

Task-Driven Semantic-Aware Green Cooperative Transmission Strategy for Vehicular Networks

Wanting Yang[✉], Member, IEEE, Xuefen Chi[✉], Linlin Zhao[✉], Member, IEEE,
Zehui Xiong[✉], Member, IEEE, and Wenchao Jiang[✉]

Abstract— Considering the infrastructure deployment cost and energy consumption, it is unrealistic to provide seamless coverage of the vehicular network. The presence of uncovered areas tends to hinder the prevalence of the in-vehicle services with large data volume. To this end, we propose a predictive cooperative multi-relay transmission strategy (PreCMTS) for the intermittently connected vehicular networks, fulfilling the 6G vision of semantic and green communications. Specifically, we introduce a task-driven knowledge graph (KG)-assisted semantic communication system, and model the KG into a weighted directed graph from the viewpoint of transmission. Meanwhile, we identify three predictable parameters about the individual vehicles to perform the following anticipatory analysis. Firstly, to facilitate semantic extraction, we derive the closed-form expression of the achievable throughput within the delay requirement. Then, for the extracted semantic representation, we formulate the mutually coupled problems of semantic unit assignment and predictive relay selection as a combinatorial optimization problem, to jointly optimize the energy efficiency and semantic transmission reliability. To find a favorable solution within limited time, we proposed a low-complexity algorithm based on Markov approximation. The promising performance gains of the PreCMTS are demonstrated by the simulations with realistic vehicle traces generated by the SUMO traffic simulator.

Index Terms— Vehicular network, store-carry-forward, proactive cooperative transmission, semantic-aware, Markov approximation.

Manuscript received 5 March 2023; revised 15 June 2023; accepted 26 July 2023. Date of publication 1 August 2023; date of current version 18 October 2023. This research is supported by National Natural Science Foundation of China under Grant 62271228, the Jilin Scientific and Technological Development Program under Grant 20210201016GX and 20230101063JC and the China Scholarship Council CSC NO. 202106170088. This research is also supported by the National Research Foundation, Singapore, and Infocomm Media Development Authority under its Future Communications Research & Development Programme. The research is also supported by the SUTD SRG-ISTD-2021-165, the SUTD-ZJU IDEA Grant (SUTD-ZJU (VP) 202102), the Ministry of Education, Singapore, under its SUTD Kickstarter Initiative (SKI 20210204) and Ministry of Education, Singapore, under its AcRF Tier 2 (MOE-T2EP2022-0017). The associate editor coordinating the review of this article and approving it for publication was R. D. Souza. (Corresponding author: Xuefen Chi.)

Wanting Yang is with the Department of Communications Engineering, Jilin University, Changchun 130012, China, and also with the Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore 487372 (e-mail: yangwt18@mails.jlu.edu.cn).

Xuefen Chi and Linlin Zhao are with the Department of Communications Engineering, Jilin University, Changchun 130012, China (e-mail: chixf@jlu.edu.cn; zhaoll13@mails.jlu.edu.cn).

Zehui Xiong and Wenchao Jiang are with the Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore 487372 (e-mail: zehui_xiong@sutd.edu.sg; wenchao_jiang@sutd.edu.sg).

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TCOMM.2023.3300869>.

Digital Object Identifier 10.1109/TCOMM.2023.3300869

I. INTRODUCTION

THE burgeon of the intelligent transportation system has spawned numerous innovative in-vehicle services to make mobility much safer and easier. Most services such as road sign recognition and situation understanding heavily rely on scene understanding [1], [2], [3], which are characterized by the large data volume. However, due to the high cost of infrastructure deployment and energy consumption, it is unrealistic to install sufficient roadside units (RSUs) to provide seamless coverage [4]. The presence of low-throughput intermittently connected vehicular networks (ICVNs) inevitably hinders the popularity of these services.

Thanks to the boom in artificial intelligence, semantic communication (SemCom)¹ has evolved from a theoretical concept to a 6G enabler. Exploiting the intelligence of vehicles, SemCom can achieve a significant reduction in transmission burden, thus mitigating the impact of ICVNs on quality of service. For example, deep learning (DL) is now commonly used to perform human-like understanding at transmitters and receivers, which are termed as semantic encoding and semantic decoding, respectively [6]. Therein, the irrelevant information about the target communication task is filtered out before transmission, and only a small data volume carrying valuable information is transmitted to receivers for the downstream inference task [7]. The promising performance gains achieved by SemCom in low channel conditions has been widely demonstrated [8], [9], [10]. Nonetheless, the black box nature of the DL-based SemCom results in low social acceptance. Moreover, the focus of existing SemCom research is mostly on the semantic processing of transceivers, where wireless environment is simplified to a channel model, such as Rayleigh channel and Rician channel. This makes it infeasible for dynamic complex vehicle networks, where the average channel gain experienced by users changes significantly. Thus, an explainable and generalized SemCom paradigm is called for.

Fortunately, the recent studies on the convergence of knowledge graph (KG) and explainable computer vision, holds promise towards the mentioned expectations. KG can provide a structured semantic representation (SR) for road traffic scenes [1], which can be seen as a container of semantic information. In contrast to the underlying raw data formats of the practical scene, the great extensibility of KG allows the

¹In our work, SemCom refers to the communication that reaches the semantic level or the effectiveness level, which are defined by Weaver in [5].

KG-based SR to be partially updated according to the dynamic changes of scenes, e.g., new roadblocks [11]. Meanwhile, the semantic information for different target tasks can be flexibly extracted in form of a sub-KG. For instance, for users interested in the road traffic, only the sub-KG related to pedestrian and traffic flow needs to be transmitted, and building-related sub-KG can be automatically filtered out. Nonetheless, a completed end-to-end model of a universal KG-based SemCom system is still a gap in the available research. Furthermore, the distinctive feature of SemCom lies in that data is assigned diverse significance [6]. From a well-established KG, both the semantic importance of each semantic unit (SU) and the number of the bits required to carry SU viewed from the physical form can be obtained, which creates the opportunity for the finer-grained semantic-aware transmission strategy design. For instance, the SUs of greater importance can be transmitted with higher power, wider bandwidth, or more reliable links [12] to enhance the semantic transmission reliability. However, this cannot be easily realized by straightforward refinements to existing schemes.

Especially, for ICVN, most research efforts are devoted into the multi-hop relay transmission for lightweight services with strict delay requirements [13], [14], where the relays serve for real-time amplification/decoding and forwarding. If they are applied to the services with large volume, much unwarranted communication overhead and cache pressure on the relays are introduced. Given this, the *one-hop* store-carry-forward scheme (SCFS) [15] is more appropriate for the considered case, where the mobility of relay vehicles is utilized to physically propagate information messages to reduce the outage areas [15]. However, the existing studies on SCFS only concentrate on the physical layer, such as minimizing the outage time [16], statistical analysis of achievable throughput gain [4]. Few of them care about the properties of the communication task, even for the content size and maximum acceptable delay. As a result, they cannot achieve on-demand fulfillment and are further away from semantic awareness. More critically, as the energy efficiency varies greatly depending on the vehicle location, the total energy consumption is strongly related to the selected relays. Thus, these ready-made SCFSs with uniform relay selection rules tend to cause different levels of energy waste for a specific task, depending on real-time on-road vehicles' location and speed. This goes against green communication in 6G.

In light of the above, to meet the 6G vision of SemCom and green communication, we propose a novel predictive cooperative multi-relay transmission strategy (PreCMTS) for large download for ICVN. In the strategy, the SR selection, relay selection, and SU assignment are all highly related to three predictable parameters: the residual dwell time of the vehicles in their associated RSUs, as well as the encounter time and V2V link lifetime of each relay with the target vehicle. The major contributions are highlighted as follows.

- We introduce a general task-driven KG-assisted SemCom system model, where both the semantic encoding and decoding are performed based on the KG. Moreover,

from the standpoint of transmission, we model the KG as a weighted directed graph (WDG), where the vertices represent the indivisible SUs that are the embedding of real-world objects and their abstract relationships, and the directed edges characterize the dependence of SUs. To enable semantic-aware transmission, the significance degree and the data size viewed from semantic and physical level for each SU are recorded as edge weights.

- To facilitate semantic extraction (SE) to get an appropriate SR, we derive the closed-form expression of the achievable throughput within the maximum acceptable delay according the current road traffic situation. Moreover, for the selected SR, we formulate the mutually coupled problems of predictive relay selection and SU assignment as a combinatorial optimization problems with the aim to minimize energy consumption while guaranteeing the semantic transmission reliability under imperfect speed prediction. Therein, the constraints of the V2V link interference, the end-to-end transmission delay², and the bottleneck of the two cascade store-carry-forward links are all considered.
- To find a favorable solution within limited time, we design a low-complexity multi-threaded search algorithm based on Markov approximation. Moreover, we devise an SU assignment algorithm following the basic SCFS in [4] as a baseline to generate the initial state. From the simulation results, in PreCMTS, the SUs with high semantic significance are more likely assigned to the direct transmission link to enhance semantic reliability and the vehicles close to the RSU are preferred to be selected as relays to pre-stores SUs to save energy compared to the baseline. The promising performance gains in terms of energy saving, semantic transmission reliability, and semantic energy efficiency are demonstrated.

In the following sections, we first review the related works about SemCom and SCFS, respectively. Then, we describe the system model and highlight the overview of the proposed semantic-aware PreCMTS in Section III. Then, the details of the proposed scheme are presented in Section IV. Section V presents the simulation results, and Section VI concludes this paper. Besides, the notations of relevant parameter symbols are listed in Table I.

II. RELATED WORKS

A. Semantic Communication

Based on our previous review works [6], [7], the existing research on SemCom can be broadly classified into four categories depending on the SE method. The first and most studied category is deep-learning (DL)-based end-to-end SemCom. The employed semantic encoder and decoder are two separate learnable neural networks, and linked through a layer for modeling random channels [17]. They are trained jointly, based on a complete data set shared by both senders and receivers. Thanks to the advancement of DL models, e.g.,

²The end-to-end delay in our work refers to the time interval between the moment when the target vehicle sends the request and the moment when the target vehicle receives all the requested data.

TABLE I
LIST OF RELEVANT NOTATIONS

Notation	Description	Notation	Description
$r_1(r_V)$	Communication radius of RSU (vehicle)	$R_1(R_V)$	Data transmission rate of V2I (V2V) link
D_i^1	Maximum duration of V2I link for vehicle v_i	T_{\max}	Maximum acceptable delay
Δ_i^T	Moment when relay vehicle v_i^R enters the communication range of the target vehicle	D_i^T	Maximum duration of the V2V link between target vehicle and relay vehicle v_i
\hat{C}_i^1	Maximum data amount that can be transmitted via V2I link to vehicle $v_i \in \mathcal{V}$	$\hat{\delta}_i$	Moment when relay v_i starts to forward data in achievable throughput analysis
\hat{C}_i^V	Maximum data amount transmitted to vehicle v_0 by relay $v_i \in \mathcal{V}_R$ for a given Φ	$t_i^{S_V}$	Moment when relay v_i starts to forward data to v_0 for a given transmission strategy
β_j	Data size of SU j	α_j	Contribution of SU j to the accuracy of SR

Transformer, the high efficient SE for text, image, and audio, achieves significantly performance gains especially at low signal-to-noise ratio (SNR) region [8], [10], [18]. Nevertheless, the back-propagation in DL paradigm requires the loss function to be differentiable, which hinders the sophisticated non-differentiable semantic metrics from being applied to guide the training. To solve this issue, deep reinforcement learning paradigm is adopted to perform SE [19]. However, the above two categories of SemCom are available only for the simplest point-to-point communication model, which cannot be directly applied to complex real-world scenarios. Moreover, the black box nature also restricts their social acceptance [6].

Meanwhile, with the development of the explainability of AI technologies, some researchers propose the knowledge base (KB)-assisted SemCom. Herein, the KB is a special database for semantic knowledge management, which consists of semantic elements embedded in the source data, the involving communication tasks, and the possible ways of reasoning by communication participants [7]. Up to now, there are two available kinds of general KB models. One is based on a hierarchical structure [20], and the other is based on graph structure [7], [21]. Moreover, there have been several technical research [22], [23] on SemCom for text transmission based on the available the interconversion technologies for text and graph. However, in the above works, the resource allocation algorithm is still following the philosophy of traditional *content-blind* resource allocation paradigm, i.e., allocating radio resources to per user according to their required data volume. In this sense, a semantic-aware transmission has not been achieved in a real sense.

In addition to the above three categories of SemCom, there is also a semantic-native SemCom paradigm, wherein the semantic information can be learned from iterative communications between intelligent agents [24]. However, this study is still stuck in the theoretical analysis based on a simple ideal model. It remains a huge challenge to put it into practice. In this sense, we focus on the KB-assisted SemCom in our work.

B. Store-Carry-Forward Scheme

The core concept of SCFS is to utilize the mobility of relays to physically propagate information messages [15], which is first proposed in [25]. Specifically, in SCFS, a relay vehicle, which will pass the target vehicle within their uncovered area, pre-stores partial data requested by the target vehicle in

advance over an available V2I link. It then carries and forwards data to the target vehicle until they encounter each other.

Initially, in [26] and [27], the authors investigate the optimization of the target vehicle speed control with the objective of minimization of the outage time. However, it deviates from the design philosophy of user experience-oriented communication nowadays, and it is unrealistic to control the vehicle's speed without consideration of the actual traffic conditions and the driver's driving habits. Additionally, the above works only focus on the unidirectional road model, and thus the mobility pattern of the vehicles is not fully exploited. To that end, a bidirectional road is considered in [28] and [16], where the mobility of vehicles can be utilized to physically propagate information messages. Different from [26] and [27], the authors in [28] and [16] propose some essential relay selection constraints on the relay link lifetime, residual dwell time, and buffer time, which can jointly determine which candidate vehicles can be picked as relays. All the above works just focus on the minimization of the outage time.

In [29] and [4], the authors derive a closed-form expression of the achievable throughput of the SCFS for a bidirectional road with vehicle flow obeying Poisson distribution. In [29], two assumptions are made. The first one is that there is no possibility that the relay vehicle is still within the available V2V connection range, but it has no data to forward to the target vehicle. The second one is that there is no interference between different V2V links, i.e., the target vehicle can maintain multiple V2V links simultaneously. Furthermore, in [4], the authors propose an elastic-segment-based V2V/V2I cooperative strategy, where the second assumption is removed, and a commonly used interference model [30] is adopted, that is, only one V2V link is active at any given time. The adopted assumption is strongly dependent on the specific scenario, and thus compromising the generality of their work.

Moreover, the existing works focus on the enhancement and evaluation of the physical layer performance. Few of them considers the demand of communication task and the energy efficiency of the communication system. To this end, a task-oriented SCFS is promising to embrace the green communication in 6G with on-demand fulfillment.

III. SYSTEM OVERVIEW

A. Scenario Description

This paper focuses on one segment of a bidirectional road, which runs through the coverage of two adjacent RSUs

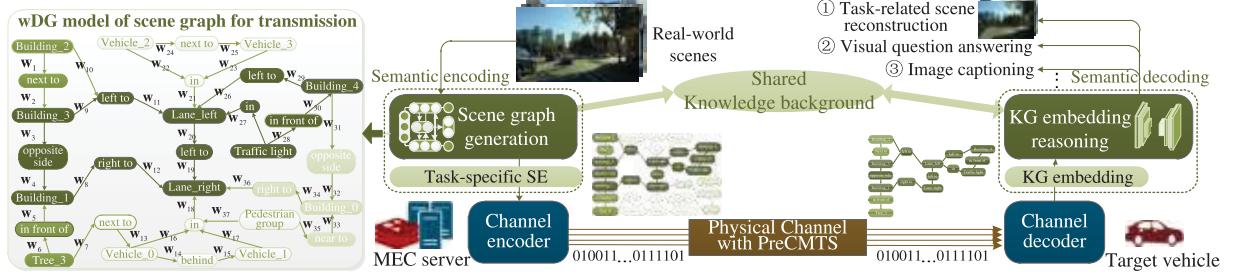


Fig. 1. KG-assisted SemCom system model.

(indexed by RSU A and RSU B, respectively). The distance between the two RSUs is denoted by H , and the coverage radius of each RSU is denoted by r_1 . Considering the restricted transmit power and the high deployment costs, we assume that there is an outage area between the RSUs, i.e., $H > 2r_1$.

Without loss of generality, we assume that a vehicle within the coverage of RSU A sends a request to multi-access edge computing (MEC) server for a large download. The maximum allowable delay of the services is denoted by T_{\max} . To complete the transmission within T_{\max} , we propose a semantic-aware PreCMTS, which is performed by a central controller (CC) at the MEC server. The vehicle sending the request is referred to as the target vehicle and denoted by v_0 . The relay candidates are the vehicles driving in the opposite direction to the target vehicle within the coverage of RSU B. The set of the relay candidates is denoted by \mathcal{V}_R , and each of them is indexed by $v_i \in \mathcal{V}_R, i \in \{1, 2, \dots, |\mathcal{V}_R|\}$. Specifically, the serial number of the relay candidate is arranged according to the sequence of them entering the target vehicle's communication range. Similar to the RSUs, the communication range is the same for all vehicles, the radius of which is denoted by r_V . Since the RSU typically has stronger communication capability than the vehicle, we have $r_1 > r_V$ [29]. Moreover, for ease of reference, we denote the set composed by the target vehicle and the relay candidates by \mathcal{V} , i.e., $\mathcal{V} = \mathcal{V}_R + \{v_0\}$. In addition, we assume that the average speed remains constant [4], and the average speed of each vehicle $v_i \in \mathcal{V}$ is denoted by $\bar{u}_i, i \in \{0, 1, 2, \dots, |\mathcal{V}_R|\}$. To facilitate an energy efficient scheme, the vehicles are required to report the information about their speed and position to the MEC server, which enables the possibility of predictive relay selection and strategic pre-store the data in the relays under better channel states.

B. Transceiver Semantic Processing Model

The proposed task-driven KG-assisted SemCom system model is shown in Fig. 1, where a two-dimensional image is taken as an example of semantic encoding input.

The semantic encoding performed at the MEC server consists of two modules. Firstly, the *scene graph generation* module bridges the gap between visual and semantic perception of the real-world scene³. Then, the well-developed scene graph, i.e., a KG, can be regarded

³It typically undergoes four steps: off-the-shelf object detectors, feature representation, feature refinement, and relationship prediction [31].

as a container for all the semantic information implied by the scene. It is composed of multiple linked triples in the form of $\langle \text{head_object}, \text{relation}, \text{tail_object} \rangle$, e.g., $\langle \text{building1}, \text{right}, \text{lane1} \rangle$. The embedding⁴ of each element in each triples are treated as an undividable SU. To facilitate semantic-aware transmission, the scene graph is re-modeled as a mathematical form of wDG as shown on the left side of Fig. 1. The SUs are treated as the vertices. The directed edges retain the dependency between the two objects. In general, the significance of SUs varies for different tasks. For instance, for users who intend to check the map of a certain place, information about pedestrian and traffic flow on the road is no longer necessary; on the contrary, for users who prefer to know the road traffic, detailed information about the surrounding buildings can be ignored. Therefore, we assign an array, $\mathbf{w}_j = [w_{j,1}, \dots, w_{j,k}, \dots, w_{j,K}]$, as the weight corresponding to an SU j , where K represents the number of the tasks and w_k is in form of a binary tuple $w_{j,k} = (\alpha_{j,k}, \beta_{j,k})$, with $\alpha_{j,k}$ denoting the quantified importance degree of SU j to task k and $\beta_{j,k}$ denoting the number of bits required to carry the information of SU j . Without loss of generality, only one task k is considered in our work. To simply the notation, the subscript k is omitted in this manuscript, and the weight for SU j is simplified to $w_j = (\alpha_j, \beta_j)$. Then, based on the wDG, the *task-specific SE* module extracts an SR in form of an edge-induced subgraph of the original wDG. To ensure the completeness of the transmission within the allowable maximum delay T_{\max} , the edges with higher semantic importance have priority to be added to the edge subset used to generate the SR, while ensuring that the total data size is less than the achievable throughput. On this premise, the cardinal number of edge subset, which determines the number of the chosen SUs, can be decided based on a specific trade-off between semantic accuracy and energy consumption. Moreover, we assume that each SU in the selected SR is encapsulated individually according to the edge weight $w_j = (\alpha_j, \beta_j)$ [33]. That is, during the transmission, the data for each SU cannot be further split.

Upon obtaining the complete task-related sub-KG, the target vehicle performs the semantic decoding, which is also accomplished by two modules. First, the *KG embedding* module is responsible for embedding the objects and relations of the

⁴The embedding is a low-dimensional vector (being in accord with Word2vec in NLP) [32], which are obtained in the KG generation via an visual translation embedding methods [31].

sub-KG into a low-dimensional vector [31]. Then, taking the low-dimensional vector as the input, the *KG-embedding reasoning* module performs the downstream semantic inference based on a cascaded sophisticated network specially designed for the particular task, such as task-related scene reconstruction, visual question answering, and image captioning [31], [32].

It should be noted that, both sides of the communication are required to share their knowledge background about historical scenes and all the possible tasks, which allows the training process for semantic encoding and decoding to match each other. Due to the limitation of space, the synchronization of knowledge background are beyond the scope of this work. Meanwhile, the communication overhead for the background knowledge, the computing resources for the KG generation and update, and the storage resources for the KG are not discussed in this paper but will be studied in the future works.

C. Wireless Link Transmission Model

Without loss of generality, we assume that every passing vehicle is equipped with one antenna [4], [29], which allows a vehicle to maintain only one link at a time, either a V2I link or a V2V link. Meanwhile, to avoid interference, if a V2V link already exists within a vehicle communication range, it will not be able to transmit data [4]. Moreover, considering the sophisticated technologies available in RSUs, such as frequency division multiplexing and multi-user beamforming, we assume that the RSU can simultaneously transfer data to multiple users without inter-user interference [4], [34].

As depicted in Section III-A, the extensively used disk model is employed to characterize the V2I and V2V connection [4], [29]. That is, any vehicle pair or vehicle-RSU pair is able to be connected if the distance between each other is less than r_V or r_I [35]. We denote the distance between any pair of transmitter and receiver by d . The large-scale channel gain, then, can be characterized by the standard power-law path loss $G_x(d) = b_x d^{-a_x}$, where a_x is the path loss exponent, b_x is the reference path loss at a unit distance, and $x \in \{I, V\}$ is set to differentiate the V2I and V2V links [36], [37]. Furthermore, considering the high mobility of vehicles and the inevitable inter-vehicle large vehicle obstructions, e.g., buses, we adopt the \mathcal{F} composite fading model to characterize the small-scale fading, where the combined effects of multi-path and shadowing are taken into account [38]. We denote the small-scale channel gain by \tilde{g} . Accordingly, the probability density function of \tilde{g} is expressed by [38]

$$f(\tilde{g}) = \frac{m^m (m_s - 1)^{m_s} \bar{g}^{m_s} \tilde{g}^{m-1}}{B(m, m_s) [m \bar{g} + (m_s - 1) \bar{g}]^{m+m_s}}, \quad (1)$$

where m , m_s , and \bar{g} represents the number of clusters of multipath, shadowing shape, and average small-scale channel gain, respectively, and $B(\cdot, \cdot)$ denotes the beta function [38]. We assume that the power control technique is adopted, where the decoding threshold of SNR for the V2I and V2V link are denoted by Υ_I and Υ_V , respectively. Upon assuming perfect capacity achieving coding, the achievable transmission rate is

expressed by

$$R_x = B_x \log (1 + \Upsilon_x), \quad (2)$$

where $x \in \{I, V\}$. Moreover, to simplify problem analysis, we assume that the bandwidth allocated to all the V2I links are fixed and the same [4], i.e., $B_I = B_V$. The instantaneous transmit power, then, is expressed as

$$p_x(d) = \begin{cases} \frac{\Upsilon_x \sigma^2}{A_x G_x(d) \tilde{g}}, & d \leq r_x \\ 0, & d > r_x, \end{cases} \quad (3)$$

where $x \in \{I, V\}$. Specifically, A_I (A_V) denotes the joint antenna gain of the transmitter and receiver of the V2I (V2V) link. Moreover, Υ_I (Υ_V) represents the decoding threshold of signal-to-noise (SNR) for the V2I (V2V) link. Meanwhile, we assume that the small-scale gains are independently and identically distributed (i.i.d.) among transmission time intervals. Then, the average transmit power with distance d between the transmitter and receiver can be expressed by

$$\begin{aligned} \bar{p}_x(d) &= \mathbb{E}_{\tilde{g}}[p_x] = \int_0^\infty \frac{\Upsilon_x \sigma^2}{A_x G_x(d) \tilde{g}} f(\tilde{g}) d\tilde{g} \\ &= \frac{\Upsilon_x \sigma^2}{A_x G_x(d)} \int_0^\infty \tilde{g}^{-1} f(\tilde{g}) d\tilde{g} \\ &= \frac{\Upsilon_x \sigma^2}{A_x G_x(d)} \mathbb{E}[\tilde{g}^{-1}], \end{aligned} \quad (4)$$

According to (1), with the aid of [39, eq. (3.194.3)], the n^{th} moment of \tilde{g} can be derived as

$$\mathbb{E}[\tilde{g}^n] = \frac{(m_s - 1)^n \bar{g}^n \Gamma(m + n) \Gamma(m_s - n)}{m^n \Gamma(m) \Gamma(m_s)} \quad (5)$$

where $\Gamma(\cdot)$ represents the gamma function. Substituting the case of $n = -1$ in (5) into (4), we can obtain the final expression of $\bar{p}_x(d)$ as below.

$$\bar{p}_x(d) = \frac{\Upsilon_x \sigma^2}{A_x G_x(d)} \frac{m \Gamma(m-1) \Gamma(m_s+1)}{(m_s-1) \bar{g} \Gamma(m) \Gamma(m_s)}. \quad (6)$$

For brevity, we rewrite (6) as $\bar{p}_x(d) = M \frac{\Upsilon_x \sigma^2}{A_x G_x(d)}$, where $M = \frac{m \Gamma(m-1) \Gamma(m_s+1)}{(m_s-1) \bar{g} \Gamma(m) \Gamma(m_s)}$ is a constant.

D. Overview of Semantic-Aware PreCMTS

As shown in Fig. 2, the proposed PreCMTS consists of three stages, namely preparation for PreCMTS, development for PreCMTS, and execution for PreCMTS, respectively.

In Stage 1, the vehicles and the MEC server exchange the necessary information for the strategy design. Specifically, when the target vehicle sends a task request to the MEC server via RSU A, it sends the information about its current position and average speed to the CC at the same time. Then, the CC broadcasts the request for cooperative transmission to the vehicles in the coverage of RSU B. After receiving the request, the relay candidates of the vehicles driving in the opposite direction to the target vehicle send their current position and average speed back to the CC.

In Stage 2, according to the information reported by the vehicles, the CC first predicts the key parameters about the

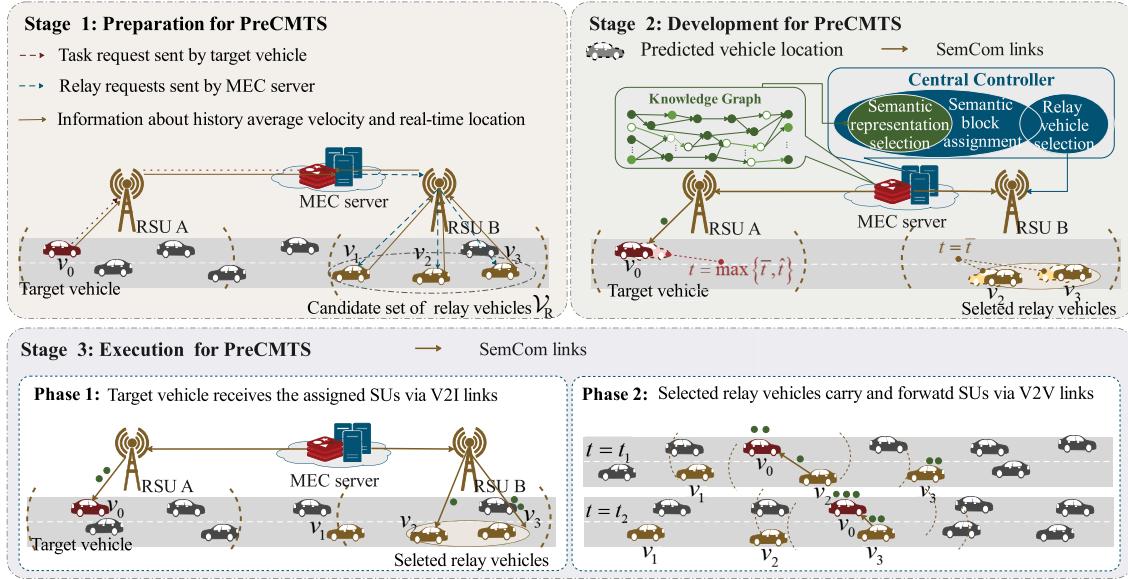


Fig. 2. Overview of the semantic-aware PreCMTS.

vehicle trajectories. Then, based on the above predictable parameters, the CC derives the achievable throughput Q_{\max} within T_{\max} , which is fed to the SE module to extract an appropriate SR. Then, we denote the computational latency for the high-dimensional optimization problem of the cooperative transmission strategy by \bar{t} . Considering the high dynamic nature of the vehicle network, the CC first predicts the locations of the relay candidates at time \bar{t} to ensure a well-matched PreCMTS to the practical world. To mitigate the impact of computing latency on the transmission delay, the SUs with large α can be transmitted to the target vehicle in advance via V2I link upon the determination of SR, before the completion of the algorithm execution. We denote the end moment of SU's advance transmission by \hat{t} . Considering the indivisibility of SUs, the CC needs to predict the location of the target vehicle at time $\max\{\bar{t}, \hat{t}\}$. Next, for the given SR, the CC develops a cooperative transmission strategy Φ^* based on the predicted locations, which jointly determines the mutually coupled relay selection and SU assignment.

In Stage 3, the transmission process begins. It can be further divided into two phases. In phase 1, according to Φ^* , the SUs assigned to the direct V2I link are transmitted to the target vehicle directly via RSU A as scheduled. At the same time, the other SUs are transmitted simultaneously to the corresponding relay vehicles in advance via RSU B. When a relay receives all the SUs assigned to it or it leaves the coverage of RSU B, the corresponding V2I link is disconnected. Then, in phase 2, the relays forward pre-stored SUs to the target vehicle in order of encounter sequence. It is to be noted that the relay selection is performed proactively in Stage 1. In this phase, only the V2V communication following strategy Φ^* happens, and neither

the target vehicle nor the CC needs to perform further relay selection.

IV. PREDICTIVE COOPERATIVE MULTI-RELAY TRANSMISSION STRATEGY

A. Preliminary

In this section, we first introduce the predictive parameters used to develop the PreCMTS, i.e., the residual dwell time of the vehicles in their associated RSUs, as well as the encounter time and V2V link lifetime of each relay with the target vehicle. To concise the notation system and without loss of the generality of the following analysis, we assume the computing latency $\bar{t} = 0$, that is, treating the moment when the target vehicle sends the request as the initial moment of the transmission process, and thus $\hat{t} = \bar{t}$. For ease of illustration, we take an example of a two-vehicle encounter process as shown in Fig. 3. We denote the initial distance to the connected RSU and position of vehicle $v_i \in \mathcal{V}$ by d_i and ℓ_i , respectively, where $0 < |l_i| = d_i \leq r_I$. If the offset of the initial distance is in the same direction as the vehicle drives, $\ell_i = d_i$, and otherwise, $\ell_i = -d_i$. We denote the average speed of vehicle $v_i \in \mathcal{V}$ by \bar{u}_i and the relative speed of the two vehicles by $\hat{u}_i = \bar{u}_i + \bar{u}_0$. Considering the fact that, in practice, the road length is much larger than the road width and the RSU height, we ignore the road width and the RSU length [16]. As such, the communication distance of a V2I and V2V link at time t can be expressed by $d_i^I(t) = |l_i + \bar{u}_i t|$ and $d_i^V(t) = |r_V - \hat{u}_i(t - \Delta_i^T)|$, respectively. Meanwhile, the residual dwell time of the vehicles in their associated RSUs can be predicted by

$$D_i^I = \frac{r_I - \ell_i}{\bar{u}_i}, \quad i \in \{0, 1, \dots, |\mathcal{V}_R|\}. \quad (7)$$

We refer to driving into the communication range of the target vehicle as an encounter with the target vehicle. The encounter

⁵The value of \bar{t} is jointly determined by the server computational capability and the expected performance gain of PreCMTS. It is to be noted that the vehicle driving out of RSU B's coverage within \bar{t} will not be able to act as a relay.

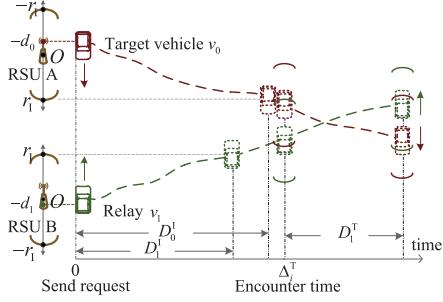


Fig. 3. Diagram of vehicle encountering.

time between relay v_i and vehicle v_0 can be predicted by

$$\Delta_i^T = \frac{H - \ell_0 - \ell_i - r_V}{\hat{u}_i}, \quad i \in \{1, \dots, |\mathcal{V}_R|\}. \quad (8)$$

Moreover, the duration of relay v_i within the coverage of vehicle v_0 can be predicted by

$$D_i^T = \frac{2r_V}{\hat{u}_i}, \quad i \in \{1, \dots, |\mathcal{V}_R|\} \quad (9)$$

Based on the above predictable parameters, the cumulative data amount transmitted via each V2I link or V2V link (without considering the existence of other vehicles communicating), can be calculated by $C_i^I = R_I D_i^I$, ($i \in \{0, 1, \dots, |\mathcal{V}_R|\}$) and $C_i^V = R_V D_i^T$, ($i \in \{1, \dots, |\mathcal{V}_R|\}$), respectively. Meanwhile, considering the mutual independence of large- and small-scale channel gain, the average overall energy consumption for a link can be calculated by

$$f_i^x(t^S, t^E) = \int_{t^S}^{t^E} \int_0^\infty \frac{\Upsilon_x \sigma^2}{A_x G_x(d_i^x(t)) \tilde{g}} f(\tilde{g}) d\tilde{g} dt \\ = \int_{t^S}^{t^E} \mathbb{E}_{\tilde{g}} [p_x(d_i^x(t))] dt, \quad (10)$$

where t^S and t^E represent the start and end time of a link, respectively. By substituting (6) into (10), the final expression of (10) can be obtained, as shown in (11) and (12), at the bottom of the page.

$$f_i^{V2I}(t^S, t^E) = \begin{cases} \frac{M\Upsilon_I \sigma^2}{A_I b_I \bar{u}_i (a_I + 1)} \left((\ell_i + \bar{u}_i t^E)^{a_I + 1} - (\ell_i + \bar{u}_i t^S)^{a_I + 1} \right), & \ell \geq 0, 0 \leq t^S \leq t^E \leq D_i^I \\ \frac{M\Upsilon_I \sigma^2}{A_I b_I \bar{u}_i (a_I + 1)} \left((-\ell_i - \bar{u}_i t^S)^{a_I + 1} - (-\ell_i - \bar{u}_i t^E)^{a_I + 1} \right), & \ell < 0, 0 \leq t^S \leq t^E \leq -\frac{\ell_i}{\bar{u}_i} \\ \frac{M\Upsilon_I \sigma^2}{A_I b_I \bar{u}_i (a_I + 1)} \left((-\ell_i - \bar{u}_i t^S)^{a_I + 1} + (\ell_i + \bar{u}_i t^E)^{a_I + 1} \right), & \ell < 0, 0 \leq t^S - \frac{\ell_i}{\bar{u}_i} \leq t^E \leq D_i^I \\ \frac{M\Upsilon_I \sigma^2}{A_I b_I \bar{u}_i (a_I + 1)} \left((\ell_i + \bar{u}_i t^E)^{a_I + 1} - (\ell_i + \bar{u}_i t^S)^{a_I + 1} \right), & \ell < 0, -\frac{\ell_i}{\bar{u}_i} \leq t^S \leq t^E \leq D_i^I \end{cases} \quad (11)$$

$$f_i^{V2V}(t^S, t^E) = \begin{cases} \frac{M\Upsilon_V \sigma^2}{A_V b_V \hat{u}_i (a_V + 1)} \left((r_V - \hat{u}_i (t^S - \Delta_i^T))^{a_V + 1} - (r_V - \hat{u}_i (t^E - \Delta_i^T))^{a_V + 1} \right), & 0 \leq t^S \leq t^E \leq \frac{D_i^T}{2} \\ \frac{M\Upsilon_V \sigma^2}{A_V b_V \hat{u}_i (a_V + 1)} \left((r_V - \hat{u}_i (t^S - \Delta_i^T))^{a_V + 1} + (\hat{u}_i (t^E - \Delta_i^T) - r_V)^{a_V + 1} \right), & 0 \leq t^S \leq \frac{D_i^T}{2} \leq t^E \leq D_i^T \\ \frac{M\Upsilon_V \sigma^2}{A_V b_V \hat{u}_i (a_V + 1)} \left((\hat{u}_i (t^E - \Delta_i^T) - r_V)^{a_V + 1} - (\hat{u}_i (t^S - \Delta_i^T) - r_V)^{a_V + 1} \right), & \frac{D_i^T}{2} \leq t^S \leq t^E \leq D_i^T \end{cases} \quad (12)$$

B. Achievable Throughput Analysis

As this subsection is to derive the achievable throughput within T_{\max} , the issue of energy saving is not considered here. With the consideration of $R_I > R_V$, we let the target vehicle maintain the V2I link until it leaves the coverage of RSU A. Moreover, since all the V2V links can provide the same average transmission rate R_V , we transform the problem of deriving the maximum achievable throughput into the problem of deriving the maximum total duration of the V2V links established sequentially between the target vehicle and the relays.

As mentioned in Section III-A, we number the relays in the sequence of encounters with the target vehicle. In this context, we have $\Delta_1^T \leq \Delta_2^T \leq \dots \leq \Delta_{|\mathcal{V}_R|}^T$. We characterize a communication link by an interval whose two endpoints represent the start time and the end time of the link. For example, without considering the effect of other existing communication links, an available V2V link of relay v_i can be represented by $[\Delta_i^T, \Delta_i^T + D_i^T]$. However, for a specific cooperative transmission, the start time and end time of each V2V link become less straightforward. To avoid the interference between V2V links, relay v_i can only communicate to vehicle v_0 after the V2V link between vehicle v_0 and relay v_{i-1} is broken. This means that the start time of the V2V link established by vehicle v_i may be later than Δ_i^T . Meanwhile, relay v_i is responsible for forwarding the pre-stored SUs to vehicle v_0 . Thus, the maximum cumulatively transmitted data amount (denoted by \hat{C}_i^I) of the previously maintained V2I link restricts the maximum data amount transmitted over the V2V link. This requires a more elaborate calculation for the end time of the V2V link.

We denote the time when relay v_i starts to forward data to vehicle v_0 in a cooperative transmission achieving the maximum throughput by $\hat{\delta}_i$. Meanwhile, the set of the total V2V links cumulatively established by vehicle v_0 when it encounters relay v_i is denoted by \mathcal{D}_i . For example, for relay v_1 , the V2V link needs to be established after the target vehicle leaves the coverage area of RSU A. Therefore, we have

$$\hat{\delta}_1 = \max \{\Delta_1^T, 0 + D_0^I\}. \quad (13)$$

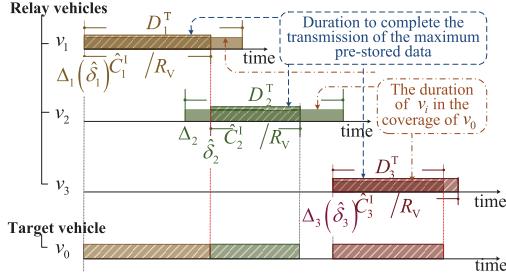


Fig. 4. Illustration of V2V-link duration.

Since the end time of a V2V link is restricted by two constraints: the maximum data amount pre-stored via the V2I link, and the time to leave the communication range of the target vehicle, the V2V link established by relay v_1 , i.e., \mathcal{D}_1 , is expressed by

$$\mathcal{D}_1 = \left[\hat{\delta}_1, \min \left(\hat{\delta}_1 + \hat{C}_1^I / R_V, \Delta_1^T + D_1^T \right) \right], \quad (14)$$

where $\hat{C}_1^I = R_I \cdot \min \left\{ D_1^I, \left[[0, \hat{\delta}_1] \right] \right\}$, since the pre-store process for the relay vehicle needs to be completed before forwarding data to the target vehicle. Moreover, as shown in Fig. 4, for subsequent relay v_i , $i \in \{2, \dots, |\mathcal{V}_R|\}$, the start time of the V2V link should be after the end time of the previous V2V link established by relay v_{i-1} . Therefore, the start time of the subsequent V2V links is expressed by

$$\hat{\delta}_i = \Delta_i^T + \left| \mathcal{D}_{i-1} \cap [\Delta_i^T, \Delta_i^T + D_i^T] \right|, \quad i \in \{2, \dots, |\mathcal{V}_R|\}, \quad (15)$$

where $|\cdot|$ represents the interval length, i.e., the link duration. Similar to the derivation of \mathcal{D}_1 , we have $\hat{C}_i^I = R_I \cdot \min \left\{ D_i^I, \left[[0, \hat{\delta}_i] \right] \right\}$. Then, the set of the total V2V links cumulatively established when vehicle v_0 encountering v_i , $i \in \{2, \dots, |\mathcal{V}_R|\}$, is represented by

$$\mathcal{D}_i = \left[\hat{\delta}_i, \min \left(\hat{\delta}_i + \hat{C}_i^I / R_V, \Delta_i^T + D_i^T \right) \right] \cup \mathcal{D}_{i-1}. \quad (16)$$

According to (16), we can obtain the maximum total duration of the V2V links, i.e., $\mathcal{D}_{|\mathcal{V}_R|}$. Furthermore, considering the delay requirement, the total duration of the V2V links is further modified to $\mathcal{D}_{|\mathcal{V}_R|} \cap [0, T_{\max}]$. Therefore, by jointly considering the direct V2I link between vehicle v_0 and RSU A, the achievable throughput within T_{\max} can be expressed by

$$Q_{\max} = R_I D_0^I + R_V \left| \mathcal{D}_{|\mathcal{V}_R|} \cap [0, d^{\max}] - [0, D_0^I] \right|. \quad (17)$$

C. Problem Formulation for Relay Selection & SU Assignment

According to Q_{\max} obtained in Section IV-B, the semantic encoder extracts an appropriate SR. We assume that the SR consists of N SUs, i.e., $j \in \{1, 2, \dots, N\}$. Considering that the direct link relies on the least predictive parameters and its start and end time is independent of other V2V links, the sudden change in vehicle speed have minimal impact on its transmission integrity. In this sense, the SUs with high importance need to prioritize the direct link, which ultimately

determine the amount of total data to be transmitted via forwarding links. Therefore, relay selection and SU allocation are two mutually coupled problems, which are jointly characterized by a $(|\mathcal{V}_R| + 1)$ -row and N -column matrix $\Phi = (\phi_{i,j} : i \in \{0, 1, 2, \dots, |\mathcal{V}_R|\}, j \in \{1, 2, \dots, N\})$. Herein, $\phi_{i,j}$ is a binary indicator, with $\phi_{i,j} = 1$ meaning that SU j is transmitted to vehicle v_i via the direct link or the relay link, and $\phi_{i,j} = 0$ otherwise. If $\sum_{j=1}^N \phi_{i,j} = 0$, it means that vehicle $v_i \in \mathcal{V}_R$ is not selected as a relay under Φ . Before defining the optimal strategy Φ^* , we first analyze the constraints that a feasible policy needs to satisfy as follows.

For a certain Φ , the start and the end time of each links are deterministic. As stated in Section IV-A, we assume that all the V2I links are established at the initial moment in the theoretical study of this paper. Thus, the end time of the V2I links are only determined by the SUs assigned to the each vehicle. As such, the end time of the V2I link of vehicle $v_i \in \mathcal{V}$ is expressed by $t_i^{E_i} = \frac{\sum_{j=1}^N \phi_{i,j} \beta_j}{R_i}$. Moreover, we denote the start time of the V2V link established by vehicle v_i by $t_i^{S_v}$, $(\Delta_i^T \leq t_i^{S_v} \leq \Delta_i^T + D_i^T)$. For vehicle v_1 , the start time of the V2V link should be after the end time of the V2I link between the target vehicle v_0 and RSU A. Thus, the expression of $t_1^{S_v}$ is shown as

$$t_1^{S_v} = \min \left\{ \max \{ \Delta_1^T, t_0^{E_i} \}, \Delta_1^T + D_1^T \right\}. \quad (18)$$

Similarly, all the subsequent V2V links should start after all their previous links are broken. We denote the end time of the V2V link established by vehicle $v_i \in \mathcal{V}_R$ by $t_i^{E_v}$, which is calculated by $t_i^{E_v} = t_i^{S_v} + \frac{\sum_{j=1}^N \phi_{i,j} \beta_j}{R_V}$. Therefore, the start time of the subsequent V2V links is expressed by

$$t_i^{S_v} = \min \left\{ \max \{ \Delta_i^T, t_{i-1}^{E_v} \}, \Delta_i^T + D_i^T \right\}, \quad \forall i \in \{2, \dots, |\mathcal{V}_R|\}. \quad (19)$$

After determining the start time, the maximum data amount can be transmitted a V2V link can be calculated by $\hat{C}_i^V = R_V (\Delta_i^T + D_i^T - t_i^{S_v})$, and the maximum pre-stored data amount via the V2I link can be calculated by $\hat{C}_i^I = R_I \cdot \min \left\{ D_i^I, \left[[0, t_i^{S_v}] \right] \right\}$. Since the total data amount of the SUs assigned is bounded by both the transmission capacity of the V2V link and V2I link, to ensure the integrity of the transmission, we have

$$\sum_{j=1}^N \phi_{i,j} \beta_j \leq \min \left\{ \hat{C}_i^I, \hat{C}_i^V \right\}, \quad \forall i \in \{1, \dots, |\mathcal{V}_R|\}. \quad (20)$$

For the same reason, the transmission of the SUs assigned to vehicle v_0 is required to be completed within the coverage of RSU A. Thus, we have

$$\sum_{j=1}^N \phi_{0,j} \beta_j \leq \hat{C}_0^I. \quad (21)$$

Additionally, considering the delay requirement, the last V2V link should end at a time earlier than the maximum acceptable delay threshold. Therefore, we have

$$\max \{ t_i^{E_v} \} \leq T_{\max}, \quad \forall i \in \{1, \dots, |\mathcal{V}_R|\}. \quad (22)$$

To evaluate feasible strategies that satisfy the above constraints, we consider two main aspects. One is the energy

consumption. According to (11) and (12), for any feasible strategy Φ , the total energy consumption of the V2I links and V2V links can be calculated by $P_{V2I} = \sum_{i=0}^{|V_R|} f_i^{V2I}(0, t_i^{E_i})$ and $P_{V2V} = \sum_{i=1}^{|V_R|} f_i^{V2V}(t_i^{SV}, t_i^{EV})$, respectively. The other is semantic reliability. Considering the possibility of sudden changes in vehicle speed, the selected SR might fail to be fully transmitted as planned. Therefore, the SUs with high importance can assign to the more reliable direct link to reduce the impact of vehicle network uncertainty as discussed at the beginning of this subsection. With this in mind, we introduce two parameters θ_T and θ_R to qualitatively characterize the reliability of the direct and forward transmission, respectively, where $\theta_T > \theta_R$. Furthermore, we define a new metric to quantify the semantic reliability of Φ based on semantic significance α_j of each SU j , which is expressed by

$$\Theta = \theta_T \sum_{j=1}^N \phi_{0,j} \alpha_j + \theta_R \sum_{i=1}^{|V_R|} \sum_{j=1}^N \phi_{i,j} \alpha_j. \quad (23)$$

In summary, the relay selection and SU assignment can be jointly formulated as a combinatorial optimization problem,

$$\min_{\Phi} \kappa_1 (P_{V2I} + P_{V2V}) - \kappa_2 \Theta, \quad (P1)$$

subject to

$$\phi_{i,j} \in \{0, 1\}, \quad \forall i \in \{0, \dots, |V_R|\}, \quad \forall j \in \{1, \dots, N\}, \quad (a)$$

$$\sum_{i=0}^{|V_R|} \phi_{i,j} = 1, \quad \forall j \in \{1, 2, \dots, N\}, \quad (b)$$

Constraints: (20) (21) (22),

where κ_1 and κ_2 are two parameters used to weigh energy consumption and semantic transmission reliability. Moreover, the constraints in (a) and (b) ensure that each SU is assigned only once. The specific solution to (P1) is provided in Section IV-D.

D. Markov Approximation and Solution

Given the multiple $\max \{\cdot\}$ and $\min \{\cdot\}$ in (P1), the explicit expression for its feasible region is challenging to derive. Also, due to the high dimensionality of Φ , the conventional numerical analysis methods and centralized search algorithms become inefficient, especially for finding a favourable solution within limited time.

To this end, we propose a Markov-chain-guided multi-thread search algorithm (M-MTSA) as shown in Fig. 5. Inspired by Markov approximation [40], we first approximate (P1) by transforming it into a continuous convex optimization problem (P2) in the probability domain based on Log-Sum-Exp approximation and the conjugate function property. The decision variables in (P2) are the probability weights corresponding to all possible Φ . Ideally, the probability corresponding to the optimal strategy Φ^* is remarkably close to one. Then, we construct a Markov chain with the state space as all possible Φ and the stationary distribution as the optimal solution of (P2). During its execution, according to constraint (b), M-MTSA maintains N threads for all the SUs, respectively. The partial strategy for thread j serves as the j th column of Φ , a one-hot vector, which determines the selected relay for SU j . According to the transition rates of the Markov

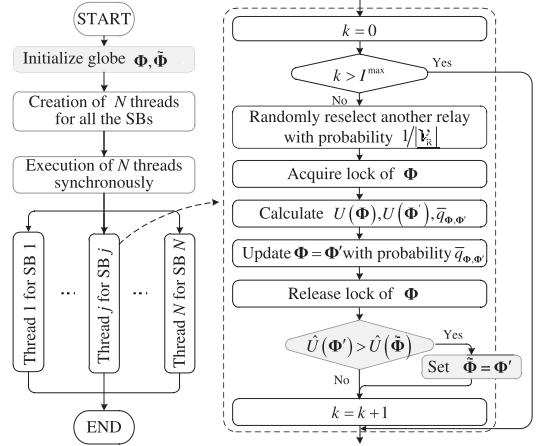


Fig. 5. The flowchart of M-MTSA.

chain, the individual relays for the SUs are constantly and distributively updated with small inter-thread message passing overhead. By the careful design of transition rates, the Markov chain jumps to better strategies over time. Next, we detail the problem transformation, Markov chain construction, and M-MTSA design in Sections IV-D.1-IV-D.3, respectively.

1) *Problem Transformation*: Recall the problem in (P1), for ease of presentation, we denote the objective function by $U(\Phi)$. Then, (P1) can be rewritten as

$$\min_{\Phi \in \mathcal{F}^*} U(\Phi), \quad (24)$$

where \mathcal{F}^* represents the feasible region. Since \mathcal{F}^* is unavailable, we transform constrained problem to unconstrained one by adding a penalty term to the objective function. Then, (P1) can be rewritten as

$$\min_{\Phi \in \mathcal{F}} U(\Phi) + \Omega \cdot \mathbf{1}_{\mathcal{C}_{\mathcal{F}^*}}(\Phi). \quad (P2)$$

In (P2), \mathcal{F} is the set of all the possible Φ satisfying constraints (a) and (b) with $|\mathcal{F}| = (|V_R| + 1)^N$. Moreover, Ω is a constant penalty factor which is significantly larger than $U(\Phi)$, and $\mathbf{1}_{\mathcal{C}_{\mathcal{F}^*}}(\Phi)$ is an indicator function defined as

$$\mathbf{1}_{\mathcal{C}_{\mathcal{F}^*}}(\Phi) = \begin{cases} 1, & \Phi \in \mathcal{C}_{\mathcal{F}^*} \\ 0, & \text{otherwise.} \end{cases} \quad (25)$$

For brevity, we rewrite the objective function in (P2) as $\hat{U}(\Phi)$. To enable the analysis from the probability domain, the log-sum-exp function is used to approximate $\min_{\Phi \in \mathcal{F}} \hat{U}(\Phi)$, i.e.,

$$\min_{\Phi \in \mathcal{F}} \hat{U}(\Phi) \approx g_{\varpi}(\hat{\mathcal{U}}) = -\varpi \log \left(\sum_{\Phi \in \mathcal{F}} \exp \left(-\frac{\hat{U}(\Phi)}{\varpi} \right) \right), \quad (26)$$

with the upper bound of $\varpi \log |\hat{\mathcal{U}}|$ for approximation gap.

Proposition 1: When $\varpi \rightarrow 0^+$, for a set \mathcal{X} of n nonnegative real variables $x_1, x_2, x_3, \dots, x_n$, we have

$$\min_{i=1,2,\dots,n} x_i - \varpi \log |\mathcal{X}| \leq g_{\varpi}(\mathcal{X}) \leq \min_{i=1,2,\dots,n} x_i. \quad (27)$$

Proof: We rearrange x_i so that they are ranked as $x_1 \leq x_2 \leq \dots \leq x_n$. Then, we have

$$\begin{aligned} g_{\varpi}(\mathcal{X}) &= -\varpi \log \left(\sum_{i=1}^n \exp \left(-\frac{x_i}{\varpi} \right) \right) \\ &= -\varpi \log \left(\exp \left(-\frac{x_1}{\varpi} \right) \exp \left(\frac{x_1}{\varpi} \right) \right. \\ &\quad \times \left. \sum_{i=2}^n \exp \left(-\frac{x_i}{\varpi} \right) \right) \\ &= x_1 - \varpi \log \left(1 + \sum_{i=2}^n \exp \left(\frac{x_1 - x_i}{\varpi} \right) \right). \quad (28) \end{aligned}$$

Therefore, the approximation gap can be expressed by

$$|g_{\varpi}(\mathcal{X}) - x_1| = \left| \varpi \log \left(1 + \sum_{i=2}^n \exp \left(\frac{x_1 - x_i}{\varpi} \right) \right) \right|. \quad (29)$$

When $x_1 = x_2 = \dots = x_n$, $|g_{\varpi}(\mathcal{X}) - x_1| = \varpi \log |\mathcal{F}|$; when $x_1 \ll x_2 \leq \dots \leq x_n$, $|g_{\varpi}(\mathcal{X}) - x_1| \rightarrow 0$. \square

Then, since $g_{\varpi}(\hat{\mathcal{U}})$ is a convex and closed function, the conjugate of its conjugate is itself, i.e., $g_{\varpi}(\hat{\mathcal{U}}) = g_{\varpi}^{**}(\hat{\mathcal{U}})$. According to the definition of conjugate function⁶, the conjugate of $g_{\varpi}(\hat{\mathcal{U}})$ can be expressed by [41, p.93]

$$g_{\varpi}^*(\mathbf{p}) = \begin{cases} -\varpi \sum_{\Phi \in \mathcal{F}} p_{\Phi} \log p_{\Phi}, & \text{if } \mathbf{p} \geq 0 \text{ and } \mathbf{1}^T \mathbf{p} = 1; \\ \infty, & \text{otherwise.} \end{cases} \quad (30)$$

Similarly, the conjugate of $g_{\varpi}^*(\mathbf{p})$, i.e., $g_{\varpi}^{**}(\hat{\mathcal{U}})$, can be obtained by solving the following problem [40].

$$\begin{aligned} &\max_{\mathbf{p} \geq 0} \sum_{\Phi \in \mathcal{F}} p_{\Phi} \hat{U}(\Phi) + \varpi \sum_{\Phi \in \mathcal{F}} p_{\Phi} \log p_{\Phi}, \\ &\text{s.t. } \sum_{\Phi \in \mathcal{F}} p_{\Phi} = 1. \end{aligned} \quad (\text{P3})$$

Therefore, the optimal value of (P3) is the same as $g_{\varpi}(\hat{\mathcal{U}})$. According to Proposition 1, it approximates the optimal value of (P2) with a gap bounded by $\varpi \log |\mathcal{F}|$, from the analysis of (P3), which is caused by the term $\varpi \sum_{\Phi \in \mathcal{F}} p_{\Phi} \log p_{\Phi}$. By addressing the Karush-Kuhn-Tucker conditions [41], the closed-form of the optimal solution to (P3) is shown as below:

$$p_{\Phi}^* = \frac{\exp \left(-\frac{\hat{U}(\Phi)}{\varpi} \right)}{\sum_{\Phi' \in \mathcal{F}} \exp \left(-\frac{\hat{U}(\Phi')}{\varpi} \right)}, \quad \forall \Phi \in \mathcal{F}. \quad (31)$$

As such, an average performance that is close to the optimal value of (P2) can be achieved via time-sharing of all the possible Φ according to individual p_{Φ}^* . Obviously, according to (31), Φ^* occupies the longest proportion of time. The point to note here is that Φ^* is what we try to find in our work, instead of the average performance itself.

2) *Markov Chain Construction:* To proceed, Markov approximation implements a well-designed Markov chain with the state space of \mathcal{F} to gradually converge to the stationary distribution shown in (31). For any stationary distribution in product form, there exists at least one continuous time-reversible ergodic Markov chain [40, Lemma 1]. Specifically, the transition rates need to meet the following two conditions:

⁶Let $g: \mathbb{R}^n \rightarrow \mathbb{R}$. The conjugate function of g is defined as $g^*(y) = \sup_{\mathbf{x} \in \text{dom}g} (y^T \mathbf{x} - g(\mathbf{x}))$ [41].

- the resulting Markov chain is irreducible, i.e., any two states are reachable from each other;
- The detailed balance equation is satisfied, i.e., $\forall \Phi, \Phi' \in \mathcal{F}$, $p_{\Phi}^* q_{\Phi, \Phi'} = p_{\Phi'}^* q_{\Phi', \Phi}$,

where $q_{\Phi, \Phi'}$ be the transition rate from state Φ to Φ' . For faster convergence and easier capturing of Φ^* , the Markov chain should be more likely to jump to the state with better performance. As such, the transition rates should depend on both $\hat{U}(\Phi)$ for the current state and $\hat{U}(\Phi')$ for the target state. With above in mind, the transition rate is designed as below:

$$q_{\Phi, \Phi'} = \frac{\alpha \exp \left(-\frac{\hat{U}(\Phi')}{\varpi} \right)}{\max \left\{ \exp \left(-\frac{\hat{U}(\Phi)}{\varpi} \right), \exp \left(-\frac{\hat{U}(\Phi')}{\varpi} \right) \right\}}, \quad (32)$$

where α is a positive constant which determines the convergence time of Markov chain. According to (32), if $\hat{U}(\Phi') > \hat{U}(\Phi)$, the state is updated with maximum transition rate of α . Otherwise, the larger difference between $\hat{U}(\Phi')$ and $\hat{U}(\Phi)$, the smaller the $q_{\Phi, \Phi'}$. Moreover, the value of $\hat{U}(\Phi)$, $\Phi \in \mathcal{F}$, determines difference of the stationary distribution among the states, thus affecting the convergence time. Specifically, the convergence time of the designed Markov chain is bounded as follows⁷:

$$\text{for } \varpi \geq 2 \left(\hat{U}_{\max} - \hat{U}_{\min} \right) \left(\ln \left(N + \frac{1}{|\mathcal{V}_R|} \right) / N - 1 \right)^{-1},$$

$$t_{\text{mix}}(\epsilon) \geq \frac{1}{2\alpha M |\mathcal{V}_R|} \ln \frac{1}{\epsilon}, \quad (33)$$

$$t_{\text{mix}}(\epsilon) \leq \frac{\frac{1}{\alpha |\mathcal{V}_R|} \cdot \exp \left(\frac{1}{\varpi} \left(2\hat{U}_{\max} - \hat{U}_{\min} \right) \right) \ln \frac{N}{\epsilon}}{N + \frac{1}{|\mathcal{V}_R|} - (N-1) \exp \left(\frac{2}{\varpi} \left(\hat{U}_{\max} - \hat{U}_{\min} \right) \right)}, \quad (34)$$

where ϵ is the parameter to judge convergence, and \hat{U}_{\max} and \hat{U}_{\min} represent the maximum and minimum values of $\hat{U}(\Phi)$. According to (33) and (34), we can observe that the larger the value of α , the smaller the upper bound on the convergence time of the Markov chain. The value of α in our work is related to numbers of vehicles and SUs, which is specified in Section IV-D.3. Moreover, the differences in the value of $\hat{U}(\Phi)$ corresponding to different $\Phi \in \mathcal{F}$ and value of ϖ also affect the convergence time.

3) *M-MTSA Design:* M-MTSA is designed as shown in Fig. 5. M-MTSA is required to perform two functions. The one is to implement the designed Markov chain in a distributed manner. The other one is to track the best solution during the Markov chain hopping process. For clarity, we use Φ , Φ' , and $\tilde{\Phi}$ to represent the current state, the next state, and the current best strategy, respectively. It should be clarified that due to the stochastic nature of the mixed time of Markov chains, M-MTSA cannot ensure that the optimal result is obtained within I iterations. In this sense, when the algorithm ends, we treat $\tilde{\Phi}$ as Φ^* approximately. With the aim to find a favorable solution within limited time, we transform the continuous-time

⁷The lower bound and upper bound are obtained based on spectral analysis and path coupling method, respectively. Due to space limitation, the proof process is omitted here. A similar process can be found in [40, Theorem 5].

Algorithm 1 SU Assignment Algorithm Based on [4] (Baseline)

Input : $R_V, R_I, D_0^I, D_i^I, \Delta_i^T, D_i^T, i \in \{1, 2, \dots, |\mathcal{V}_R|\}$

Output: Φ

- 1 Initialize $\Phi = 0$.
- 2 Set $\Phi_{0,j} = 1, \forall j \in \{1, 2, \dots, N\}$ /* Assign all the SUs to the target vehicle */
- 3 **for** $i = 0 : |\mathcal{V}_R| - 1$ **do**
- 4 Check if constraint (20) or (21) is satisfied.
- 5 **if** the related constraint is not satisfied **then**
- 6 **do**
- 7 Select an SU $j^* = \arg \min_{j \in \mathcal{S}_i} \{\beta_j\}$, where $\mathcal{S}_i = \{j \mid \Phi_{i,j} = 1, \forall j \in \{1, 2, \dots, N\}\}$
- 8 Set $\Phi_{i,j^*} = 0, \Phi_{i+1,j^*} = 1$
- 9 Calculate the remaining data amount that cannot be transmitted
- 10 $\delta\Lambda = \sum_{j=1}^N \phi_{i,j} \beta_j - \hat{C}_i^{V2I}, i = 0$
- 11 $\delta\Lambda = \sum_{j=1}^N \phi_{i,j} \beta_j - \min \left\{ \hat{C}_i^I, R_V \left(\Delta_i^T + D_i^T - \hat{\delta}_i \right) \right\}, i \in \{1, \dots, |\mathcal{V}_R|\}$
- 12 **while** $\delta\Lambda \leq 0$ **do**
- 13 **if** $i == 0$ **then**
- 14 $\hat{\delta}_1 = \min \left\{ \max \left\{ \Delta_1^T, \frac{\sum_{j=1}^N \phi_{0,j} \beta_j}{R_I} \right\}, \Delta_1^T + D_1^T \right\}$
- 15 **else**
- 16 $\hat{\delta}_{i+1} = \min \left\{ \max \left\{ \Delta_{i+1}^T, \hat{\delta}_i + \frac{\sum_{j=1}^N \phi_{i,j} \beta_j}{R_V} \right\}, \Delta_{i+1}^T + D_{i+1}^T \right\}$
- 17 **end**
- 18 **end**
- 19 **end**

channel-hopping Markov chain to a discrete-time Markov chain via uniformization [42]. Specifically, all the threads randomly reselect another relay for their individual SUs with the probability of $\frac{1}{|\mathcal{V}_R|}$ in parallel. Then, one of the threads acquires the lock of Φ , and calculates $\hat{U}(\Phi)$ and $\hat{U}(\Phi')$. The state jumps from Φ to Φ' with the probability of $\bar{q}_{\Phi, \Phi'} = \exp \left(-\frac{\hat{U}(\Phi')}{\varpi} \right) / \max \left\{ \exp \left(-\frac{\hat{U}(\Phi)}{\varpi} \right), \exp \left(-\frac{U(\Phi')}{\varpi} \right) \right\}$. With the assumption that each thread has an equal probability of obtaining the lock of Φ , the transition probability from Φ to Φ' can be specified as $q_{\Phi, \Phi'} = \frac{1}{N|\mathcal{V}_R|} \bar{q}_{\Phi, \Phi'}$, which is consistent with the form of (32), i.e., $\alpha = \frac{1}{N|\mathcal{V}_R|}$. Meanwhile, if $\hat{U}(\Phi') > \hat{U}(\tilde{\Phi})$, the thread updates global $\tilde{\Phi}$ to Φ' . Assume that the optimal solution can be found after I iterations. The complexity of M-MTSA is $\mathcal{O}(IN)$. Compared to the centralized search algorithm with the complexity of $\mathcal{O}((|\mathcal{V}_R| + 1)^N)$, the complexity is greatly reduced.

Moreover, we devise an SU assignment algorithm as the baseline following the idea of the elastic-segment-based V2V/V2I cooperative strategy [4], where the relays are selected in the order of encounter with the target vehicle until the requested data transmission is completed. Moreover, we treat the strategy generated by this algorithm as the initial feasible state of M-MTSA for easier capture of the optimal solution. Next, we present the details about the SU assignment algorithm, which is outlined in Algorithm 1. Considering the preference for V2I links, at the beginning, all the SUs are assigned to the target vehicle v_0 . Then, Algorithm 1 pre-checks whether the transmission of the assigned SUs can be completed, (i.e., constraint (21)). If not, the excessive SUs are moved to be transmitted by the next relay vehicle to encounter. To mitigate the idleness of V2I link caused by

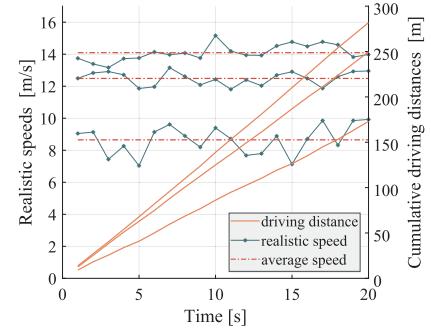


Fig. 6. Realistic trajectory and speed generated with SUMO.

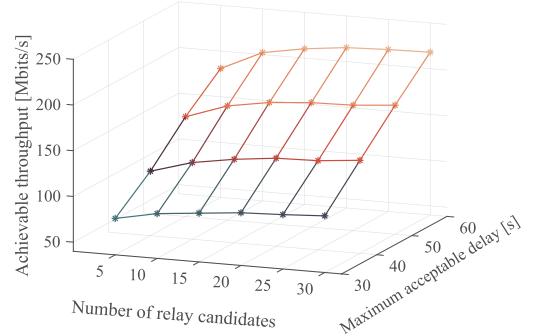


Fig. 7. Achievable throughput with different T_{\max} and $|\mathcal{V}_R|$.

the non-divisibility of SU, the SUs with small data volume are moved in priority. The detail of the process is outlined in Lines 6–12. Then, Algorithm 1 calculates the transmission start moment after the encounter with the next relay vehicle, which is shown in Lines 13–17. After that, Algorithm 1 checks if constraint (20) for the relay vehicle is satisfied. Then, Algorithm 1 repeats the above process.

V. SIMULATION

A. Simulation Setup

In the simulation, we focus on a segment of a road with two RSUs. The parameters related to the communication scenarios are summarized in Table II. In our system, the small-scale fading occurring in each transmission slot, with a duration of 1 ms, is generated by utilizing realizations of the square of the (random) small-scale channel coefficients according to [38, Eq.(1)]. The initial positions of both the target vehicle and the relay vehicles are randomly generated with a uniform distribution within $(-r_1, r_1)$. The vehicle trajectories are generated with SUMO, where the average routing speed and the traffic density are set 13.89 m/s and 10 vehicle/km per lane. Moreover, vehicles can be distinguished according to the setting of attribute parameters, such as acceleration, deceleration, sigma, and maximum speed. Three representative trajectories are shown in Fig. 6, where we can see that although the speed varies noticeably on the small time scale, the distance driven cumulatively from a large time scale is close to the distance driven with its average speed. This validates the rationality of analysis based on historical average speed in the proposed PreCMTS in an intuitive way.

TABLE II
MAIN SIMULATION PARAMETERS

Parameters	Settings	Parameters	Settings	Parameters	Settings
RSU coverage radius	$r_1 = 500$ m [4]	Vehicle coverage radius	$r_v = 300$ m [4]	Distance between two RSUs	$H = 1500$ m [4]
V2I channel model	$b_1 = 1$ $a_1 = 2.2$ [34]	V2V channel model	$b_{v2v} = 1$ $a_{v2v} = 2$ [34]	Average small-scale channel gain	$\bar{g} = 1$ dB
Fading severity	$m = 6$ [38]	Shadowing shape	$m_s = 6$ [38]	Link bandwidth	$B = 1$ MHz
Noise	$\sigma^2 = -110$ dBm/Hz	SNR threshold for RSU	$\Gamma_1 = 15.27$ dB	SNR threshold for vehicle	$\Gamma_v = 11.44$ dB
Joint antenna gain	$G_1 = G_v = 1$ [43]	Reliability for direct transmission	$\theta_T = 1.5$	Reliability for relay transmission	$\theta_R = 0.5$

TABLE III
SET OF SUs

SR1	SU	a	b	c	d	e	f	g	h	i	j	k	l	m	n
	β	11	14	15	24	3	20	23	4	1	22	7	9	4	8
α	0.86	1.08	1.17	1.87	0.23	1.56	1.79	0.31	0.08	1.71	0.54	0.70	0.31	0.62	
	SU	a	b	c	d	e	f	g	h	i	j	k	l	m	n
SR2	β	11	14	15	24	3	20	23	4	1	22	7	9	4	8
	α	0.86	1.08	1.17	1.87	0.23	1.56	1.79	0.31	0.08	1.71	0.54	0.70	0.31	0.62
SR2	SU	o	p	q	r	s									
	β	17	12	5	1	9									
α	0.06	0.29	0.62	0.10	0.59										

TABLE IV
THE FOUR RESULTS OF BASELINE AND PRECMTS WITH DIFFERENT T_{\max}

Baseline	PreCMTS ($T_{\max} = 40$ s)				PreCMTS ($T_{\max} = 50$ s)				PreCMTS ($T_{\max} = 60$ s)												
	Veh.	SU	Veh.	SU	Veh.	SU	Veh.	SU	Veh.	SU	Veh.	SU	Veh.	SU							
v_0^*	a, b, c, d, f, g, j, l, 1	v_0	b, c, d, g, j, l, n	v_9	f	v_0	b, c, d, j, l, n	v_1	h, k	v_8	g	v_0	b, c, d, l, n	v_1	h	v_9	f	v_{15}	a, j	v_{19}	k
v_1	e, h, i, k, m, n	v_1	e, h, i, k	v_{12}	m			v_3	e, i	v_{12}	m	v_6	e	v_{12}	m	v_7	i	v_{13}	g		

* The shaded cells indicate the target vehicle, the other vehicle indexes indicate the selected relay vehicles in each strategy.

B. Performance Evaluation

The achievable throughput within T_{\max} not only determines the maximum data volume of the SR that can be supported for the current scenario, but also the optimizable space of PreCMTS given a selected SR. In this sense, before the evaluation of the proposed PreCMTS, we first show the average achievable throughput of 50 simulations with randomly generated initial positions under different maximum acceptable delay T_{\max} and different numbers of relay candidates $|\mathcal{V}_R|$ in Fig. 7. From Fig. 7, we can observe that as the number of relay candidates increases and the delay requirement is relaxed, there is an increase in the achievable throughput. Specifically, with the increase of the maximum acceptable delay, the rise in the achievable throughput, as the number of vehicles in RSU B increases, is more significant. This is because the vehicles are randomly distributed within RSU B. When the value of T_{\max} is small, a high percentage of vehicles encounter the target vehicle exceed the maximum acceptable delay, which fails to contribute to the throughput. Moreover, since only one V2V link can exist at any moment, when the number of vehicle candidates reaches a certain threshold, the increase in throughput becomes flat. Furthermore, the threshold value of the number of relay candidates decreases with the increase of the maximum acceptable delay.

Next, we take an example with 20 relay candidates to evaluate the performance of the PreCMTS with $\kappa_1 = 0.5$, $\kappa_2 = 0.1$. To demonstrate its superiority, we evaluate the PreCMTS under different maximum acceptable delay, i.e., $T_{\max} = 40$ s, $T_{\max} = 50$ s, and $T_{\max} = 60$ s with the

baseline devised according to [4]. Based on (17), we have $Q_{\max} = 186.6$ Mbits within 40 s. Accordingly, with out loss of generality, we generate a SR indexed by SR 1 as Table III with the total volume of 165 Mbits and semantic accuracy of 12.84 to perform the simulation. The four strategies derived in the above four cases are presented in Table. IV, where the SUs within the SR 1 that each vehicle should transmit are determined. A more intuitive presentation is in Fig. 8. Moreover, the practical trajectory information of all the relay candidates are shown in Table. V and Fig. 10. At last, the performance of PreCMTS in terms of energy efficient and semantic reliability are shown in Fig. 9, respectively.

Overall, as shown in Fig. 9(a), the simulated cumulative energy consumption basically coincides the theoretical estimated value, which supports the rationality of performance evaluation in terms of energy consumption. Moreover, it is evident that the overall energy consumption experiences a significant reduction when the delay requirements are relaxed following the optimization of the proposed PreCMTS. Specifically, by comparing Fig. 8 and Table V, we can observe that in the baseline, the target vehicle consistently maintains the V2I link within the coverage area of RSU A. Owing to $R_1 > R_v$, the baseline achieves the lowest transmission delay. However, due to the low channel gain at the edge of coverage, such a V2I link-first mechanism would cause significant energy waste. In our PreCMTS, by optimizing the relay selection and SU assignment within T_{\max} , it selectively assigns partial SUs to the store-carry-forward links. In this way, the energy efficiency in RSU A can remarkably increase, which can be

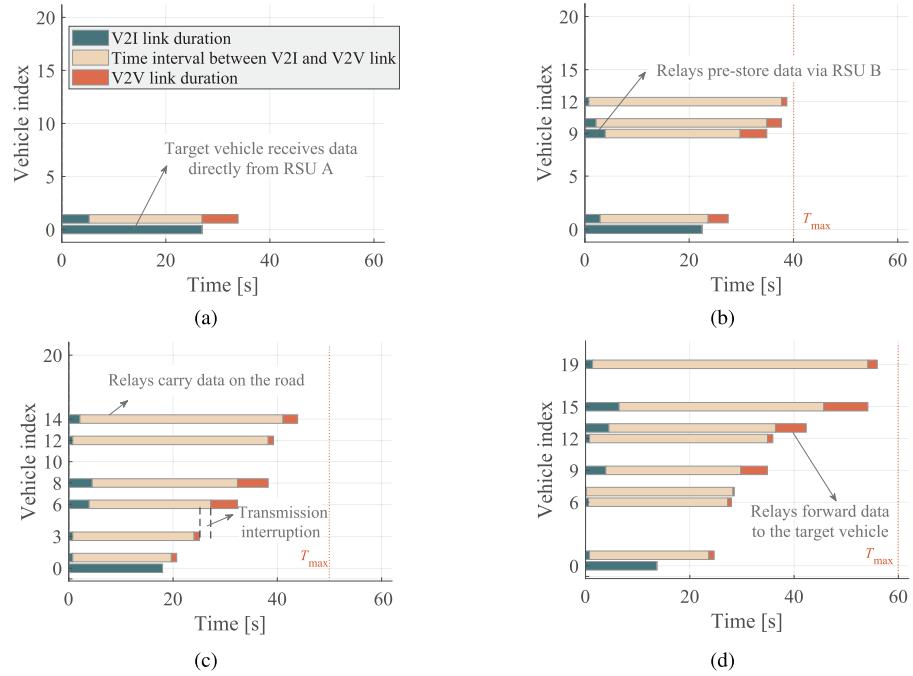


Fig. 8. Comparison of the four strategies. (a) Baseline; (b) PreCMTS with $T_{\max} = 40$ s; (c) PreCMTS with $T_{\max} = 50$ s; (d) PreCMTS with $T_{\max} = 60$ s.

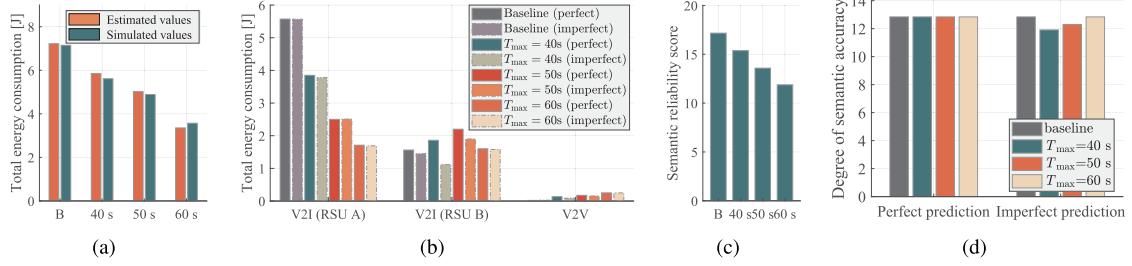


Fig. 9. Performance analysis. (a) Comparison of theoretical and simulated values of cumulative energy consumption. (b) Comparison of energy consumption of the strategies. (c) Comparison of semantic reliability scores; (d) Comparison of degree of semantic accuracy.

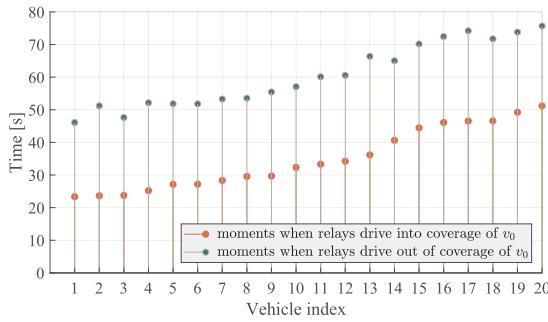


Fig. 10. Encounter time between relay and target vehicle.

verified by calculating the ratio of the V2I link duration of v_0 (or the data volume assigned to v_0) and the energy consumption of V2I link in RSU A in the four strategies according to Figs. 8 and 9(b). Moreover, from Fig. 8(b)-(d), with the relaxation of time delay requirements, the relays closer to RSU B, such as v_{14} and v_{15} are more likely to be selected under different delay requirements to enhance energy efficiency. This also allows the fact shown in Fig. 9(b) that the total energy consumption of the V2I links in RSU B does not increase monotonically with the data volume transmitted. Meanwhile,

TABLE V
INITIAL LOCATIONS AND AVERAGE SPEEDS OF VEHICLES

Veh.	v_0	v_1	v_2	v_3	v_4	v_5	v_6
\bar{l}_i	200	382	484	403	438	340	336
\bar{u}_i	10.97	15.44	10.81	14.14	11.28	13.31	13.41
Veh.	v_7	v_8	v_9	v_{10}	v_{11}	v_{12}	v_{13}
\bar{l}_i	317	260	308	214	253	220	281
\bar{u}_i	13.10	14.03	12.30	13.30	11.41	11.81	8.89
Veh.	v_{14}	v_{15}	v_{16}	v_{17}	v_{18}	v_{19}	v_{20}
\bar{l}_i	0.12	-39	-50	-10	-112	-202	-254
\bar{u}_i	13.63	12.38	11.80	10.72	12.90	13.45	13.53

the vehicles far from the RSU B such as v_2 and v_4 are missed in the all the strategies, even if this leads to an avoidable transmission interruption implied by Fig. 10. This is the key reason why the PreCMTS can achieve higher energy efficiency compared with the existing schemes. Moreover, comparing Figs. 8 and 10, most V2V links are established after a period of time when the relay encounters the target vehicle. This means that the transmission distance of V2V links is generally shorter and thus more energy can be saved. Therefore, the energy consumption of V2V links is significantly smaller than that of the V2I links. Moreover, as shown in Fig. 9(c), as more SUs are assigned to the relays, the semantic reliability score becomes smaller, which is consistent with the definition

TABLE VI
THE THREE RESULTS OF PRECMTS UNDER DIFFERENT SRs AND T_{\max}

PreCMTS ($\kappa_1 = 0.1$; $\kappa_2 = 0.5$)				PreCMTS ($T_{\max} = 50$ s, SR 2)							
Veh.	SU	Veh.	SU	Veh.	SU	Veh.	SU	Veh.	SU	Veh.	SU
v_0^*	b, c, d, f, g, j, l, n	v_{12}	m	v_0	b, c, d, f, g, j	v_3	m, n, q	v_4	e, h, r	v_{14}	o
v_1	e, h, i, k	v_{14}	a	v_1	p, s	v_9	k	v_{10}	l	v_{16}	a
Baseline (SR2)											
Veh.	SU	Veh.	SU	Veh.	SU	Veh.	SU	Veh.	SU	Veh.	SU
v_0	b, c, d, f, g, j, o	v_3	k, l, m, n, q	v_0	b, c, d, f	v_3	m, n	v_{11}	i	v_{16}	a, e, j
v_1	a, p, s	v_4	e, h, i, r	v_1	p, s	v_9	k, l	v_{13}	o	v_{14}	g
PreCMTS ($T_{\max} = 60$ s, SR 2)											
Veh.	SU	Veh.	SU	Veh.	SU	Veh.	SU	Veh.	SU	Veh.	SU
v_0	b, c, d, f, g, j, o	v_3	k, l, m, n, q	v_0	b, c, d, f	v_3	m, n	v_{11}	i	v_{16}	a, e, j
v_1	a, p, s	v_4	e, h, i, r	v_1	p, s	v_9	k, l	v_{14}	g	v_{20}	q

* The shaded cells indicate the target vehicle, the other vehicle indexes indicate the selected relay vehicles in each strategy.

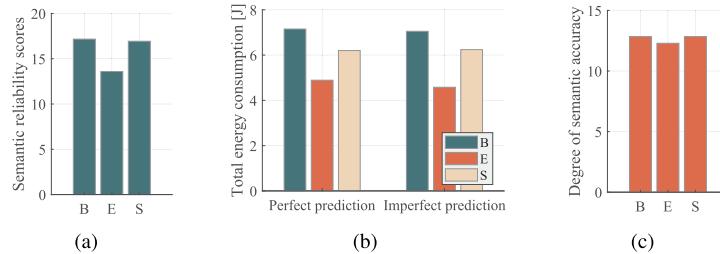


Fig. 11. Comparison of PreCMTS with different κ_1 and κ_2 , where “B” represents the baseline, “E” represents the PreCMTS with $\kappa_1 = 0.5$ and $\kappa_2 = 0.1$, and “S” represents the PreCMTS with $\kappa_1 = 0.1$ and $\kappa_2 = 0.5$. (a) Semantic reliability scores; (b) Energy consumption; (c) Degree of semantic accuracy with imperfect speed prediction.

in (23). However, since we minimize energy consumption while optimizing semantic reliability, SUs with less semantic importance are preferentially assigned to other relays, which can be seen by comparing Tables III and IV. Since the SUs with large semantic importance are mostly assigned to the direct link, the transmission of them is completed with priority. Therefore, in some cases with imperfect speed prediction, there is no remarkable decrease in the degree of the semantic accuracy as shown in Fig. 9(d). Moreover, in some cases, the SR can still be fully transmission with imperfect speed prediction, but it consumes more energy. This shows the adaptability of the proposed strategy to sudden changes in the vehicular environment.

In addition, by adjusting the values of κ_1 and κ_2 , the attention of PreCMTS to the energy consumption and semantic reliability of SR transmission can be adjusted. In Fig. 11, we compare two PreCMTS under the settings of $\kappa_1 = 0.5$, $\kappa_2 = 0.1$, and $\kappa_1 = 0.1$, $\kappa_2 = 0.5$, respectively. The two specific strategy results can be found in Table. IV and Table VI. As shown in Fig. 11(a), with the increase of the value of κ_2 , the semantic reliability score is improved significantly, and the energy consumption is increased slightly. Moreover, as the SU with high semantic importance such SU g and SU f are assigned to v_0 , the degree of semantic accuracy achieved by PreCMTS S is higher than that achieved by PreCMTS E, as shown in Fig. 11(c). In addition, it is to be noted that the above simulation only supports that the values of κ_1 and κ_2 can effectively influence the strategy results. However, as there is no definite linear relationship between the semantic importance of an SU and its data volume, the optimal combination of values for κ_1 and κ_2 deserves further investigation.

Furthermore, we evaluate the performance of PreCMTS under different SRs. According to (17), when the maximum

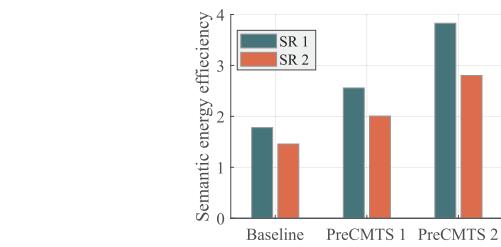


Fig. 12. Comparison of the PreCMTS with SR 1 and SR 2, where PreCMTS 1 represents the PreCMTS with $T_{\max} = 50$ s, and PreCMTS 2 represents the PreCMTS with $T_{\max} = 60$ s.

acceptable delay extends to 50 s and 60 s, $Q_{\max} = 225.6$ Mbits and $Q_{\max} = 264.6$ Mbits, respectively. We generate a second SR indexed by SR 2 as shown in Table. III. Considering that the SUs with small semantic contribution is filtered out in priority, the SUs with relative small value of α_j are added in SR 2. Moreover, for a intuitive performance comparison, we define a new metric called semantic energy efficiency as the ratio of the degree of semantic accuracy and total energy consumption, i.e., $EE_S = \sum_{j=1}^N \alpha_j / (P_{V2V} + P_{V2I})$. As shown in Fig. 12, as the added SUs is with relatively small semantic importance and random data volume, the semantic energy efficiency of the strategies with SR 2 is clearly lower than that achieved by the corresponding strategies with SR 1. Meanwhile, with the extension of the acceptable delay, the semantic energy efficiency increases remarkably for both SR 1 and SR 2. Specifically, the semantic energy efficiency of the PreCMTS with SR 1 has a greater enhancement than that with SR 2. This means that the PreCMTS achieves superior performance at lower system loads. Thus, a trade-off between the energy efficiency and semantic accuracy degree can be made on a case-by-case basis.

VI. CONCLUSION

In this paper, we have proposed a predictive cooperative multi-relay transmission strategy for bidirectional road scenarios. Specifically, we have introduced a general task-driven KG-assisted SemCom system for complex vehicular network. To facilitate semantic-aware transmission, we have modeled the KG into a wDG. Next, for an appropriate SR, we have derived the closed-form expression for the achievable throughput for within the maximum acceptable delay. Moreover, we have formulated the relay vehicle selection and SU assignment as a combinatorial optimization problem to optimize energy efficiency and semantic reliability. To finding a favorable solution within limited time, we have solved the problem with a low-complexity M-MTSA based on Markov approximation, where the solution is iteratively optimized. To demonstrate the feasibility of the PreCMTS, we have simulated it with realistic vehicle traces generated by SUMO. The high energy efficiency, semantic transmission reliability, and semantic energy efficiency of PreCMTS have been demonstrated with simulations.

REFERENCES

- [1] J. Luettin, S. Monka, C. Henson, and L. Halilaj, "A survey on knowledge graph-based methods for automated driving," in *Proc. Iberoamerican Knowl. Graph Semantic Web Conf.* Cham, Switzerland: Springer, 2022, pp. 16–31.
- [2] J. E. Kim, C. Henson, K. Huang, T. A. Tran, and W.-Y. Lin, "Accelerating road sign ground truth construction with knowledge graph and machine learning," in *Intelligent Computing*. Cham, Switzerland: Springer, 2021, pp. 325–340.
- [3] L. Halilaj, I. Dindorkar, J. Lüttin, and S. Rothermel, "A knowledge graph-based approach for situation comprehension in driving scenarios," in *Proc. Eur. Semantic Web Conf.* Cham, Switzerland: Springer, 2021, pp. 699–716.
- [4] X. Liu, Z. Xu, Y. Meng, W. Wang, J. Xie, and Y. Li, "An elastic-segment-Based V2V/V2I cooperative strategy for throughput enhancement," *IEEE Trans. Veh. Technol.*, vol. 71, no. 5, pp. 5272–5283, May 2022.
- [5] W. Weaver, "Recent contributions to the mathematical theory of communication," *ETC, Rev. Gen. Semantics*, vol. 74, no. 1, pp. 261–281, 1953.
- [6] W. Yang et al., "Semantic communications for future internet: Fundamentals, applications, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 213–250, Jan. 2022.
- [7] W. Yang et al., "Semantic communication meets edge intelligence," *IEEE Wireless Commun.*, vol. 29, no. 5, pp. 28–35, Oct. 2022.
- [8] C.-H. Lee, J.-W. Lin, P.-H. Chen, and Y.-C. Chang, "Deep learning-constructed joint transmission-recognition for Internet of Things," *IEEE Access*, vol. 7, pp. 76547–76561, 2019.
- [9] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Trans. Signal Process.*, vol. 69, pp. 2663–2675, 2021.
- [10] Z. Weng, Z. Qin, and G. Y. Li, "Semantic communications for speech signals," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2021, pp. 1–6.
- [11] H. Qiu, A. Ayara, and B. Glimm, "A knowledge architecture layer for map data in autonomous vehicles," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2020, pp. 1–6.
- [12] Z. Q. Liew, Y. Cheng, W. Y. B. Lim, D. Niyato, C. Miao, and S. Sun, "Economics of semantic communication system in wireless powered Internet of Things," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 8637–8641.
- [13] L. Liu, M. Zhao, M. Yu, M. A. Jan, D. Lan, and A. Taherkordi, "Mobility-aware multi-hop task offloading for autonomous driving in vehicular edge computing and networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 2, pp. 2169–2182, Feb. 2023.
- [14] H. Zhang, X. Zhang, and D. K. Sung, "An efficient cooperative transmission based opportunistic broadcast scheme in VANETs," *IEEE Trans. Mobile Comput.*, vol. 22, no. 3, pp. 1327–1342, Mar. 2023.
- [15] P. Kolios, V. Friderikos, and K. Papadaki, "Load balancing via store-carry and forward relaying in cellular networks," in *Proc. IEEE Global Telecommun. Conf.*, Dec. 2010, pp. 1–6.
- [16] Y. Wang, Y. Liu, J. Zhang, H. Ye, and Z. Tan, "Cooperative store-carry-forward scheme for intermittently connected vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 1, pp. 777–784, Jan. 2017.
- [17] H. Xie and Z. Qin, "A lite distributed semantic communication system for Internet of Things," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 142–153, Jan. 2021.
- [18] Q. Zhou, R. Li, Z. Zhao, C. Peng, and H. Zhang, "Semantic communication with adaptive universal transformer," *IEEE Wireless Commun. Lett.*, vol. 11, no. 3, pp. 453–457, Mar. 2022.
- [19] K. Lu et al., "Rethinking modern communication from semantic coding to semantic communication," *IEEE Wireless Commun.*, vol. 30, no. 1, pp. 158–164, Feb. 2023.
- [20] M. K. Farshbafan, W. Saad, and M. Debbah, "Curriculum learning for goal-oriented semantic communications with a common language," *IEEE Trans. Commun.*, vol. 71, no. 3, pp. 1430–1446, Mar. 2023.
- [21] C. Kurisummoottil Thomas and W. Saad, "Neuro-symbolic causal reasoning meets signaling game for emergent semantic communications," 2022, *arXiv:2210.12040*.
- [22] Y. Wang et al., "Performance optimization for semantic communications: An attention-based reinforcement learning approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2598–2613, Sep. 2022.
- [23] F. Zhou, Y. Li, X. Zhang, Q. Wu, X. Lei, and R. Q. Hu, "Cognitive semantic communication systems driven by knowledge graph," in *Proc. IEEE Int. Conf. Commun.*, May 2022, pp. 4860–4865.
- [24] H. Seo, J. Park, M. Bennis, and M. Debbah, "Semantics-native communication via contextual reasoning," *IEEE Trans. Cognit. Commun. Netw.*, vol. 9, no. 3, pp. 604–617, Jun. 2023.
- [25] K. Fall, "A delay-tolerant network architecture for challenged internets," in *Proc. Conf. Appl., Technol., Archit., Protocols Comput. Commun.*, Aug. 2003, pp. 27–34.
- [26] D. Wu, G. Zhu, and D. Zhao, "Adaptive carry-store forward scheme in two-hop vehicular delay tolerant networks," *IEEE Commun. Lett.*, vol. 17, no. 4, pp. 721–724, Apr. 2013.
- [27] S. H. Bouk, S. H. Ahmed, B. Omoniwa, and D. Kim, "Outage minimization using bivious relaying scheme in vehicular delay tolerant networks," *Wireless Pers. Commun.*, vol. 84, no. 4, pp. 2679–2692, Oct. 2015.
- [28] O. Trullols-Cruces, M. Fiore, and J. Barcelo-Ordinas, "Cooperative download in vehicular environments," *IEEE Trans. Mobile Comput.*, vol. 11, no. 4, pp. 663–678, Apr. 2012.
- [29] J. Chen, A. Zafar, G. Mao, and C. Li, "On the achievable throughput of cooperative vehicular networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–7.
- [30] A. Agarwal and P. R. Kumar, "Capacity bounds for ad hoc and hybrid wireless networks," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 34, no. 3, pp. 71–81, Jul. 2004.
- [31] G. Zhu et al., "Scene graph generation: A comprehensive survey," 2022, *arXiv:2201.00443*.
- [32] W. Zhang, J. Chen, J. Li, Z. Xu, J. Z. Pan, and H. Chen, "Knowledge graph reasoning with logics and embeddings: Survey and perspective," 2022, *arXiv:2202.07412*.
- [33] P. Zhang et al., "Toward wisdom-evolutionary and primitive-concise 6G: A new paradigm of semantic communication networks," *Engineering*, vol. 8, pp. 60–73, Jan. 2022.
- [34] P. Wu, L. Ding, Y. Wang, and H. Zheng, "V2V-assisted V2I mmWave communication for cooperative perception with information value-based relay," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2021, pp. 1–6.
- [35] W. Zhang et al., "Multi-hop connectivity probability in infrastructure-based vehicular networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 4, pp. 740–747, May 2012.
- [36] L. Su et al., "Content distribution based on joint V2I and V2V scheduling in mmWave vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 71, no. 3, pp. 3201–3213, Mar. 2022.
- [37] L. Xu et al., "Socially driven joint optimization of communication, caching, and computing resources in vehicular networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 1, pp. 461–476, Jan. 2022.
- [38] S. K. Yoo et al., "A comprehensive analysis of the achievable channel capacity in \mathcal{F} composite fading channels," *IEEE Access*, vol. 7, pp. 34078–34094, 2019.
- [39] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. Amsterdam, The Netherlands: Elsevier, 2007.

- [40] M. Chen, S. C. Liew, Z. Shao, and C. Kai, "Markov approximation for combinatorial network optimization," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6301–6327, Oct. 2013.
- [41] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
- [42] O. Ibe, *Markov Processes for Stochastic Modeling*, 2nd ed. London, U.K.: Elsevier, 2013.
- [43] C. Chen, J. Hu, T. Qiu, M. Atiquzzaman, and Z. Ren, "CVCG: Cooperative V2V-aided transmission scheme based on coalitional game for popular content distribution in vehicular ad-hoc networks," *IEEE Trans. Mobile Comput.*, vol. 18, no. 12, pp. 2811–2828, Dec. 2019.



Wanting Yang (Member, IEEE) received the B.S. and Ph.D. degrees from the Department of Communications Engineering, Jilin University, Changchun, China, in 2018 and 2023, respectively. From 2021 to 2022, she was a Visiting Student with the Singapore University of Technology and Design, sponsored by the Chinese Scholarship Council. Her research interests include wireless communications, predictive resource allocation, semantic communication, learning, martingale, and URLLC. She served as a Technical Program Committee Member for flagship conferences, such as WCNC, GLOBECOM, and VTC.



Xuefen Chi received the B.Eng. degree in applied physics from the Beijing University of Posts and Telecommunications, Beijing, China, in 1984, and the M.S. and Ph.D. degrees from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China, in 1990 and 2003, respectively. She was a Visiting Scholar with the Department of Computer Science, Loughborough University, U.K., in 2007, and the School of Electronics and Computer Science, University of Southampton, Southampton, U.K., in 2015. She is currently a Professor with the Department of Communications Engineering, Jilin University, China. Her research interests include machine-type communications, indoor visible light communications, random access algorithms, delay-QoS guarantees, and network modeling theory and its applications.



Linlin Zhao (Member, IEEE) received the B.Eng., M.S., and Ph.D. degrees from the Department of Communications Engineering, Jilin University, Changchun, China, in 2009, 2012, and 2017, respectively. From 2017 to 2019, she was a Post-Doctoral Researcher with the Department of Communications Engineering, Jilin University. She is currently an Associate Professor with the Department of Communications Engineering, Jilin University, and a Post-Doctoral Research Fellow with the State Key Laboratory of Internet of Things for Smart City, University of Macau. Her current research interests include throughput optimal random access algorithms, resource allocation schemes, and delay and reliability analysis and optimization, especially for reliability analysis of ultra-reliable low-latency communications. She was a recipient of the Best Ph.D. Thesis Award of Jilin University in 2017 and acquired the Macau Young Scholars Program in 2019. She has served as the Registration Co-Chair for IEEE ICCC 2019.



Zehui Xiong (Member, IEEE) received the Ph.D. degree from Nanyang Technological University (NTU), Singapore. He was a Visiting Scholar with Princeton University and the University of Waterloo. He is currently an Assistant Professor with the Singapore University of Technology and Design and also an Honorary Adjunct Senior Research Scientist with the Alibaba-NTU Singapore Joint Research Institute, Singapore. He has published more than 200 research papers in leading journals and flagship conferences and many of them are ESI highly cited papers. His research interests include wireless communications, the Internet of Things, blockchain, edge intelligence, and Metaverse. He was a recipient of the IEEE Early Career Researcher Award for Excellence in Scalable Computing, the IEEE Technical Committee on Blockchain and Distributed Ledger Technologies Early Career Award, the IEEE Internet Technical Committee Early Achievement Award, the IEEE TCSVC Rising Star Award, the IEEE TCI Rising Star Award, the IEEE TCCLD Rising Star Award, the IEEE Best Land Transportation Paper Award, the IEEE CSIM Technical Committee Best Journal Paper Award, the IEEE SPCC Technical Committee Best Paper Award, the IEEE VTS Singapore Best Paper Award, the Chinese Government Award for Outstanding Students Abroad, and the NTU SCSE Best Ph.D. Thesis Runner-Up Award. He has won over ten best paper awards in international conferences and is listed in the World's Top 2% Scientists identified by Stanford University. He is also serving as the Editor or the Guest Editor for many leading journals, including IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE INTERNET OF THINGS JOURNAL, IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, and IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING. He is also serving as the Associate Director for the Future Communications Research and Development Program. In 2023, he was featured on the list of Forbes Asia 30 Under 30.



Wenchao Jiang received the Ph.D. degree from the Department of Computer Science and Engineering, University of Minnesota, Twin Cities, in 2019. He is currently an Assistant Professor with the Pillar of Information System Technology and Design, Singapore University of Technology and Design. His research interests include the Internet of Things, wireless and low-power embedded networks, and mobile computing.