semantic information, further augments its predictive capabilities, enabling it to optimize beamforming strategies in a way that accounts for both static and dynamic aspects of the V2X environment.

4.3.1 *Robustness of LMM*

Then, we assess the robustness of the proposed predictive beamforming framework, leveraging the LMM, under various V2X communication scenarios. The robustness of LMM is evaluated against several critical parameters, including the number of CVs, antenna configurations, and average vehicle velocity. The results, presented in Figures 3(b)–(d), highlight the consistent superiority of LMM over conventional approaches and alternative learning-based methods.

As demonstrated in Figure 3(b), LMM demonstrates remarkable robustness and performance across a broad range of CV counts, consistently outperforming both the conventional WMMSE algorithm and singlemodal learning techniques. This enhanced performance stems from LMM's unique ability to integrate road geometry information with semantic data, enabling a more comprehensive understanding of the communication environment. The multi-modal nature of LMM enables it with the capacity to capture and process spatial-temporal dependencies inherent in vehicular networks, allowing it to manage interference and optimize beamforming design more effectively than its counterparts. In scenarios involving a higher number of CVs, the traditional SDMA scheme and the RSMA method with WMMSE experience a sharp decline in performance. This can be attributed to the escalating complexity of the underlying optimization problem as the user population grows. WMMSE, being an iterative algorithm, struggles to efficiently optimize beamforming under such circumstances. Consequently, the performance of WMMSE saturates and fails to scale effectively, leading to suboptimal interference management. In contrast, LMM, with its predictive capabilities and ability to leverage both current and historical data, exhibits robust performance even as the network density increases, reaffirming its effectiveness in complex, dynamic V2X environments.

In Figure 3(c), we turn our attention to the impact of antenna configurations. The results illustrate a consistent improvement in the MMFR as the number of antennas increases, highlighting the crucial role that spatial resolution plays in enhancing communication performance. The use of more antenna elements enhances the spatial separation between signals, allowing for more effective interference mitigation and improved beamforming accuracy. LMM utilizes on this spatial diversity, achieving superior MMFR values across all antenna settings when compared to conventional WMMSE and singlemodal learning-based methods. LMM's robust performance across different antenna settings highlights its flexibility and adaptability, making it an ideal solution for V2X scenarios that involve diverse physical and communication infrastructures.

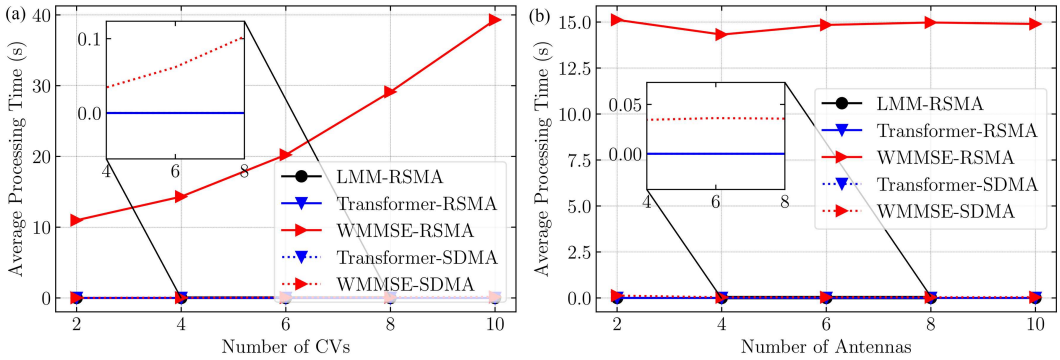

Figure 4 (Color online) Average processing time of LMM. (a) Average processing time vs. number of CVs; (b) average processing time vs. number of antennas.

Furthermore, Figure 3(d) examines the robustness of LMM under varying average vehicle velocities. In high-mobility V2X environments, where vehicle speeds can fluctuate considerably, maintaining reliable communication becomes increasingly challenging due to the rapid changes in CSI. As vehicle velocity increases, the predictability of channel conditions decreases. Conventional approaches, such as WMMSE, which rely solely on historical CSI feedback, struggle to adapt to these rapidly changing conditions, resulting in significant performance degradation. Similarly, singlemodal learning-based methods that lack the ability to incorporate contextual information, such as road geometry, face limitations in maintaining high prediction accuracy. Despite these challenges, LMM continues to outperform its counterparts, demonstrating remarkable robustness in high-velocity scenarios. The multi-modal learning framework employed by LMM allows it to predict channel variations more effectively by integrating not only historical CSI but also road geometry and semantic information. This additional contextual understanding enables LMM to make more informed decisions regarding beamforming design, leading to improved communication reliability and performance in fast-changing environments. The results shown in Figure 3(d) confirm that LMM achieves consistently higher MMFR values than both WMMSE and singlemodal learning approaches, even as vehicle speeds increase. This robustness under diverse traffic conditions highlights LMM's suitability for deployment in a wide range of V2X use cases, including high-speed highways and urban traffic scenarios with frequent velocity changes.

4.3.2 *Efficacy of LMM*

Subsequently, we present a comprehensive analysis of the efficacy of our proposed predictive beamforming framework, LMM, in comparison to traditional approaches. The results, as illustrated in Figures 4(a) and (b), provide critical insights into the practical viability of LMM for real-time V2X communication systems, particularly in terms of scalability and efficiency. The analysis focuses on two key factors that significantly influence the processing time: (i) the number of CVs involved in the communication system, and (ii) the configuration and complexity of the antenna array at the transmitter. These two parameters are crucial in determining the computational demands of the beamforming design process.

Figure 4(a) specifically investigates the average processing time required for beamforming under different numbers of CVs. As expected, the processing time for the conventional WMMSE algorithm grows substantially with an increasing number of CVs. This rapid growth can be attributed to the inherently iterative nature of WMMSE, which involves solving complex optimization problems in each iteration, resulting in a significant computational overhead as the number of vehicles scales up. In contrast, learning-based approaches, including the proposed LMM, exhibit a remarkable resilience to such increases in CV count, maintaining an average processing time in the millisecond (ms) range. This behavior demonstrates the efficiency and scalability of LMM.

Figure 4(b) extends the analysis by evaluating the average processing time under varying antenna configurations at the transmitter. Interestingly, the results reveal that the processing time for LMM remains relatively consistent across different antenna configurations. This stability in performance can be attributed to the parallel processing capabilities inherent in matrix operations involved in beamforming design. The ability of LMM to sustain millisecond-level processing times, even in the presence of complex antenna configurations, highlights its potential for deployment in real-world V2X systems. In practical terms, this translates to LMM being capable of making rapid beamforming decisions that can meet the

low latency requirements of V2X communications.

5 Conclusion

In this study, we focused on enhancing the effectiveness, adaptability, and robustness of beamforming design in V2X communication networks. By integrating an LMM with RSMA, we tackled the inherent challenges posed by complex road geometries in the beamforming process. Our proposed framework, named LMM, introduces a novel multi-modal learning architecture that dynamically extracts features from the historical CSI and the complex road geometry. Through extensive simulations, we demonstrated that LMM significantly outperforms existing schemes, including the conventional WMMSE algorithm and the latest predictive beamforming techniques utilizing Transformers. Moreover, the RSMA scheme integrated within our model exhibits clear superiority over conventional SDMA, particularly under conditions of channel uncertainty in V2X scenarios. The integration of an LMM and RSMA not only improves beamforming accuracy but also enhances system robustness, leading to substantial performance gains in V2X communications.

In the future, several key directions could further enhance the impact of these advancements. As V2X networks scale, centralized training may face limitations in terms of latency and scalability. Moving towards edge and federated learning frameworks, where local data are processed on-site, allows for faster adaptation to local traffic conditions, improving the overall system efficiency and privacy. Moreover, the integration of V2X systems with non-terrestrial networks (NTNs), such as satellite edge networks, presents an exciting direction for extending coverage and connectivity in remote or under-served areas.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant Nos. 62471208, 62371369), Guangdong Provincial Natural Science Foundation (Grant No. 2024A151510098), Shenzhen Science and Technology Program (Grant No. JCYJ20240813094627037), National Research Foundation, Singapore and Infocomm Media Development Authority under its Future Communications Research & Development Programme, National Natural Science Foundation for Distinguished Young Scholar (Grant No. 62425103), National Key R&D Program of China (Grant No. 2022YFC3301300), XPLORER PRIZE, Innovative Research Groups of the National Natural Science Foundation of China (Grant No. 62121001), Discipline Innovation and Talent Introduction Base of Colleges and Universities in Shaanxi Province, Natural Science Basis Research Plan in Shaanxi Province of China (Grant No. 2021JQ-205), and Dongguan Key Laboratory of Intelligent Equipment and Smart Industry, School of Advanced Engineering, Great Bay University.

References

1 Herrera-Quintero L F, Vega-Alfonso J C, Banse K B A, et al. Smart ITS sensor for the transportation planning based on IoT approaches using serverless and microservices architecture. IEEE Intell Transp Syst Mag, 2018, 10: 17–27
2 Zhao C, Lv Y, Jin J, et al. DeCAST in TransVerse for parallel intelligent transportation systems and smart cities: three decades and beyond. IEEE Intell Transp Syst Mag, 2022, 14: 6–17
3 Zhang S, Zhang S. Efficient federated connected electric vehicle scheduling system: a noncooperative online incentive approach. IEEE Trans Intell Transp Syst, 2025, 26: 3934–3946
4 Zhang S, Wang S, Yu S, et al. Collision avoidance predictive motion planning based on integrated perception and V2V communication. IEEE Trans Intell Transp Syst, 2022, 23: 9640–9653
5 Zhang S, Zhang S, Yuan W, et al. Efficient rate-splitting multiple access for the Internet of vehicles: federated edge learning and latency minimization. IEEE J Sel Areas Commun, 2023, 41: 1468–1483
6 Huang S, Zhang M, Gao Y, et al. MIMO radar aided mmwave time-varying channel estimation in MU-MIMO V2X communications. IEEE Trans Wireless Commun, 2021, 20: 7581–7594
7 Ding Z G, Schober R, Fan P Z, et al. Next generation multiple access for IMT towards 2030 and beyond. Sci China Inf Sci, 2024, 67: 166301
8 Huang Y Z, Zhang Z Y, Che J Z, et al. Self-attention reinforcement learning for multi-beam combining in mmWave 3D-MIMO systems. Sci China Inf Sci, 2023, 66: 162304
9 Mao Y, Dizdar O, Clerckx B, et al. Rate-splitting multiple access: fundamentals, survey, and future research trends. IEEE Commun Surv Tut, 2022, 24: 2073–2126
10 Clerckx B, Mao Y, Jorswieck E A, et al. A primer on rate-splitting multiple access: tutorial, myths, and frequently asked questions. IEEE J Sel Areas Commun, 2023, 41: 1265–1308
11 Mao Y, Clerckx B, Li V O K. Rate-splitting multiple access for downlink communication systems: bridging, generalizing, and outperforming SDMA and NOMA. J Wireless Com Netw, 2018, 2018: 133
12 Qi C, Wang X. Precoding design for energy efficiency of multibeam satellite communications. IEEE Commun Lett, 2018, 22: 1826–1829
13 Li X, Lu H, Zeng Y, et al. Near-field modeling and performance analysis of modular extremely large-scale array communications. IEEE Commun Lett, 2022, 26: 1529–1533
14 Feng C, Lu H, Zeng Y, et al. Near-field modeling and performance analysis for extremely large-scale IRS communications. IEEE Trans Wireless Commun, 2024, 23: 4976–4989
15 Li S, Wei Z, Yuan W, et al. Faster-than-nyquist asynchronous NOMA outperforms synchronous NOMA. IEEE J Sel Areas Commun, 2022, 40: 1128–1145
16 Li Z, Wang S, Zhang S, et al. Edge-assisted V2X motion planning and power control under channel uncertainty. IEEE Trans Veh Technol, 2023, 72: 9641–9646
17 Zhang S, Zhang S, Mao Y, et al. Transformer-based channel prediction for rate-splitting multiple access-enabled vehicle-to-everything communication. IEEE Trans Wireless Commun, 2024, 23: 12717–12730
18 Chen Z, Zhang Z, Yang Z, et al. Channel deduction: a new learning framework to acquire channel from outdated samples and coarse estimate. IEEE J Sel Areas Commun, 2025, 43: 944–958

19 Clerckx B, Mao Y, Yang Z, et al. Multiple access techniques for intelligent and multifunctional 6G: tutorial, survey, and outlook. Proc IEEE, 2024, 112: 832–879
20 Zhang S, Mao Y, Chen Z, et al. Federated learning-assisted predictive beamforming for extremely large-scale antenna array systems with rate-splitting multiple access. IEEE J Sel Top Signal Process, 2025, 19: 461–476
21 Mishra A, Mao Y, Dizdar O, et al. Rate-splitting multiple access for downlink multiuser MIMO: precoder optimization and PHY-layer design. IEEE Trans Commun, 2022, 70: 874–890
22 Krishnamoorthy A, Schober R. Downlink MIMO-RSMA with successive null-space precoding. IEEE Trans Wireless Commun, 2022, 21: 9170–9185
23 Amor D B, Joham M, Utschick W. Rate splitting in FDD massive MIMO systems based on the second order statistics of transmission channels. IEEE J Sel Areas Commun, 2023, 41: 1351–1365
24 Dizdar O, Sattarzadeh A, Yap Y X, et al. RSMA for overloaded MIMO networks: low-complexity design for max-min fairness. IEEE Trans Wireless Commun, 2024, 23: 6156–6173
25 Dizdar O, Mao Y, Clerckx B. Rate-splitting multiple access to mitigate the curse of mobility in (massive) MIMO networks. IEEE Trans Commun, 2021, 69: 6765–6780
26 Dizdar O, Mao Y J, Xu Y N, et al. Rate-splitting multiple access for enhanced URLLC and eMBB in 6G: invited paper. In: Proceedings of the 17th International Symposium on Wireless Communication Systems (ISWCS), 2021. 1–6
27 Zhu J, Chen Y, Pei X, et al. Rate-splitting multiple access with finite blocklength and high mobility for URLLC transmissions. IEEE Wireless Commun Lett, 2024, 13: 1518–1522
28 Zhang S Y, Zhang S Y, Yeung L K. Energy-efficient federated edge learning for internet of vehicles via rate-splitting multiple access. In: Proceedings of International Symposium on Wireless Communication Systems (ISWCS), 2022. 1–6
29 Zhang S, Zhang S, Yuan W, et al. Rate-splitting multiple access-based satellite-vehicular communication system: a noncooperative game theoretical approach. IEEE Open J Commun Soc, 2023, 4: 430–441
30 Yuan W, Liu C, Liu F, et al. Learning-based predictive beamforming for UAV communications with jittering. IEEE Wireless Commun Lett, 2020, 9: 1970–1974
31 Liu C, Yuan W, Li S, et al. Learning-based predictive beamforming for integrated sensing and communication in vehicular networks. IEEE J Sel Areas Commun, 2022, 40: 2317–2334
32 Liu C, Yuan W, Wei Z, et al. Location-aware predictive beamforming for UAV communications: a deep learning approach. IEEE Wireless Commun Lett, 2021, 10: 668–672
33 Wang Y, Gao Z, Zheng D, et al. Transformer-empowered 6G intelligent networks: from massive MIMO processing to semantic communication. IEEE Wireless Commun, 2023, 30: 127–135
34 Xu M, Niyato D, Chen J, et al. Generative AI-empowered simulation for autonomous driving in vehicular mixed reality metaverses. IEEE J Sel Top Signal Process, 2023, 17: 1064–1079
35 Xu M, Du H, Niyato D, et al. Unleashing the power of edge-cloud generative AI in mobile networks: a survey of AIGC services. IEEE Commun Surv Tut, 2024, 26: 1127–1170
36 Liu Y, Du H, Niyato D, et al. Deep generative model and its applications in efficient wireless network management: a tutorial and case study. IEEE Wireless Commun, 2024, 31: 199–207
37 Xie G, Xiong Z, Zhang X, et al. GAI-IoV: bridging generative AI and vehicular networks for ubiquitous edge intelligence. IEEE Trans Wireless Commun, 2024, 23: 12799–12814
38 Du H, Zhang R, Liu Y, et al. Enhancing deep reinforcement learning: a tutorial on generative diffusion models in network optimization. IEEE Commun Surv Tut, 2024, 26: 2611–2646
39 Yang W, Xiong Z, Quek T Q S, et al. Streamlined transmission: a semantic-aware XR deployment framework enhanced by generative AI. IEEE Netw, 2024, 38: 29–38
40 Meng X, Liu F, Masouros C, et al. Vehicular connectivity on complex trajectories: roadway-geometry aware ISAC beamtracking. IEEE Trans Wireless Commun, 2023, 22: 7408–7423
41 Kong X R, Braunl T, Fahmi M, et al. A superalignment framework in autonomous driving with large language models. 2024. ArXiv:2406.05651
42 Wen L C, Fu D C, Li X, et al. DiLu: a knowledge-driven approach to autonomous driving with large language models. 2023. ArXiv:2309.16292
43 Ullah A, Qi G L, Hussain S, et al. The role of LLMs in sustainable smart cities: applications, challenges, and future directions. 2024. ArXiv:2402.14596
44 Lee W, Park J. LLM-empowered resource allocation in wireless communications systems. 2024. ArXiv:2408.02944
45 Shao J W, Tong J W, Wu Q, et al. WirelessLLM: empowering large language models towards wireless intelligence. 2024. ArXiv:2405.17053
46 Sheng Y, Huang K, Liang L, et al. Beam prediction based on large language models. 2024. ArXiv:2408.08707
47 Zheng T, Dai L. Large language model enabled multi-task physical layer network. 2024. ArXiv:2412.20772
48 Yang Z, Chen M, Liu Y, et al. A joint communication and computation framework for digital twin over wireless networks. IEEE J Sel Top Signal Process, 2024, 18: 6–17
49 Rubinstein R, Kroese D. Simulation and the Monte Carlo Method. 3rd ed. Hoboken: John Wiley & Sons, 2016
50 Federal Highway Administration. Next Generation Simulation (NGSIM). Accessed on: 2024-03-13. https://ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm
51 Yang H, Zhao J, Xiong Z, et al. Privacy-preserving federated learning for UAV-enabled networks: learning-based joint scheduling and resource management. IEEE J Sel Areas Commun, 2021, 39: 3144–3159