Select2Drive: Pragmatic Communications for Real-Time Collaborative Autonomous Driving

Jiahao Huang, *Student Member, IEEE*, Jianhang Zhu, *Member, IEEE*, Rongpeng Li, *Senior Member, IEEE*, Zhifeng Zhao, *Member, IEEE*, and Honggang Zhang

***Abstract*—Vehicle-to-everything communications-assisted autonomous driving has witnessed remarkable advancements in recent years, with pragmatic communications (PragComm) emerging as a promising paradigm for real-time collaboration among vehicles and other agents. Simultaneously, extensive research has explored the interplay between collaborative perception and decision-making in end-to-end driving frameworks. In this work, we revisit the collaborative driving problem and propose the Select2Drive framework to optimize the utilization of limited computational and communication resources. Particularly, to mitigate cumulative latency in perception and decision-making, Select2Drive introduces distributed predictive perception by formulating an active prediction paradigm and simplifying high-dimensional semantic feature prediction into a computationally efficient, motion-aware reconstruction. Given the "less is more" principle that an over-broadened perceptual horizon possibly confuses the decision module rather than contributing to it, Select2Drive utilizes area-of-importance-based PragComm to prioritize the communication of critical regions, thus boosting both communication efficiency and decision-making efficacy. Empirical evaluations on the V2Xverse and real-world DAIR-V2X datasets demonstrate that Select2Drive achieves a 2.60% and 1.99% improvement in offline perception tasks under limited bandwidth (resp., pose error conditions). Moreover, it delivers at most 8.35% and 2.65% enhancement in closed-loop driving scores and route completion rates, particularly in scenarios characterized by dense traffic and high-speed dynamics.**

***Index Terms*—Collaborative perception, pragmatic communications, data-based approaches, connected and autonomous vehicles.**

I. Introduction

DUE to the inherent limitations of Autonomous Driving (AD), such as restricted visibility [1], unpredictability of other road users [2], and difficulties in determining optimal paths [3], Vehicle-to-Everything (V2X) Communications have become an indispensable ingredient in the Internet of Vehicles. By enabling the exchange of complementary information among vehicles, roadside units (RSUs), and even pedestrians, V2X communications promises a broadened perceptual horizon for individual autonomous vehicles [4], contributing to timely identification of emergent objects beyond visual observations [5] and swiftly making proper responses [6]. Conventionally, early studies in the field of V2X communications focused on the realization of ubiquitous connectivity for accomplishing collaborative perception [7]. However, the associated communication costs scale linearly with the size of the perceptual region and the time duration grows quadratically with the number of collaborating agents [8], placing significant demands on even next-generation communication systems [9]. Meanwhile, collaborative perception within a small number of neighboring agents and a limited timeframe only yields marginal performance improvement over single-agent perception [10]. Fortunately, for V2X communications-assisted AD (V2X-AD), its sole reliance on reliable communications and the ignorance of the lasting impact of perception results on autonomous driving decisions still leave enormous room for optimization.

Pragmatic Communications (PragComm), which aims to deliver compact latent representations tailored to specific downstream decision-making tasks, can better take into account both collaborative perception according to sensor data and subsequent driving decisions simultaneously [18]. Widely known as pragmatic compression or effective communications, the PragComm is commonly deployed as a compression paradigm in the context of V2X-AD [10]. These methods operate under a fundamental assumption: during each time interval τ, all participating agents first broadcast Basic Safety Messages (BSMs) and subsequently decide whether to engage in communication [13] or exchange valuable perception blocks [14]. However, this approach presumes an idealized scenario in which the entire process, regardless of the number of point-to-point communication links, can be completed within each τ. Apparently, this assumption is impractical due to inevitable latency from transmission and inference delays.[1]

Received 21 January 2025; revised 21 July 2025 and 5 September 2025; accepted 14 September 2025. Date of publication 29 September 2025; date of current version 1 December 2025. This work was supported in part by the National Key Research and Development Program of China under Grant 2024YFE0200600, in part by Zhejiang Provincial Natural Science Foundation of China under Grant LR23F010005, and in part by Huawei Cooperation Project under Grant TC20240829036. The Associate Editor for this article was R. Qin. *(Corresponding author: Rongpeng Li.)*

Jiahao Huang, Jianhang Zhu, and Rongpeng Li are with Zhejiang University, Hangzhou 310027, China (e-mail: 22331083@zju.edu.cn; zhujh20@zju.edu.cn; lirongpeng@zju.edu.cn).

Zhifeng Zhao is with the Zhejiang Laboratory, Hangzhou 310012, China, and also with Zhejiang University, Hangzhou 310027, China (e-mail: zhaozf@zhejianglab.org).

Honggang Zhang is with Macau University of Science and Technology, Macau, China (e-mail: hgzhang@must.edu.mo).

This article has supplementary downloadable material available at https://doi.org/10.1109/TITS.2025.3611377, provided by the authors.

Digital Object Identifier 10.1109/TITS.2025.3611377

[1]Latency here specifically refers to the minimum response time required for a background vehicle to collect, process, and transmit data until the information is fused at the ego vehicle.

On the other hand, despite advancements in collaborative perception, a critical gap lies in understanding how perception enhancements impact integrated, system-level driving performance. Typically, Imitation Learning (IL) [1] instead of Reinforcement Learning (RL) [19] is adopted owing to the remarkable performance of Behavior Cloning (BC) in accident scenarios on predefined routes. Counterintuitively, as shown in findings from Ref. [20], particularly under augmented, collaborative perception, an expanded field of vision does not consistently improve decision-making, advocating for a "*less is more*" principle in V2X-AD. In other words, for closed-loop driving tasks, isolated perception modules often fail to seamlessly benefit subsequent planning and control stages, while incurring troublesome *error propagation*, since inaccuracies in perception accumulate through the system [21]. Therefore, in order to address latency-induced collaborative perception inconsistencies and ensure a consistent driving improvement, PragComm shall be redeveloped beyond simple context compression.

In this paper, we propose Select2Drive, a revamped PragComm-based framework that not only accounts for the compensation of overall latency but also incorporates calibrations tailored for eliminating error propagation in V2X-AD. Particularly, on top of a formulated delivery model that contributes to evaluating the underlying physical transmission plausibility [22], Select2Drive introduces a novel Distributed Predictive Perception (DPP) module, which is capable of predicting future semantic features using low-level indicators. Notably, despite the conceptual simplicity, implementing DPP is non-trivial, as the limited computational capability requires precise forecasting of future states from high-dimensional voxel flow or pseudo-maps, focusing on minimizing disparities between predicted and current heatmaps. Furthermore, inspired by the underscored benefits of constrained observational horizons [23], Select2Drive investigates the feasibility of decision-making strategies using minimal observation content. This finally culminates in an Area-of-Importance-based PragComm (APC) framework, which prioritizes communications in driving-critical regions. While providing key distinctions with highly relevant literature in Table I, our key contribution could be summarized as follows:

- To significantly boost the closed-loop driving performance under the impact of communication and computational latency, we propose a PragComm-based, IL-enabled real-time collaborative driving framework Select2Drive. Beyond simple information compression, the DPP and APC components therein can effectively incorporate background vehicle information while avoiding redundant computational burden and minimizing unnecessary communication.
- The calibrated DPP component integrates a predictive mechanism and a motion-aware affine transformation, which leverages low-dimensional motion flow to infer future semantic features. Avoiding direct prediction of high-dimensional Bird's Eye View (BEV) semantic features effectively mitigates timeliness challenges without introducing significant computational cost.
- Bearing the "*less is more*" principle in mind, we introduce a revamped APC component that restricts the communication region to the Area-of-Importance (AoIm), effectively alleviating the *covariate shift* induced by BC on constrained datasets. Therefore, Select2Drive enables prioritized communication in driving-critical regions and solves the latency-induced fusion inconsistencies from collaborative perception.
- Building upon the CARLA Simulator [24] and prior studies [11], we develop a comprehensive simulation platform[2] that transitions collaborative perception approaches from offline datasets to closed-loop driving scenarios [25] while offering an extensible interface for multi-vehicle cooperative driving. Through extensive experiments on both collaborative perception tasks and online closed-loop driving tasks, we demonstrate significantly improved performance (e.g., 2.60% higher perception accuracy in a simulated dataset V2Xverse, 1.99% higher perception accuracy in a real-world dataset DAIR-V2X, 8.35% higher closed-loop driving scores, and 2.65% larger route completion rates) of Select2Drive across diverse communications-limited scenarios.

The remainder of this paper is organized as follows. We introduce related works in Section II. We introduce system models and formulate the problem in Section III. In Section IV, we elaborate on the details of our proposed prediction paradigm. In Section V, we present the experimental results and discussions. Finally, Section VI concludes this paper.

II. Related Works

A. End-to-End Autonomous Driving

Recent advancement in learning-based end-to-end autonomous driving, which directly translates environmental observations into control signals [1] and conceptually addresses the cascading errors of traditional modular design [26], has positioned this domain as a pivotal research focus. Nevertheless, existing methods highlight a gap between theoretical assumptions and practical implementation. For example, Ref. [27] demonstrates the performance of collaborative perception algorithms in simulated environments but these are rarely applied to real-world driving tasks. Ref. [28] assumes accurate agent position data, which is often impractical in real-world scenarios. Our approach bridges this gap by integrating theoretical strategies with higher-fidelity implementations, which utilize perception data directly from emulated raw sensor inputs for more realistic analysis.

1) Learning Approaches: End-to-end driving approaches can be classified into RL-based or supervised learning-based IL [21].[3] Compared to RL-based solutions, IL progressively benefits from the increasing perception performance, leading to a stable enhancement in the learning of driving tasks through BC [1]. Notably, BC demonstrates effective performance for in-distribution states within the training dataset but

[2]The open-source codes can be found at https://github.com/zjunice once the manuscript has been accepted.

[3]Traditional RL/IL methods are typically limited to lower-dimensional problems. Therefore, the methodologies discussed in this paper specifically refer to Deep Learning (DL)-driven RL/IL approaches.

struggles to generalize to Out-Of-Distribution (OOD) states due to compounding action errors, a phenomenon termed *covariate shift* [29]. To mitigate this, we intentionally add noise to expert control signal to ensure more states within the training distribution [30].

2) AD: Ref. [6] proposes a visually cooperative driving framework that aggregates voxel representations from multiple collaborators to improve decision-making. Ref. [2] demonstrates that besides challenges in predicting the motion of out-of-view or non-interactive objects, single-agent driving systems inherently struggle with occluded or distant regions, often leading to catastrophic failures. To address these limitations, V2X-AD adopts a multi-agent collaborative paradigm leveraging V2X communications, enabling vehicles to share information and collaboratively make informed decisions [27]. Despite the remarkable progress, the latest evaluation platform [11] remains constrained by idealized communication assumptions.

B. Pragmatic Communications

Commonly formulated as an extension of the Markov Decision Process (MDP) framework [31], PragComm shifts the focus from accurate bit transmission or precise semantic interpretation to capturing key information and creating compact representations for specific downstream tasks.

1) V2X Communications: Information exchange for V2X cooperation can be posed as an image-transmission task whereby vehicles periodically capture and disseminate camera frames. Considering RGB image sharing, a front-view camera operating at 10 Hz with 2048 $\times$ 1024 resolution and 24-bit color depth produces approximately 48 Mb per frame; lossless PNG compression reduces this to about 18.85 Mb [32]. As of 2024, 3GPP specifies up to 53 Mbps for User Equipment (UEs) information sharing in V2X applications [9], implying a maximum image-sharing rate of roughly 2.81 Hz, which is insufficient for exchanging raw camera data in the near term. Therefore, efficient filtering and compression of perception data are essential for real-time performance. PragComm is contingent on the underlying capability of V2X communications, such as IEEE 802.11p-based DSRC [33] and the 3GPP Cellular-based V2X (C-V2X) [34]. Both architectures define BSMs [35], [36], transmitted periodically at up to 10 Hz to convey critical state information such as position, dynamics, and vehicle status. Correspondingly, high-frequency BSMs can serve as a foundation for high-dimensional semantic feature communication, minimizing redundant transmissions. For DSRC-based transmission, bandwidth-limited channel conditions highlight the necessity to investigate the impact of communication latency on collaborative perception, while the reliance on inter-node routing in C-V2X-based transmission necessitates a focus on systemic overall delays.

2) PragComm in V2X-AD: Ref. [14] establishes a PragComm-based framework towards achieving a balance between perception performance and communication costs in V2X-AD. It employs a two-step strategy: (1) semantic feature extraction from raw sensory data to low-level heatmaps as indicators; (2) selective transmission of high-value semantic features for fusion to optimize communication efficiency.

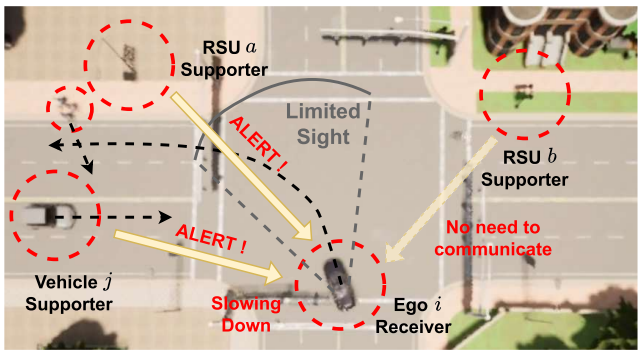


Fig. 1. Overview of V2X-AD. Contingent on pragmatic communications of driving-critical information with nearby supporters (e.g., vehicles and RSUs), the Ego vehicle maintains safe AD.

However, considering the heterogeneity in distance and content, PragComm in V2X-AD encounters difficulties spanning from localization uncertainty [16] to clock synchronization and dynamic delay compensation [15]. For example, even minimal delays can profoundly undermine the timeliness of transmitted information, potentially incurring catastrophic outcomes [37]. Meanwhile, prior methodologies primarily focus on reconstructing the distribution of proximal objects. While enhancing perception, these methods often misalign with driving policy optimization, necessitating integrated frameworks for cohesive performance. In that regard, Ref. [31] underscores that the decoupling of learning and communication yields suboptimal results. Therefore, there emerges a strong incentive to revamp PragComm for AD.

Compared to the literature, Select2Drive employs DPP, which diverges from traditional approaches by integrating a prediction mechanism at the supporter level to alleviate the impact of inevitable delays without imposing considerable computational burdens. Meanwhile, Select2Drive takes advantage of APC to bridge the disconnection between perception modules and low-level controllers by explicitly incorporating prior trajectory information into communication strategies. Therefore, Select2Drive not only further minimizes communication overhead but also sharpens the model's focus on task-critical information, ultimately enhancing driving performance.

III. System Model and Problem Formulation

Beforehand, primary notations used in this paper are summarized in Table II. In the subsequent discourse, intermediate variables output by the Deep Neural Network (DNN) will be denoted using a script font (e.g., $\mathcal{F}_i^t$), while directly observable variables will be represented in a standard font (e.g., $D_i^{t_r}$). In addition, a DNN will be denoted as a function $\Phi(\cdot)$.

A. System Model

In this paper, we consider a collaborative perception-based AD scenario with multiple vehicles (i.e., agents). Particularly, as shown in Fig. 2, let t represent the moment when an agent i initiates a decision-making cycle. At time t, agent i can perceive raw data (e.g., RGB images and 3D point clouds)

TABLE I
A COMPARISON BETWEEN SELECT2DRIVE AND RELATED WORKS

References	Realistic Communications	Latency Considered	Perception Involved	Driving-Task Oriented	Brief Description
[11]	○	○	●	●	Integrates basic collaborative perception into closed-loop driving, lacks communication frameworks and real-world latency simulation.
[12]	○	○	●	○	Proposes Dual-Perception Network (DP-Net), a lightweight network enabling simultaneous individual/cooperative 3D detection with State-Of-The-Art (SOTA) performance.
[6]	●	○	○	●	A blind-spot warning mechanism without engaging in precise collaborative perception and lack of generalization ability.
[13], [14]	○	○	●	○	Fetches the most valuable information for exchange under the premise of an ideal communication assumption, susceptible to latency issues.
[15], [16]	○	●	●	○	A centralized estimation upon the timing of incoming information, imposes significant challenges on mobile devices' computational and storage capacities.
[17]	●	●	●	○	A centralized, latency-based collaborator selection mechanism, incorporating the receiver's historical data into perception, proves inefficient in utilizing communication resources effectively.
Ours	●	●	●	●	Implementation of a distributed prediction mechanism to mitigate overall latency, and pre-filtering invaluable information based on driving context before communication.

Notations: ○ *indicates not included;* ● *indicates fully included.*

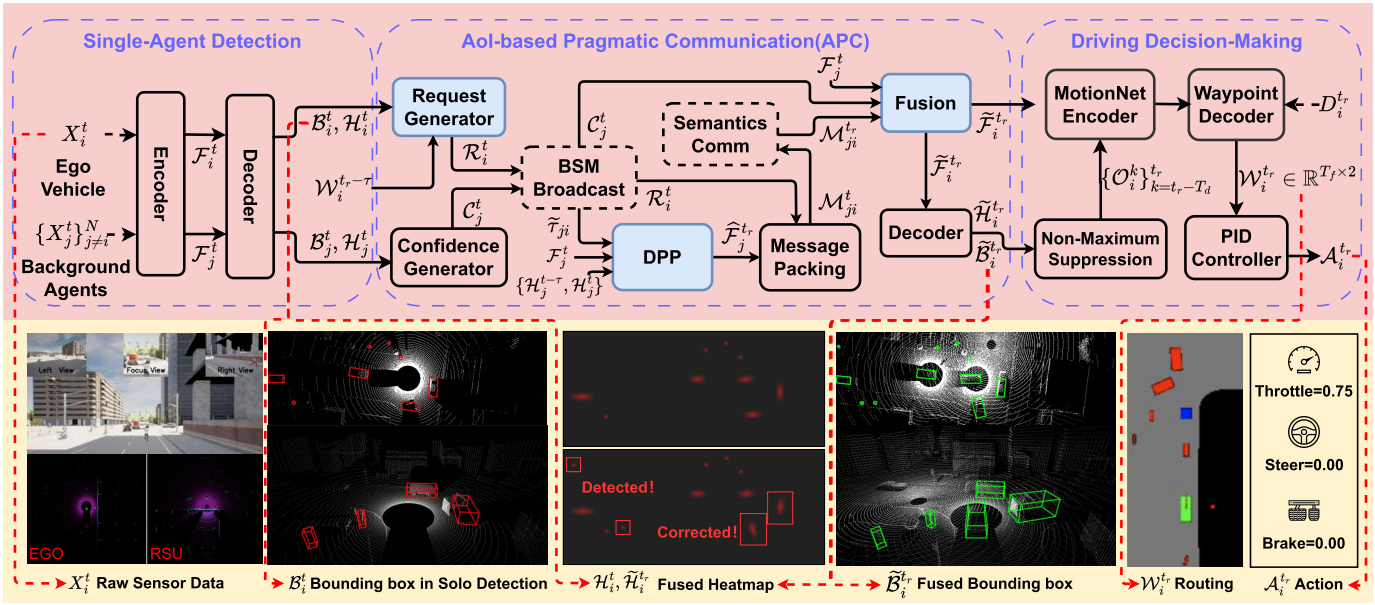


Fig. 2. System model of our V2X-AD framework encompassing perception, decision-making, and control stages. The upper section provides a detailed closed-loop flowchart, illustrating the iterative cycle from perception to action, while incorporating feedback into subsequent iterations. The lower section visually depicts the complete decision-making process, emphasizing the sequential flow of actions and data exchange.

at a fixed interval τ, while the communications possibly occur between any ego agent i and one of its supporting neighbors j (i.e., background vehicles and RSUs). Afterwards, agent i aims to maximize the accomplishment rate of its IL-based driving task with driving plan $\mathcal{W}_i^{t_r}$, contingent on the fusion of its own observed raw data X_i^t and exchanged information $\{\mathcal{M}_{ji}^t\}_{\mathcal{N}_i^t}$ from neighboring agents $j \in \mathcal{N}_i^t$. Basically, such a scenario can be classified as a pragmatic communications-based MDP.

1) Confidence-Driven Message Packing: After obtaining the raw sensor data X_i^t, each vehicle leverages an encoder Φ_{encoder}, which consists of a series of 2D convolutions and max-pooling layers, to yield the latest available semantic features $\mathcal{F}_i^t$ that merge RGB images and 3D point clouds into a unified global coordinate system, namely

$$\mathcal{F}_i^t = \Phi_{\text{encoder}}(X_i^t) \in \mathbb{R}^{H\times W\times D}, \tag{1}$$

where H, W, and D denote the dimensions of the pseudo-image. Typically, $D \gg 1$ even only 3D point clouds are utilized. Subsequently, a decoder Φ_{decoder}, composed of several deconvolution layers, is employed to generate a probability heatmap $\mathcal{H}_i^t$ and a bounding box regression map $\mathcal{B}_i^t$. The heatmap $\mathcal{H}_i^t$ represents the spatial likelihood of an object (e.g., vehicles, pedestrians, or traffic signs) being present in an image or frame, while the regression map $\mathcal{B}_i^t$ provides precise localization details (e.g., center coordinates, width, and height) for detected objects. Therefore,

$$\mathcal{H}_i^t, \mathcal{B}_i^t = \Phi_{\text{decoder}}(\mathcal{F}_i^t) \in \mathbb{R}^{H\times W\times C}, \mathbb{R}^{H\times W\times 8C}, \tag{2}$$

TABLE II
A SUMMARY OF MAJOR NOTATIONS USED IN THIS PAPER

Notation	Definition
$X_i^t, \mathcal{F}_i^t$	Raw sensor data and latest available semantic features of agent i at time t
$\mathcal{H}_i^t, \mathcal{B}_i^t$	Heatmap and bounding box from agent i
$\mathcal{H}_j^{t-\tau}, \mathcal{H}_j^t$	Historical heatmaps from agent j
$\widehat{\mathcal{H}}_j^{t_r}, \widehat{\mathcal{F}}_j^{t_r}$	Forecasted heatmap and processed semantic features from agent j
$\mathcal{C}_j^t, \mathcal{R}_i^t$	Confidence map from agent j and request map from agent i
$\widetilde{\mathcal{F}}_i^{t_r}, \widetilde{\mathcal{H}}_i^{t_r}, \widetilde{\mathcal{B}}_i^{t_r}$	Fused semantic feature and collaborated perception of agent i
$\{\mathcal{O}_i^k\}_{k=t_r-T_d}^{t_r}$	T_d frames of historical Bird's Eye View (BEV) occupancy maps in the view of agent i
$\mathcal{W}_i^{t_r}, \mathcal{A}_i^{t_r}$	Estimated trajectory and expected driving action of agent i
$\Delta\tau$	Broadcast period of request map $\mathcal{R}_i^t$
$\delta\tau_{ji}, \widetilde{\delta}\tau_{ji}$	Overall transmission latency between agent j and agent i, and the related estimation
$\tau_{ji}, \widetilde{\tau}_{ji}$	Real systematic latency between agent j and agent i, and the related estimation

with C representing the number of object categories and $C = 3$ if three categories, i.e. vehicles, bicycles, and pedestrians, are detected.

Afterward, the agent i sends low-dimensional BSMs, including an Φ_{Gen}-induced confidence map $\mathcal{C}_i^t$ and a request map $\mathcal{R}_i^t$, as:

$$\mathcal{C}_i^t = \Phi_{\text{Gen}}(\mathcal{H}_i^t) \in [0,1]^{H\times W}, \tag{3}$$

$$\mathcal{R}_i^t = 1 - \mathcal{C}_i^t \in [0,1]^{H\times W}, \tag{4}$$

where Φ_{Gen} denotes a maximum operation in the third dimension followed by a Gaussian filter. Under the ideal latency-free assumption, for the supporting vehicle $j \in \mathcal{N}_i^t$, confidence-driven messages for feedback are given as:

$$\mathcal{M}_{ji}^t = \mathcal{F}_j^t \times \mathcal{P}_{ji}^t \in \mathbb{R}^{H\times W\times D}. \tag{5}$$

Here, $\mathcal{P}_{ji}^t = \mathbf{1}\left(\mathcal{R}_i^t \odot \mathcal{C}_j^t \geq p_{\text{thre}}\right) \in \mathbb{R}^{H\times W}$ indicates a spatial selection mechanism for $\mathcal{F}_j^t$, and p_{thre} is a hyperparameter controlling the extent of collaboration. The indicator $\mathbf{1}(\cdot)$ equals 1 if the condition is met; while nulls otherwise. The operator $\odot$ denotes element-wise multiplication.

2) Latency Model: The acquisition, communication and post-processing of $\mathcal{M}_{ji}^t$ inevitably incur some latency, such as the computational latency involved in semantic extraction τ_j^{ext} and post-processing for decision-making τ_i^{dm}, the asynchronous inter-agent timing differences and latency jitter τ_{ji}^{asyn}, and the more prominent communication latency[4] τ_{ji}^{tx}.

As depicted in Fig. 3, to quantify τ_{ji}^{tx}, a verification mechanism proposed in [38] is employed. Notably, in the multi-channel alternating switch mode therein, the communication process is structured into a Synchronization Interval (SI), denoted as τ, which is further divided into a Service Channel Interval (SCHI) and a Control Channel Interval (CCHI), each lasting $\Delta\tau$. During the SCHI, BSMs, such as ($\mathcal{R}_i^t$ and $\mathcal{C}_i^t$) are broadcast, while semantic information $\mathcal{M}_{ji}^t$

[4]Notably, due to the low-dimensional nature of BSMs, the latency for transmitting BSMs is assumed to be negligible relative to that for transmitting high-dimensional semantics $\mathcal{M}_{ji}^t$ (i.e., τ_{ji}).

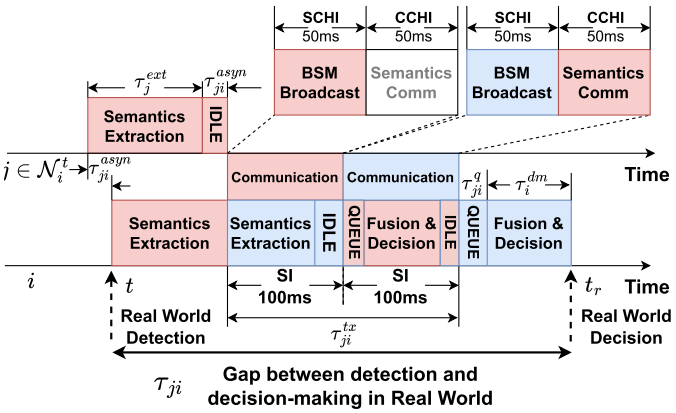


Fig. 3. Flow chart from the perspective of agent i. The red box delineates the decision cycle initiated at time t, while the blue box represents the subsequent cycle commencing at time $t+\tau$. The interval between consecutive perception and communication phases is uniformly set to τ.

is transmitted during the subsequent CCHI. The minimum transmission time for $\mathcal{M}_{ji}^t$ is given by

$$\tau_{ji}^{\text{tx}} = \tau_{ji}^{\text{pr}} + \tau_{ji}^{\text{net}}. \tag{6}$$

Here, τ_{ji}^{pr} represents the propagation latency, computed as per 3GPP TR 38.901 [39], that is,

$$\tau_{ji}^{\text{pr}} = \text{size}(\mathcal{M}_{ji}^t) \Big/ \left(b_{ji}\log_2(1+10^{0.1(p_{ji}^{\text{tx}}-p_{ji}^{\text{loss}}-p_{ji}^{\text{noise}})})\right), \tag{7}$$

where b_{ji} is the bandwidth allocated per agent, p_{ji}^{tx} is the transmission power, p_{ji}^{noise} is the noise power, and p_{ji}^{loss} denotes the path loss calculated by $p_{ji}^{\text{loss}} = 28 + 22\log_{10}(d_{ji}) + 20\log_{10}(f_c)$ [39], with d_{ji} being the inter-agent distance (in meters) and f_c the carrier frequency (in GHz). The term τ_{ji}^{net} accounts for the processing time at network nodes (e.g., routers, switches, base stations) before forwarding data to the next hop. Owing to the direct communication characteristics of DSRC-based transmission, τ_{ji}^{tx} is predominantly constrained by τ_{ji}^{pr}, with negligible τ_{ji}^{net} [4]. Conversely, the multicast service in C-V2X-based transmission effectively mitigates τ_{ji}^{pr} while introducing substantial τ_{ji}^{net} delays, primarily attributed to computational burdens at network nodes caused by access and handover overhead [40].

In a nutshell, the overall latency τ_{ji} can be expressed as:

$$\tau_{ji} = \tau_j^{\text{ext}} + \tau_{ji}^{\text{asyn}} + \tau_{ji}^{\text{tx}} + \tau_i^{\text{dm}} + \tau_{ji}^{\text{q}}, \tag{8}$$

where the term τ_{ji}^{q} denotes the queueing latency [32] for the ego agent to sequentially process multiple agent interactions. For notational simplicity, t_r denotes the moment when the agent i obtains the message $\mathcal{M}_{ji}^t$ from another agent j during the cycle starting at t. Thus, $t_r = t + \tau_{ji}$.

3) Information Fusion and Decision-Making: After the communications, the ego vehicle would aggregate all available information $\{\mathcal{M}_{ji}^t\}_{i\cup\mathcal{N}_i^t}$[5] to derive fused features $\mathcal{F}_i^{t_r}$

$$\mathcal{F}_i^{t_r} = \Phi_{\text{fuse}}\left(\{\mathcal{Z}_{ji}^{t_r} \odot \mathcal{M}_{ji}^t\}_{i\cup\mathcal{N}_i^t}\right) \in \mathbb{R}^{H\times W\times D}, \tag{9}$$

[5]For simplicity of representation, we denote $\mathcal{M}_{ii}^t = \mathcal{F}_i^t$ and $\tau_{ii} = 0$.

where Φ_{fuse} is implemented with a feed-forward network and $\mathcal{Z}_{ji}^{t_r} = \Phi_{\text{MHA}}\left(\mathcal{F}_i^t, \mathcal{M}_{ji}^t, \mathcal{M}_{ji}^t\right) \odot \mathcal{C}_j^t \in \mathbb{R}^{H\times W}$ indicates a Scale-Dot Product Attention (SDPA) [41] generated with per-location multi-head attention Φ_{MHA}. Next, by decoding the fused feature $\mathcal{F}_i^{t_r}$ through a predefined decoder Φ_{decoder} as:

$$\mathcal{H}_i^{t_r}, \mathcal{B}_i^{t_r} = \Phi_{\text{decoder}}(\mathcal{F}_i^{t_r}) \in \mathbb{R}^{H\times W\times C}, \mathbb{R}^{H\times W\times 8C}, \tag{10}$$

where $\mathcal{H}_i^{t_r}$ and $\mathcal{B}_i^{t_r}$ represent the heatmap and bounding box regression map obtained with fused semantic information $\mathcal{F}_i^{t_r}$, respectively. 3D objects are then detected via non-maximum suppression [42] and rasterized into a binary BEV occupancy map $\mathcal{O}_i^{t_r}$. Using T_d-length historical occupancy maps $\{\mathcal{O}_i^k\}_{k=t_r-T_d}^{t_r}$ as well as the navigation information $D_i^{t_r}$, the ego vehicle leverages a learnable planner Φ_{plan} encompassing a MotionNet encoder, a goal encoder and corresponding waypoint decoder [26] to generate a driving plan consisting of a series of waypoints $\mathcal{W}_i^{t_r}$. Mathematically, it can be described as:

$$\mathcal{W}_i^{t_r} = \Phi_{\text{plan}}(\{\mathcal{H}_i^{t_r}, \mathcal{B}_i^{t_r}\}_{k=1}^{T_d}, \mathcal{F}_i^{t_r}, D_i^{t_r}) \in \mathbb{R}^{2\times T_f}. \tag{11}$$

The optimal driving action $\mathcal{A}_i^{t_r}$, comprising steering, throttle, and brake commands, is then determined via lateral and longitudinal Proportional–Integral–Derivative (PID) controllers $\Phi_{\text{controller}}$ as:

$$\mathcal{A}_i^{t_r} = \Phi_{\text{controller}}(\mathcal{W}_i^{t_r}) \in [0,1]^2 \cup \{0,1\}^1. \tag{12}$$

B. Problem Formulation

This paper aims to maximize the achievable driving performance through calibrated pragmatic communication. Particularly, the PragComm-based V2X-AD problem can be consistently formulated as:

$$\begin{aligned} &\max_{\theta,\eta} \sum_{i=1}^{N} \mathcal{E}\left[\mathcal{W}_i^{t_r}, \Phi_{\text{plan}}\left(\Phi_{\text{percep}}(X_i^t, \{\mathcal{M}_{ji}^t\}_{\mathcal{N}_i^t})\right)\right], \\ &\text{s.t. } \mathcal{M}_{ji}^t = \Phi_{\text{process}}(\mathcal{F}_j^t, \mathcal{H}_i^t), \\ &\quad \mathcal{M}_{ji}^t \le b_{ji} * \Delta\tau \text{ for } j \in \mathcal{N}_i^t, \end{aligned} \tag{13}$$

where Φ_{percep} represents all involved perception-related DNNs in Eqs. (1) to (10). Specially, Φ_{process} corresponds to the DNN-based PragComm components outlined in Eqs. (4) and (5), and the operator $\mathcal{E}(\cdot)$ indicates metrics [43] (e.g., route completion rates, infraction penalty and driving scores) in the driving scenarios, considering both safety rate and traffic efficiency. While the request map, as derived in Eq. (4), provides an intuitive foundation, it lacks task-specific optimization, such as the prioritization of information relevant to navigation information $D_i^{t_r}$, route planning [44], or salient objects [45]. Even worse, as mentioned earlier in Section II-B, under ideal channel conditions, communication resources are insufficient to achieve latency-free communication, even for extremely compressed messages.

On the other hand, for the highly volatile AD environment, due to the existence of computation and communications latency τ_{ji}, the currently available observations in Eq. (5) might become outdated at time t_r. Instead, directly transmitting the forecast semantic features, predicted at time t, to complement the possible impact of latency τ_{ji} is preferable. Nevertheless, although the estimation of the overall latency $\widetilde{\tau}_{ji}$ is achievable through a synchronization mechanism as in Section III-A.2, the prediction of high-dimensional $\mathcal{F}_j^t$ might impose a significant computational burden on the mobile device. Therefore, beyond simple information compression, Φ_{process} (particularly Eq. (5)) shall be carefully investigated to effectively incorporate predicted background vehicle information at the expense of reduced computational overhead and minimal communications.

IV. Select2Drive: Driving-Oriented Collaborative Perception

In this section, we introduce a Select2Drive framework that prioritizes the communication of decision-critical, timely content into the collaborative driving process. To obtain computationally efficient prediction, we reformulate it as a dimensionality reduction-based reconstruction problem and devise a DPP to extract the inherent transformation $\overrightarrow{\mathcal{H}}_j^{t_r}$, which represents the motion flow of objects from $\mathcal{F}_j^t$ to $\mathcal{F}_j^{t_r}$, and subsequently infer $\mathcal{F}_j^{t_r}$ from an affine approximation of $\mathcal{F}_j^t$. Furthermore, to ensure that improvements in perception performance consistently translate to enhanced outcomes in offline driving simulations, we design the message-packing mechanism (i.e., APC) on top of DPP.

A. Distributed Predictive Perception (DPP)

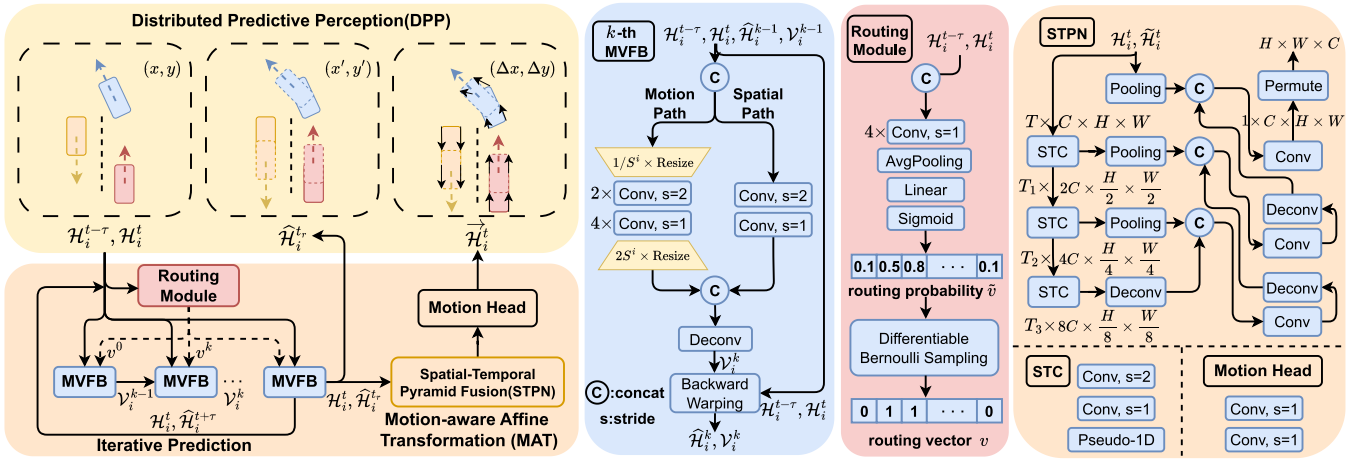
As illustrated in Fig. 4, instead of directly predicting $\overrightarrow{\mathcal{H}}_j^{t_r}$ from high-level semantics $\mathcal{F}_j^t$, which exhibits significant sensitivity to continuous latency τ_{ji}, we first downsample the semantics $\{\mathcal{F}_j^t, \mathcal{F}_j^{t_r}\}$ to low-level heatmap [46] $\{\mathcal{H}_j^t, \mathcal{H}_j^{t_r}\}$ with the decoder Φ_{decoder} in Eq. (2). As mentioned earlier, due to the temporary unavailability of $\mathcal{H}_j^{t_r}$, we leverage a video prediction-inspired *iterative prediction* method to learn a predicted version $\widehat{\mathcal{H}}_j^{t_r}$. Next, we introduce a *motion-aware affine transformation* mechanism to extract motion information $\overrightarrow{\mathcal{H}}_j^{t_r} \in \mathbb{R}^{H\times W\times 2}$, which corresponds to the 2-dimensional positional shifts $(\Delta x, \Delta y)$ for every object initially located at (x, y).

1) Iterative Prediction: We discretize the estimated latency $\widetilde{\tau}_{ji}$ into discrete steps $n_{ji}^t = \lfloor \widetilde{\tau}_{ji}/\tau \rfloor$ [37]. Thus, $t_r' = t + n_{ji}^t \times \tau$, consistent with the decision-making cycle τ in Section III-A.2. We iteratively generate a heatmap sequence $\{\widehat{\mathcal{H}}_j^{t+\tau}, \ldots, \widehat{\mathcal{H}}_j^{t_r'}\}$ through n_{ji}^t steps as:

$$\widehat{\mathcal{H}}_j^{t+\tau}, \ldots, \widehat{\mathcal{H}}_j^{t_r'} \overset{\text{iteratively}}{\underset{\text{for } n_{ji}^t}{\rightleftharpoons}} \Phi_{\text{DMVFN}}\left(\mathcal{H}_j^{t-\tau}, \mathcal{H}_j^t\right), \tag{14}$$

while employing $\widehat{\mathcal{H}}_j^{t_r'}$ to approximate $\widehat{\mathcal{H}}_j^{t_r}$.

Before delving into the implementation details, Table III summarizes popular candidates and compares model parameter count, computational complexity (measured in FLOPs per inference), mean latency (evaluated empirically and estimated on the vehicle platform [55]), and the performance impact of individual module modifications. In short, we first prioritize real-world deployability under a 10 Hz decision frequency, and subsequently select the optimal models in the end-to-end perception task. Such a procedure leads to an integration of



(a) Main architecture (b) k-th MVFB (c) Routing Module (d) STPN and Motion Head

Fig. 4. Overview of the proposed DPP framework.

TABLE III

PARAMETERS AND COMPUTATIONAL OVERHEAD OF MAJOR MODULES. MODULES HIGHLIGHTED WITH **BOLD** ARE SELECTED AS THE BACKBONE OF OUR MODEL FOR INTEGRATION TO MEET THE 100 MS DECISION INTERVAL REQUIREMENT

Module	Params (M)	FLOPs (G)	Execution time on Desktop (ms)	Execution time on Vehicle (ms)	Δ Performance to Ours (%)
Future Confidence Forecast Module (Φ_{DMVFN})					
DMVFN [47]	3.6	2.1	1.17	1.10	0
PredRNN++ [48]	24.6	169.8	94.80	89.45	+0.62
TAU [49]	38.7	85.0	47.45	44.77	+0.77
MAU [50]	10.5	29.1	16.25	15.33	+0.27
PhyDNet [51]	5.8	80.7	45.00	42.46	+0.44
Semantic Feature Extraction Module (Φ_{encoder}, Φ_{decoder})					
PointPillar [52]	8.2	119	66.46	62.71	0
CenterPoint [53]	8.2	170	94.94	89.58	+0.10
Motion Perception Module (Φ_{MAT}, Φ_{plan})					
MotionNet [54]	1.7	10.2	5.70	5.38	0
LSTM	3.5	30.46	17.02	16.06	−4.56
Intermediate Feature Fusion Module (Φ_{fuse})					
SDPA [41]	0.007	0.29	0.16	0.15	0
Max Fusion	0	0	0.05	0.05	−3.07

[1] Execution time on Desktop is measured on an RTX 4090 (1, 321 TOPS) in single-step, one-to-one driving scenarios. Vehicle time is estimated by scaling against NVIDIA THOR [57] (2, 000 TOPS) with 70% utilization to account for operating system overhead. Notably, Max Fusion solely involves element-wise maximum operations without floating-point computations, resulting in null FLOPs.

[2] In the context of the perception task, "Δ Performance to Ours" quantifies the performance gap when a specific module is substituted, with our method serving as the established baseline.

PointPillar [52], DMVFN [47], MotionNet [54], and SDPA [41]. Notably, it trades a maximum 0.87% performance loss for a 50.43% (resp., 70.54 ms) reduction in decision-making latency on vehicles, resulting in a total latency of 69.34 ms. Specifically, DMVFN excels by operating without extra inputs and avoiding redundant convolutions, making it ideal for dense decision-making in autonomous driving.

As depicted in Fig. 4, to ensure remarkable performance in resource-constrained settings, the DMVFN employs $K = 9$ Multi-scale Voxel Flow Blocks (MVFBs) coupled with a dynamic routing module. Particularly, to effectively capture large-scale motion while maintaining spatial fidelity, each MVFB $k \in \{1, \cdots, K\}$ incorporates a dual-branch network structure, encompassing a motion path and a spatial path, to downsample the inputs by a scaling factor S^k for a larger receptive field while preserving fine-grained spatial details [56]. Subsequently, the outputs from both paths are concatenated to predict the voxel flow $\mathcal{V}_j^k$, which is then applied through backward warping [57] to generate a synthesized frame $\widehat{\mathcal{H}}_j^k$.

Without loss of generality, taking the example of inputting $(\mathcal{H}_j^{t-\tau}, \mathcal{H}_j^t)$, each MVFB k is achieved by processing these two historical frames, a synthesized frame $\widehat{\mathcal{H}}_j^{k-1}$ and the voxel flow $\mathcal{V}_j^{k-1}$ generated by the $k-1$ MVFB block. Thus, we have

$$\widehat{\mathcal{H}}_j^k, \mathcal{V}_j^k = \Phi_{\text{MVFB}}^k(\mathcal{H}_j^{t-\tau}, \mathcal{H}_j^t, \widehat{\mathcal{H}}_j^{k-1}, \mathcal{V}_j^{k-1}, S^k). \tag{15}$$

When $k = 1$, $\widehat{\mathcal{H}}_j^0$ and $\mathcal{V}_j^0$ are set to zero.

On the other hand, the routing module is designed to dynamically balance the activation of each MVFB block, enabling adaptive selection according to the input variability. Contingent on a lightweight DNN, the routing module is optimized using Differentiable Bernoulli Sampling (DBS), to prevent the routing module from converging to trivial

solutions (e.g., consistently activating or bypassing specific blocks). Specifically, DBS incorporates Gumbel-Softmax [58] to determine the selection $v^k \in \{0,1\}$ of k-th MVFB through a stochastic classification task governed by $\tilde{v}^k$ as:

$$v^k = \frac{\exp\left(\frac{1}{\beta}(\tilde{v}^k + G_k)\right)}{\exp\left(\frac{1}{\beta}(\tilde{v}^k + G_k)\right) + \exp\left(\frac{1}{\beta}(2 - \tilde{v}^k - G_k)\right)}, \tag{16}$$

where $G_k \in \mathbb{R}$ follows the Gumbel(0,1) distribution. The temperature parameter β starts with a high value to allow exploration of all possible paths and gradually decreases to approximate a one-hot distribution, ensuring effective and controllable routing. To ensure the participation of DBS in gradient computation during end-to-end training, the Straight-Through Estimator (STE) [59], which approximates the discrete sampling process in the backward pass, can be employed to further maintain compatibility with standard gradient descent optimization.

In summary, the prediction process for the k-th MVFB is formulated as:

$$\widehat{\mathcal{H}}_j^k, \mathcal{V}_j^k = \begin{cases} \Phi_{\text{MVFB}}^k(\mathcal{H}_j^{t-\tau}, \mathcal{H}_j^t, \widehat{\mathcal{H}}_j^{k-1}, \mathcal{V}_j^{k-1}, S^k), & v^k = 1; \\ \widehat{\mathcal{H}}_j^{k-1}, \mathcal{V}_j^{k-1}, & v^k = 0. \end{cases} \tag{17}$$

To enhance the video prediction model's capacity for capturing dynamic information in traffic flow scenarios within the original training framework, we combine the standard ℓ_1 loss, which controls the contribution of each stage and is regulated by a discount factor γ, with the VGG loss $\mathcal{L}_{\text{Vgg}}$ [60] with a weight ε. The VGG loss revolves around leveraging the feature extraction capabilities of pre-trained VGG networks to quantify perceptual differences between images. Mathematically,

$$\mathcal{L}_{\text{DMVFN}} = \sum\nolimits_{k=1}^{K} \gamma^{K-k} \ell_1(\mathcal{H}_j^{t+\tau}, \widehat{\mathcal{H}}_j^k) + \varepsilon \mathcal{L}_{\text{Vgg}}, \tag{18}$$

where $\mathcal{L}_{\text{Vgg}} = \sum_{m=1}^{M} \gamma_m \sum_{h,w,c=1}^{H_m,W_m,C_m} \frac{1}{H_m W_m C_m} \left(\phi_m(\mathcal{H}_j^{t+\tau})_{h,w,c} - \phi_m(\widehat{\mathcal{H}}_j^{t+\tau})_{h,w,c}\right)^2$. Here, $M = 5$ indicates the number of VGG layers we chose in the off-the-shelf VGG-19 network [60]. At the m-th layer, $\phi_m(\zeta)$ refers to the feature representation of input ζ and contributes to total loss with corresponding weight γ_m, and $\phi_m(\zeta)_{h,w,c}$ specifies the value of the feature map at the h-th row, w-th column, and c-th channel for the input ζ [60]. H_m, W_m, and C_m represent the height, width and channel count of the feature map at the m-th layer, respectively.

2) Motion-Aware Affine Transformation (MAT): Building upon the foundational work of [54], Φ_{MAT} computes the motion prediction flow $\overrightarrow{\mathcal{H}}_j^{t_r}$, which explicitly encodes relative positional shifts between $\mathcal{H}_j^t$ and $\widehat{\mathcal{H}}_j^{t_r}$.

$$\overrightarrow{\mathcal{H}}_j^{t_r} = \Phi_{\text{MAT}}(\mathcal{H}_j^t, \widehat{\mathcal{H}}_j^{t_r}), \tag{19}$$

As depicted in Fig. 4, the MotionNet for Φ_{MAT} consists of two primary components: a Spatial-Temporal Pyramid Network (STPN) and a motion head, implemented by a two-layer 2D convolution module. The STPN is designed to extract multi-scale spatio-temporal features through its Spatial-Temporal Convolution (STC) block. The STC integrates standard 2D convolutions with a pseudo-1D convolution, which serves as a degenerate 3D convolution with kernel size $T_m \times 1 \times 1$, where $\{T_m\}_{m=1,2,3}$ corresponds to the temporal dimension, enabling efficient feature extraction across both spatial and temporal dimensions. Spatially, the STPN computes feature maps at multiple scales with a scaling factor of 2, while temporally, it progressively reduces the temporal resolution to capture hierarchical temporal semantics. Following this, global temporal pooling, and a feature decoder with lateral connections and upsample layers are employed to aggregate and refine the extracted temporal features, ensuring robust motion representation.

To precisely estimate the motion flow $\overrightarrow{\mathcal{H}}_j^{t_r}$, the loss function for Φ_{MAT} is defined using the smooth ℓ_1 loss as:

$$\mathcal{L}_{\text{MAT}} = \left\| \sum_k \sum_{(x,y),(x',y') \in o_k} f_A\left(f_\Delta(\overline{\mathcal{H}}_{(x,y)}^t, \overline{\mathcal{H}}_{(x',y')}^{t_r})\right) - \overrightarrow{\mathcal{H}}_j^{t_r} \right\|, \tag{20}$$

where $f_\Delta(\overline{\mathcal{H}}_{(x,y)}^t, \overline{\mathcal{H}}_{(x',y')}^{t_r}) \in \mathbb{R}^2$ represents the aggregated motion (i.e., $(\Delta x, \Delta y) = (x', y') - (x, y)$) of object k within instance o_k over the interval $[t, t_r]$, which is derived through grid-level comparisons between the Ground-Truth (GT) heatmaps $\overline{\mathcal{H}}_i^t$ and $\overline{\mathcal{H}}_i^{t_r}$. The operator $f_A(\cdot)$ indicates a simple affine operation to map the increment $(\Delta x, \Delta y)$ to the x-th column, y-th row into a $H \times W$ matrix. Subsequently, the transformation of the semantic feature $\mathcal{F}_j^t$ can be directly performed with the help of motion flow $\overrightarrow{\mathcal{H}}_j^{t_r}$ as

$$\widehat{\mathcal{F}}_j^{t_r}(x,y) = \mathcal{F}_j^t\left[x + \overrightarrow{\mathcal{H}}_j^{t_r}(x,y,0), y + \overrightarrow{\mathcal{H}}_j^{t_r}(x,y,1)\right]. \tag{21}$$

B. AoIm-Based Pragmatic Communications

To incorporate driving-related information within the PragComm procedure, we initiate by generating the request map $\mathcal{R}_i^t$. Given the inherent ambiguity of relying solely on navigation information $D_i^{t_r}$ [61], the request map is constructed as a Gaussian distribution centered on the nearest waypoint (W_x, W_y) within prior waypoint plan $\mathcal{W}_i^{t_r-\tau}$, inspired by [45]. The formulation is given by:

$$\mathcal{R}_i^t(x,y) = \frac{1}{\sigma_F \sqrt{2\pi}} \exp\left(-\frac{(x - W_x)^2 + (y - W_y)^2}{2\sigma_F^2}\right), \tag{22}$$

where σ_F, termed the *Focus Radius*, is a hyperparameter controlling the width of the Gaussian distribution.

With the assistance of DPP, we further emphasize the dynamic information during message packing by computing $\Delta\mathcal{C}_j^t(x,y) = \left|\Phi_{\text{Gen}}(\widehat{\mathcal{H}}_j^{t_r})(x,y) - \mathcal{C}_j^t(x,y)\right|$ as an alert signal, which has been proven to be practical in prior works [37]. The message $\mathcal{M}_{ji}^t$ is then packed as:

$$\mathcal{M}_{ji}^t = \widehat{\mathcal{F}}_j^{t_r} \times \mathcal{P}_{ji}^t, \tag{23}$$

where $\mathcal{P}_{ji}^t = \mathbf{1}\left(\max\left(\mathcal{R}_i^t \odot \Phi_{\text{Gen}}(\widehat{\mathcal{H}}_j^{t_r}), \Delta\mathcal{C}_j^t / n_{ji}^t\right) \geq p_{\text{thre}}\right)$. Subsequently, the information delivery, fusion and decision-making procedures can be conducted as in Section III-A.3. In summary, Select2Drive can be executed as in Algorithm 1.

Algorithm 1 Select2Drive

Input: Raw sensor data and last planned waypoints $\{X_i^t, \mathcal{W}_i^{t_r-\tau}\}_{i\cup\mathcal{N}_i^t}$ of ego i and its neighboring agents $j \in \mathcal{N}_i^t$,

Output: Next driving action for each agent $\{\mathcal{A}_i^{t_r}\}_{i\cup\mathcal{N}_i^t}$

for *each agent* i **do**
  Generate intermediate semantic features $\mathcal{F}_i^t$ along with solo-perception results $\mathcal{H}_i^t, \mathcal{B}_i^t$ based on X_i^t using Eqs. (1)(2);
  Exchange request map $\mathcal{R}_i^t$ based on prior driving plan $\mathcal{W}_i^{t_r-\tau}$ using Eq. (22) and estimate latency $\widetilde{\tau}_{ji}$ of sending message to neighbor $j \in \mathcal{N}_i^t$;
  for *neighboring agent* $\mathcal{N}_i^t$ **do**
    $n_{ji}^t \leftarrow \lfloor \widetilde{\tau}_{ji}/\tau \rfloor$;
    Predict future heatmap $\widehat{\mathcal{H}}_j^{t_r}$ based on historical information $\mathcal{H}_j^{t-\tau}, \mathcal{H}_j^t$ through n_{ji}^t iterations of DMVFN in Eq. (17);
    Extract motion flow $\overrightarrow{\mathcal{H}}_j^{t_r}$ between $\mathcal{H}_j^t$ and $\widehat{\mathcal{H}}_j^{t_r}$ with MAT in Eq. (19);
    Apply affine approximation $\overrightarrow{\mathcal{H}}_j^{t_r}$ on semantic features $\mathcal{F}_j^t$ to estimate high-level semantic information $\widehat{\mathcal{F}}_j^{t_r}$ with Eq. (21); ▷ DPP
    Send packed Message $\mathcal{M}_{ji}^t$ based on $\widehat{\mathcal{F}}_j^{t_r}$ and confidence map $\mathcal{C}_i^t$ generated with Eq. (4) using Eq. (23); ▷ APC
  end
  Fuse received message $\{\mathcal{M}_{ji}^t\}_{i\cup\mathcal{N}_i^t}$ and ego information $\mathcal{F}_i^t$ to obtain collaborated semantics $\widetilde{\mathcal{F}}_i^{t_r}$ using Eq. (9);
  Generate next driving plan $\mathcal{W}_i^{t_r}$ and make driving decision $\mathcal{A}_i^{t_r}$ based on $\widetilde{\mathcal{F}}_i^{t_r}$ using Eqs. (10)(12);
end
return Next action $\{\mathcal{A}_i^{t_r}\}_{i\cup\mathcal{N}_i^t}$

C. Training Methods

In order to train the DNNs in Select2Drive, we assume the existence of a dataset $\mathcal{T} = \{\xi_k\}_{k=0\ldots N}$, which comprises trajectories $\xi_k = \{(X_i^t, \mathcal{S}_i^{t_r}, \overline{\mathcal{W}_i}^{t_r})\}_{t=0\ldots T}$ representing sequences of state-action pairs, with actions $\overline{\mathcal{W}_i}^{t_r} = \pi_E(\mathcal{S}_i^{t_r})$ derived from an expert policy π_E, where the real state $\mathcal{S}_i^{t_r} = (\overline{\mathcal{H}}_i^t, \overline{\mathcal{B}}_i^t, D_i^{t_r})$. The training process is structured around two interconnected parts (i.e., the perception-related DNN and the planning policy). For the former part, a η-parameterized DNN Φ_{percep}, which encompasses the encoder Φ_{encoder}, decoder Φ_{decoder}, fuser Φ_{fuse} as well as the incorporated intermediate DNNs, especially DMVFN and MAT in DPP, is learned through minimizing the PointPillar perception loss through supervised learning,

$$\min_{\eta} \mathcal{L}(\eta) = \mathbb{E}_{(X_i^t, \mathcal{S}_i^{t_r})\in\mathcal{T}}[(\mathcal{S}_i^{t_r} - \mathcal{S}_i^{t_r})^2] + \mathcal{L}_{\text{DMVFN}} + \mathcal{L}_{\text{MAT}}, \quad (24)$$

where $\mathcal{S}_i^{t_r} = (\mathcal{H}_i^{t_r}, \mathcal{B}_i^{t_r}, D_i^{t_r})$ represents the estimated state.

On the other hand, the latter planning policy DNN Φ_{plan} parameterized by θ is trained using IL to minimize the l_2-norm deviation [62] between the low-level planning strategies and the expert policy π_E as:

$$\min_{\theta} \mathcal{L}(\theta) = \mathbb{E}_{(\mathcal{S}_i^{t_r}, \overline{\mathcal{W}_i}^{t_r})\in\mathcal{T}}[(\mathcal{W}_i^{t_r} - \overline{\mathcal{W}_i}^{t_r})^2], \quad (25)$$

where $\mathcal{W}_i^{t_r} = \Phi_{\text{plan}}(\mathcal{S}_i^{t_r}, \mathcal{F}_i^{t_r})$ represents the waypoint plan using the estimated state[6] $\mathcal{S}_i^{t_r}$ along with fused semantic features $\mathcal{F}_i^{t_r}$ given by the perception-related DNN with converged parameters η. Since the optimization objective of Φ_{plan} differs from that of Φ_{percep}, the planner is trained for the closed-loop task using features from the converged perception model.

TABLE IV
MAINLY USED PARAMETERS IN THIS PAPER

Parameter	Value
DSRC-based transmission	
Interval $\Delta\tau$ for SCHI and CCHI	50 ms
Fixed Decision Interval τ	100 ms
Allocated Bandwidth b_{ji}	$1 \sim 20$ MHz
Transmit Power p_{ji}^{tx}	23 dBm
Power of Noise p_{ji}^{noise}	$U(-95, -110)$ dBm
Carrier Frequency f_c	5.9 GHz
C-V2X-based transmission(ms)	
Fixed transmission Latency $\tau_{ji}^{\text{pr}} + \tau_{ji}^{\text{net}}$	$0 \sim 600$
Shared Latency-related parameters	
Packet Loss	5%
Asynchronous latency τ_{ji}^{asyn}	$U(-100, 100)$ ms
Queueing latency τ_{ji}^{q}	$U(0, 50)$ ms
Semantic Extraction Time τ_j^{ext}	$U(40, 50)$ ms
Decision-Making Time τ_i^{dm}	$U(20, 30)$ ms
Hyperparameters	
Height, Width, Channel $\{H, W, D\}$	$[192, 576, 64]$
Request Map Threshold p_{thre}	0.05
Focus Radius σ_F	15 m
Number of frames for planning T_d	5
Number of waypoints to plan T_f	10
Scaling factors $\{S^k\}_{k=1}^9$	$[4, 4, 4, 2, 2, 2, 1, 1, 1]$
Discount factor γ, VGG weight ε	$0.8, 0.5$
Index $\{m\}_M$ of VGG Layers	[2, 7, 12, 21, 30]
Corresponding weights $\{\gamma_m\}_M$	[0.38, 0.21, 0.27, 0.18, 6.67]
Temporal factors in STPN T_1, T_2, T_3	$[2, 2, 1]$

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. Experimental Settings

In this section, we evaluate the performance of Select2Drive in a high-fidelity environment based on CARLA simulator, which facilitates sensor rendering and the computation of physics-based updates to the world state. It adheres to the ASAM OpenDRIVE standard [63] for defining road networks and urban environments. Table IV outlines the principal experimental parameters, with communications-related parameters primarily mentioned in [38]. The values of τ_j^{ext} and τ_i^{dm} are obtained from Table III. Specifically, τ_j^{ext} denotes the aggregated latency of Φ_{encoder} and Φ_{process} with an average of 43.71 ms, whereas τ_i^{dm} represents the cumulative delay of Φ_{decoder}, Φ_{fuse}, and Φ_{plan}, averaging 24.34 ms. Furthermore, the value of the queueing latency τ_{ji}^{q} corresponds to 30.42 ms under

[6]The occupancy map $\mathcal{O}_i^{t_r}$ is derived through a non-trainable suppression mechanism and rasterization process applied to $\mathcal{S}_i^{t_r}$.

(a) Pedestrian crossing (b) Bicycle Overtaking (c) Pedestrian crossing from blind spot

Fig. 5. Visualization for closed-loop driving upon several accident scenarios. The visualization delineates the ego vehicle's position with a green box, the planned trajectory with a red dot, detected obstacles as red squares, and the next waypoint along the route as a blue square.

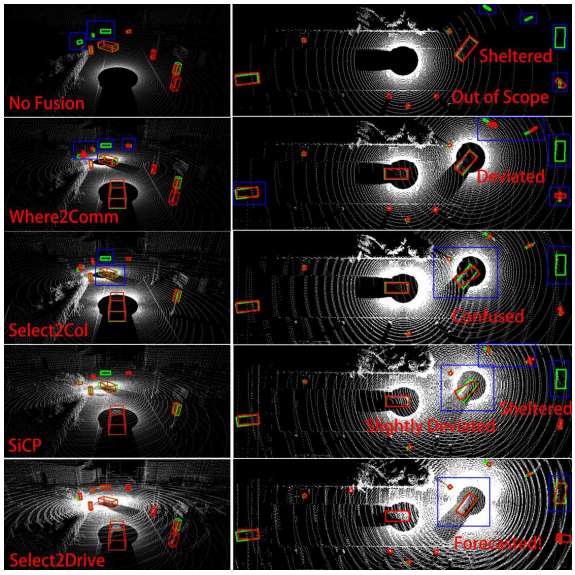
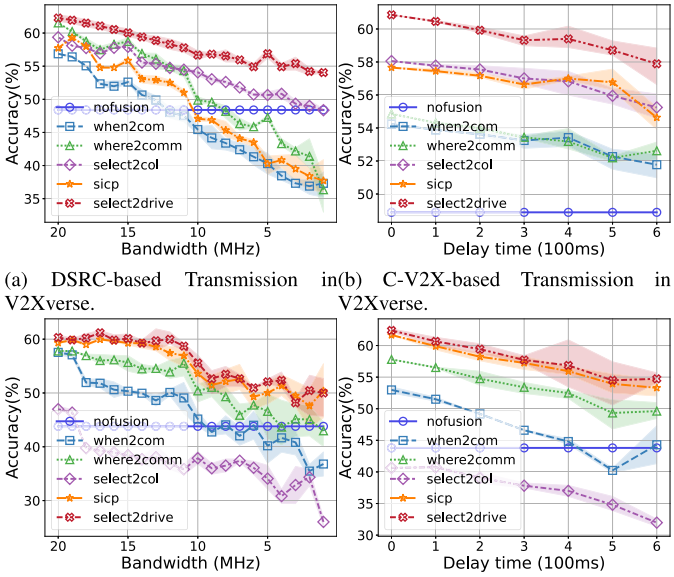


Fig. 6. Visualization of collaborative perception in bandwidth-constrained (5 MHz) scenarios. The red box illustrates the ego vehicle's predicted positions for surrounding objects, whereas the green box indicates the GT positions of those objects.



(a) DSRC-based Transmission in V2Xverse. (b) C-V2X-based Transmission in V2Xverse.

(c) DSRC-based Transmission in DAIR-V2X. (d) C-V2X-based Transmission in DAIR-V2X.

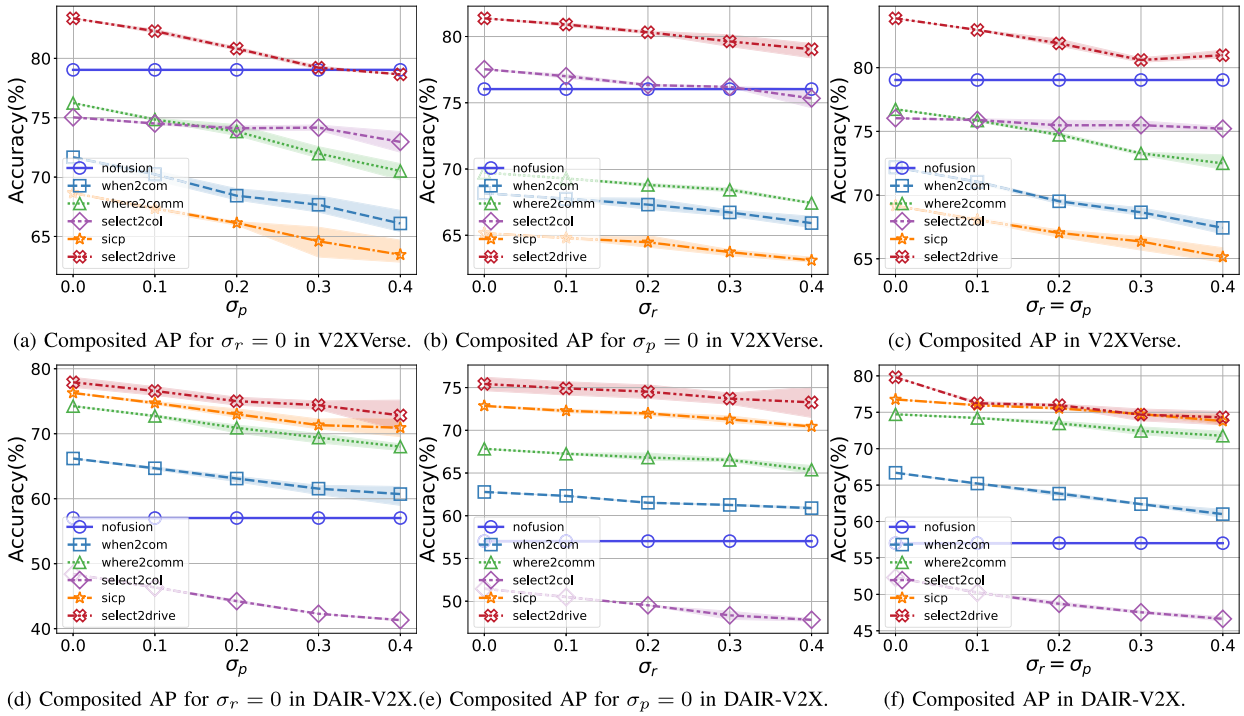
Fig. 7. Robustness of DPP to the communication constraints in the perception task.

conditions with 5 or more communicable agents, assuming a DSRC channel throughput of 20 Mbps and an arrival rate of 10 Hz modeled as an M/M/1 model [64].

The overall latency τ_{ji} is simulated separately with assumed bandwidth constraints in DSRC [33] and C-V2X [34] as in Section III-A.2. Since DSRC-based transmission encounters hidden node issues, which can lead to packet collisions [4], it is modeled by constraining τ_{ji}^{pr} with limited bandwidth, while $\tau_{ji}^{\text{net}} = 0$ due to its direct communication nature. In contrast, in C-V2X, the impact of τ_{ji}^{net} is more pronounced due to the possible handover procedures, and τ_{ji}^{tx} fluctuates within a bounded range. Motivated by the practice in [65], we simulate packet loss by applying random dropout on the transmitted message $\mathcal{M}_{ji}^{t}$, and model jitter τ_{ji}^{asyn} with varying variance levels. Specifically, when packet loss occurs, the features received by the ego vehicle are replaced with Gaussian noise, whereas jitter shifts the receive timestamp of semantic features, causing them to arrive earlier or later than expected.

As the V2X-AD framework naturally divides into perception and subsequent driving tasks, we evaluate our proposed method across two distinct stages.

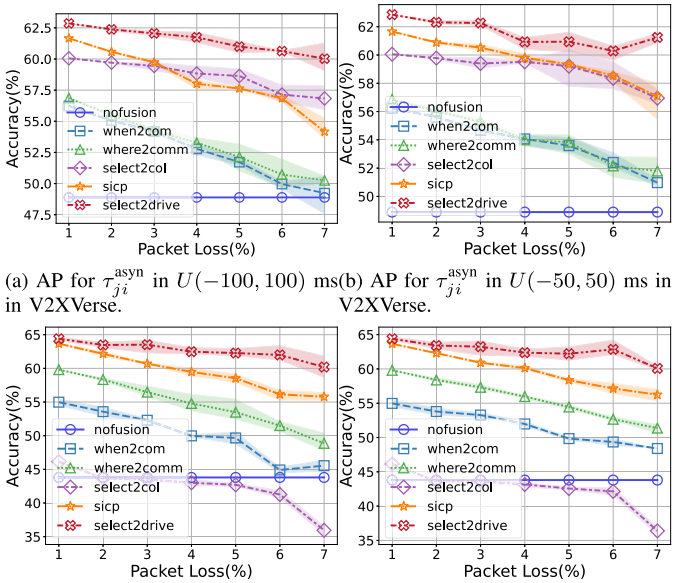
- *For planning policy*, we mainly simulate the closed-loop driving task through online route completion tasks. All decision-making policies are pre-trained on V2Xverse [11], while online tasks are tested on the 31 Town05 Short Routes in the CARLA Leaderboard [25] version 0.9.10, where the ego vehicle collaborates with the nearest agents (including vehicles and RSUs). Following [43], we employ three key evaluation metrics, including *route completion rate*, *infraction penalty*, and *driving score*, as mentioned in Section III. Route completion rates quantify the agent's ability to successfully complete navigation tasks, calculated as the percentage of the planned route traversed. Infraction penalty evaluates traffic rule compliance through a geometric penalty function that accounts for both violation severity and frequency. Driving scores integrate the aforementioned factors along with collision rates, serving as a more comprehensive performance metric. Consequently, the ranking in the table primarily follows the driving score criterion.
- *For perception capability*, we leverage the V2Xverse [11] and DAIR-V2X [66] datasets for offline perception



(a) Composited AP for $\sigma_r = 0$ in V2XVerse. (b) Composited AP for $\sigma_p = 0$ in V2XVerse. (c) Composited AP in V2XVerse.

(d) Composited AP for $\sigma_r = 0$ in DAIR-V2X.(e) Composited AP for $\sigma_p = 0$ in DAIR-V2X. (f) Composited AP in DAIR-V2X.

Fig. 8. Robustness of DPP to the vehicle pose noise in the perception task, where uniform latency τ_{ji}^{tx} is set to 100 ms.



(a) AP for τ_{ji}^{asyn} in $U(-100, 100)$ ms in V2XVerse. (b) AP for τ_{ji}^{asyn} in $U(-50, 50)$ ms in V2XVerse.

(c) AP for τ_{ji}^{asyn} in $U(-100, 100)$ ms in DAIR-V2X. (d) AP for τ_{ji}^{asyn} in $U(-50, 50)$ ms in DAIR-V2X.

Fig. 9. Robustness of DPP to the jitter and packet loss in the perception task, where uniform latency τ_{ji}^{tx} is set to 100 ms.

performance evaluation. The former dataset comprehensively incorporates RSUs compared to the widely used OPV2V dataset [5], extending beyond vehicle-to-vehicle (V2V) communications, while the latter is the latest real-world vehicle-infrastructure cooperative dataset. Notably, though some vehicles might not participate in the communications of perceived data to the ego vehicle, to account for their potential communications to others, we equally allocate the available bandwidth among all the ego's neighboring vehicles capable of communications. For evaluation, consistent with [17], we adopt *Average Precision (AP) at Intersection over Union (IoU)* thresholds of 0.3, 0.5, and 0.7 for vehicles, two-wheeled vehicles, and pedestrians, denoted as AP30, AP50, and AP70. As for communication volume, we calculate it as $\log_2(H \times W \times D \times \|\mathcal{P}_{ji}^t\|_1 \times 32/8)$ [14]. To enhance clarity, the Composited AP is a weighted sum of AP30, AP50, and AP70, with respective weights of 0.3, 0.3, and 0.4. Also, to streamline representation, perception results for vehicles, bicycles, and pedestrians are merged with weights of 0.4, 0.4, and 0.2 in Latency-induced scenarios shown in Fig. 7 and Fig. 9. To mitigate excessive oscillations in the curves caused by positioning noise, the weights become 0.8, 0.1, and 0.1 in Positioning-induced scenarios shown in Fig. 8. To quantitatively assess the statistical impact of latency fluctuations, we present performance curves accompanied by their corresponding confidence intervals. The solid line represents the mean values, while the shaded regions delineate the upper and lower bounds of the confidence interval.

For baseline methodologies, we reproduce When2Com [13], Where2Comm [14] and Select2Col [17], as well as the State-Of-The-Art (SOTA) SiCP [12], as the baseline of the perception task. Additionally, a no-fusion baseline is included to evaluate performance in the absence of collaborative mechanisms. Meanwhile, the baselines in the driving task include an IL-based planner trained atop the collaborative perception methods above. Additionally, prominent single-agent end-to-end methodologies are also considered, such as TCP [44] and the SOTA Interfuser [26].

TABLE V

CLOSED-LOOP DRIVING PERFORMANCE. THE APPROACH WITH THE BEST AVERAGE DRIVING SCORE IS HIGHLIGHTED IN **BOLD**, WHILE THE SECOND-BEST AND THE THIRD-BEST ARE MARKED WITH ITALICS AND UNDERLINES, RESPECTIVELY

Method	Avg. Driving Scores↑	Avg. Route Completion Rate (%)↑	Avg. Infraction Penalty↓	Collisions With Pedestrians↓	Collisions With Vehicles↓	Collisions With Layout↓	Off-road Infractions↓	Mean Speed↑
No Communications								
Interfuser [26]	35.372	79.254	0.434	0.052	0.492	0.568	0.223	0.586
TCP [44]	38.214	50.526	0.817	0.029	0.079	0.069	0.004	1.066
No Fusion	38.481	84.732	0.432	0.109	0.379	0.603	0.105	0.569
Bandwidth = 20 MHz (uniform latency = 0 ms)								
When2Com [13]	30.571	41.840	0.646	0.028	0.923	0.450	0.416	0.218
Where2Comm [14]	35.811	82.266	0.394	0.156	0.390	0.393	0.115	0.791
Select2Col [17]	35.178	69.045	0.492	0.126	0.572	0.371	0.106	0.442
SiCP [12]	43.289	80.159	0.466	0.111	0.205	0.852	0.071	1.082
Select2Drive wo APC	40.991	82.535	0.411	0.148	0.447	0.404	0.126	0.978
Select2Drive	46.904	82.284	0.446	0.140	0.270	0.008	0.083	1.211
Bandwidth = 10 MHz (uniform latency = 100 ms)								
When2Com	29.915	43.725	0.632	0.051	0.953	0.410	0.385	0.257
Where2Comm	33.704	48.560	0.651	0.054	0.640	0.351	0.115	0.668
Select2Col	29.794	70.232	0.414	0.189	0.516	0.348	0.118	0.448
SiCP	41.849	78.755	0.418	0.124	0.215	0.755	0.079	1.107
Select2Drive wo APC	40.725	66.760	0.496	0.052	0.469	0.482	0.142	1.066
Select2Drive	45.062	81.157	0.456	0.117	0.310	0.376	0.095	0.976
Bandwidth = 5 MHz (uniform latency = 200 ms)								
When2Com	27.204	38.392	0.652	0.043	1.114	0.377	0.514	0.507
Where2Comm	31.976	51.161	0.527	0.119	0.514	0.399	0.188	0.306
Select2Col	28.391	64.284	0.447	0.096	0.586	0.345	0.137	0.486
SiCP	40.511	66.415	0.405	0.132	0.229	0.795	0.075	0.912
Select2Drive wo APC	38.853	54.574	0.627	0.044	0.626	0.340	0.331	0.860
Select2Drive	43.823	70.588	0.520	0.088	0.373	0.497	0.126	1.211

[1] The experimental results present averaged measurements across 31 independent routes, evaluated under varying seed parameters while maintaining fixed shared parameters as specified in Table IV.

B. Quantitative Results

1) Driving Task: Quantitatively, as illustrated in Fig. 5, the proposed approach effectively leverages PragComm-based driving-critical information for emergent obstacle perception and timely collision avoidance.

Table V demonstrates a marked performance enhancement of 8.35% (resp., 3.62) in driving scores using the proposed methodology compared with the SOTA approach SiCP. Under bandwidth constraints, the collaborative driving paradigm maintains a 8.18% (resp., 3.31) Driving scores advantage compared to the latest available single-agent SOTA TCP method [44]. Notable gains can be observed for the road completion rate (2.65% improvement) and infraction penalty (4.29% reduction). These results confirm the universal superiority and robustness of our method in real-world scenarios. Ablation studies confirm the contribution of the APC methodology, yielding a 14.43% improvement in the driving score and empirically validating our "less is more" hypothesis. Conversely, latency-agnostic methods [13], [14] exhibit significant performance degradation under high-latency conditions.

2) Perception Performance: Fig. 6 presents qualitative findings. It can be observed that the predictive approach in Select2Drive facilitates the timely acquisition of projected data from surrounding vehicles, thus effectively addressing blind zone perception. Due to the neglect of latency, Where2Comm employs outdated information aggregation, and consequently generates notable perceptual inaccuracies. Meanwhile, Select2Col and SiCP offer partial remediation, yet remain susceptible to blind zone perception loss stemming from temporal constraints.

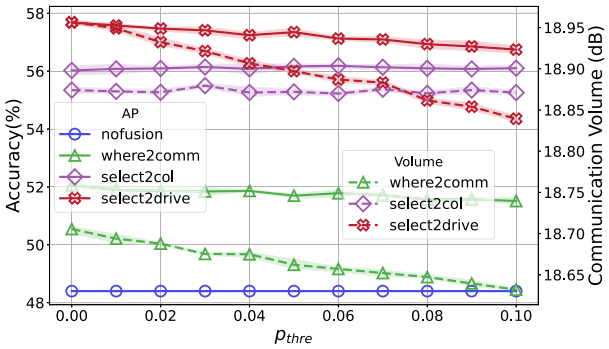
Fig. 7 illustrates the perceptual performance across various methodologies under DSRC-based and C-V2X-based transmission scenarios. Under ideal communication conditions, a comprehensive multi-object perception evaluation indicates our method outperforms the other baselines. Meanwhile, in realistic V2X scenarios where existing methods degrade to performance levels comparable to non-communicative baselines, DPP maintains consistent performance advantages: achieving gains of 2.60% (10 MHz bandwidth) and 2.32% (300 ms latency) over the second-best Select2Col method in V2Xverse, while demonstrating improvements of 1.99% (10 MHz) and 0.47% (300 ms) against the second-best SiCP approach in DAIR-V2X. Even under severely constrained bandwidth conditions, our method still maintains superior accuracy. The performance variation across different random seeds remains within 3% for all primary methods, with the illustrated gains calculated as the average difference.

As illustrated in Fig. 8, DPP demonstrates robust stability in the presence of angular noise σ_r. The angular noise, parameterized by a standard deviation σ_r in degrees, is formally modeled using a von Mises (or circular normal) distribution for the angle α with a concentration parameter κ given by the relation $\kappa = (180/(\pi \cdot \sigma_r))^2$. For cases of low dispersion (i.e., large κ), this distribution is well-approximated by a

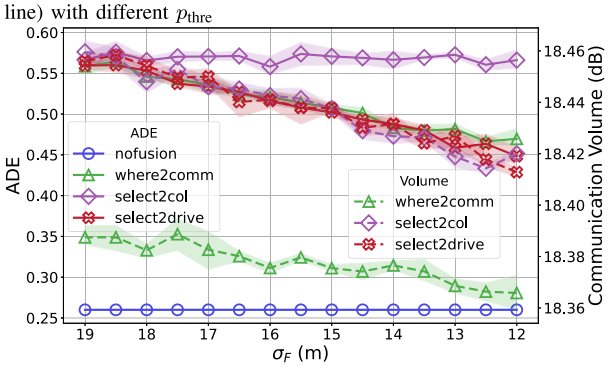
normal distribution with a zero mean and a variance of $\sigma^2 = (\pi \cdot \sigma_r/180)^2$. Meanwhile, its performance is slightly compromised when subjected to positioning noise σ_p (i.e., absolute deviations in (x, y, z)). Nevertheless, under moderate noise conditions, our method achieves significant gains of 3.27% ($\sigma_p = 0.1$, $\sigma_r = 0$), 1.87% ($\sigma_r = 0.1$, $\sigma_r = 0$), and 3.93% ($\sigma_p = \sigma_r = 0.1$) compared to non-collaborative perception schemes in V2Xverse. Meanwhile, the gain in DAIR-V2X is 19.55% ($\sigma_p = 0.1$, $\sigma_r = 0$), 18.94% ($\sigma_r = 0.1$, $\sigma_p = 0$), and 14.84% ($\sigma_p = \sigma_r = 0.1$). Even under severe noise conditions, our method achieves significant gains of 1.79% ($\sigma_p = 0.2$, $\sigma_r = 0$), 4.29% ($\sigma_r = 0.2$, $\sigma_p = 0$), and 1.95% ($\sigma_p = \sigma_r = 0.2$) compared to non-collaborative perception schemes in V2Xverse. Meanwhile, the gain in DAIR-V2X is 17.98% ($\sigma_p = 0.2$, $\sigma_r = 0$), 17.48% ($\sigma_r = 0.2$, $\sigma_p = 0$), and 18.94% ($\sigma_p = \sigma_r = 0.2$). This suggests the necessity of precise positional information, while our approach exhibits strong correction capabilities.

Fig. 9 represents the influence of different packet losses along with latency jitter on the performance. Specifically, when subjected to jitter with a variance of 100 ms, DPP exhibits only a marginal precision reduction of less than 2%, while the second-best SiCP approach experiences a significant performance degradation of 6.6% in the V2Xverse benchmark. Under elevated packet loss rates, our method exhibits reduced performance degradation and demonstrates superior robustness to burst jitter. This resilience is attributed to our DPP approach, which leverages temporal information from both preceding and succeeding frames. This multi-frame processing capability effectively mitigates communication failures caused by isolated spikes and enables cross-frame joint prediction to compensate for partial packet loss.

3) Hyperparameter Research: As depicted in Fig. 10, we investigate the impact of the hyperparameter p_{thre} in Eq. (5) and σ_F in Eq. (22). The hyperparameter p_{thre} predominantly regulates the stringency of message exchange. A higher value diminishes the volume of information involved in aggregation, potentially inducing a degradation in perception performance. To achieve an optimal trade-off between communication efficiency and perception accuracy, we empirically set $p_{\text{thre}} = 0.05$, referring to previous benchmarks [8]. It can be observed from Fig. 10(a), this configuration yields a decrease in communication overhead of 0.07 dB (4.74%), accompanied by a marginal 0.33% degradation in composite AP. To determine the optimal σ_F, we conduct an offline expert trajectory replication task. Performance is quantitatively assessed using the Average Displacement Error (ADE) [67], which measures the mean Euclidean distance between the system's predicted trajectories and the ground truth expert demonstrations. A lower value of σ_F enforces more stringent filtering of content extraneous to the driving task, thereby enhancing offline simulation performance during imitation of expert behaviors. However, under limited observational perspectives, the fidelity of expert imitation serves only as a reference metric rather than a direct determinant of ultimate performance, as the absence of collaboration leads to significant occlusions, illustrated in Fig. 6. Consequently, Fig. 10(b) indicates that setting $\sigma_F = 15$



(a) Composited AP (solid line) and Communication Volume (dotted line) with different p_{thre}



(b) ADE (solid line) and Communication Volume (dotted line) with different σ_F

Fig. 10. Quantitative analysis on the influence of p_{thre} and σ_F.

yielded an ADE reduction of 0.057 (10.20%), alongside a modest communication decrease of 0.023 dB (1.58%).

VI. Conclusion

In this work, we have presented Select2Drive, a PragComm-based real-time collaborative driving framework, which introduces two key components (i.e. DPP and APC) to address the critical timeliness challenges in V2X-AD systems. In particular, the DPP algorithm integrates predictive modeling and motion-aware affine transformation to infer future high-dimensional semantic features, maintaining robust perception performance even under severe positioning noise or constrained communication scenarios. Simultaneously, APC enhances decision-making efficiency by restricting communication to critical regions and minimizing unnecessary data exchanges, thereby mitigating potential confusion in decision-making. Extensive evaluations have been conducted on both collaborative perception tasks and online closed-loop driving simulations. The experimental results demonstrate that our communication-efficient optimization framework is well-suited for real-time collaborative perception tasks, achieving significant performance improvements across a wide range of scenarios. In the future, we will explore integrating generative models to enhance driving policy robustness across diverse scenarios.

References

[1] K. Renz, K. Chitta, O.-B. Mercea, A. S. Koepke, Z. Akata, and A. Geiger, "PlanT: Explainable planning transformers via object-level representations," in *Proc. 6th Conf. Robot Learn.*, Jan. 2022, pp. 459–470.

[2] H. Shao, L. Wang, R. Chen, S. L. Waslander, H. Li, and Y. Liu, "ReasonNet: End-to-end driving with temporal and global reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 13723–13733.

[3] Z. Peng, Q. Li, K. M. Hui, C. Liu, and B. Zhou, "Learning to simulate self-driven particles system with coordinated policy optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2021, pp. 10784–10797.

[4] R. Sedar, C. Kalalas, F. Vázquez-Gallego, L. Alonso, and J. Alonso-Zarate, "A comprehensive survey of V2X cybersecurity mechanisms and future research paths," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 325–391, 2023.

[5] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. Ma, "OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 2583–2589.

[6] J. Cui, H. Qiu, D. Chen, P. Stone, and Y. Zhu, "Coopernaut: End-to-end driving with cooperative perception for networked vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17231–17241.

[7] T.-H. Wang, S. Manivasagam, M. Liang, B. Yang, W. Zeng, and R. Urtasun, "V2VNet: Vehicle-to-vehicle communication for joint perception and prediction," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 605–621.

[8] R. Xu, H. Xiang, Z. Tu, X. Xia, M. Yang, and J. Ma, "V2X-ViT: Vehicle-to-everything cooperative perception with vision transformer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jan. 2022, pp. 107–124.

[9] *Physical Channels and Modulation*, document TS 38.211, 3GPP, Mar. 2024. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3180

[10] Y. Hu et al., "Pragmatic communication in multi-agent collaborative perception," 2024, *arXiv:2401.12694*.

[11] G. Liu et al., "Towards collaborative autonomous driving: Simulation platform and end-to-end system," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 8, pp. 6566–6584, Aug. 2025.

[12] D. Qu et al., "SiCP: Simultaneous individual and cooperative perception for 3D object detection in connected and automated vehicles," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2024, pp. 8905–8912.

[13] Y. Liu, J. Tian, N. Glaser, and Z. Kira, "When2com: Multi-agent perception via communication graph grouping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4105–4114.

[14] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," in *Proc. 36th Conf. Neural Inf. Process. Syst. (Neurips)*, Nov. 2022, pp. 4874–4886.

[15] S. Wei et al., "Asynchrony-robust collaborative perception via bird's eye view flow," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023, pp. 28462–28477.

[16] Z. Lei et al., "Robust collaborative perception without external localization and clock devices," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Yokohama, Japan, May 2024, pp. 7280–7286.

[17] Y. Liu et al., "Select2Col: Leveraging spatial–temporal importance of semantic information for efficient collaborative perception," *IEEE Trans. Veh. Technol.*, vol. 73, no. 9, pp. 12556–12569, Sep. 2024.

[18] D. Gündüz, F. Chiariotti, K. Huang, A. E. Kalør, S. Kobus, and P. Popovski, "Timely and massive communication in 6G: Pragmatics, learning, and inference," *IEEE BITS Inf. Theory Mag.*, vol. 3, no. 1, pp. 27–40, Mar. 2023.

[19] D. Chen, B. Zhou, V. Koltun, and P. Krähenbühl, "Learning by cheating," in *Proc. Conf. Robot Learn.*, 2020, pp. 66–75.

[20] C. Sun, P. He, R. Wang, and C. Zheng, "Revisiting communication efficiency in multi-agent reinforcement learning from the dimensional analysis perspective," 2025, *arXiv:2501.02888*.

[21] P. S. Chib and P. Singh, "Recent advancements in end-to-end autonomous driving using deep learning: A survey," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 1, pp. 103–118, Jan. 2024.

[22] S. So, J. Petit, and D. Starobinski, "Physical layer plausibility checks for misbehavior detection in V2X networks," in *Proc. 12th Conf. Secur. Privacy Wireless Mobile Netw.*, Jun. 2019, pp. 84–93.

[23] J. Liu et al., "MaskMA: Towards zero-shot multi-agent decision making with mask-based collaborative learning," 2023, *arXiv:2310.11846*.

[24] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. Conf. Robot Learn.*, PMLR, 2017, pp. 1–16.

[25] CARLA. *CARLA Leaderboard*. Accessed: Jan. 16, 2025. [Online]. Available: https://leaderboard.carla.org/leaderboard/

[26] H. Shao, L. Wang, R. Chen, H. Li, and Y. Liu, "Safety-enhanced autonomous driving using interpretable sensor fusion transformer," 2022, *arXiv:2207.14024*.

[27] R. Hao et al., "Research challenges and progress in the end-to-end V2X cooperative autonomous driving competition," 2025, *arXiv:2507.21610*.

[28] C. Zhang, F. Steinhauser, G. Hinz, and A. Knoll, "Occlusion-aware planning for autonomous driving with vehicle-to-everything communication," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 1, pp. 1229–1242, Jan. 2024.

[29] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 627–635.

[30] F. Codevilla, M. Müller, A. López, V. Koltun, and A. Dosovitskiy, "End-to-end driving via conditional imitation learning," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 4693–4700.

[31] T.-Y. Tung, S. Kobus, J. P. Roig, and D. Gündüz, "Effective communications: A joint learning and communication framework for multi-agent reinforcement learning over noisy channels," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2590–2603, Aug. 2021.

[32] J. M. Gimenez-Guzman, I. Leyva-Mayorga, and P. Popovski, "Semantic V2X communications for image transmission in 6G systems," *IEEE Netw.*, vol. 38, no. 6, pp. 48–54, Nov. 2024.

[33] D. Jiang and L. Delgrossi, "IEEE 802.11p: Towards an international standard for wireless access in vehicular environments," in *Proc. IEEE Veh. Technol. Conf. (VTC Spring)*, May 2008, pp. 2036–2040.

[34] S. Chen et al., "Vehicle-to-everything (v2x) services supported by LTE-based systems and 5G," *IEEE Commun. Standards Mag.*, vol. 1, no. 2, pp. 70–76, Feb. 2017.

[35] J. B. Kenney, "Dedicated short-range communications (DSRC) standards in the United States," *Proc. IEEE*, vol. 99, no. 7, pp. 1162–1182, Jul. 2011.

[36] D. Garcia-Roger, E. E. González, D. Martín-Sacristán, and J. F. Monserrat, "V2X support in 3GPP specifications: From 4G to 5G and beyond," *IEEE Access*, vol. 8, pp. 190946–190963, 2020.

[37] Z. Lei, S. Ren, Y. Hu, W. Zhang, and S. Chen, "Latency-aware collaborative perception," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jan. 2022, pp. 316–332.

[38] Y. Yao et al., "Multi-channel based Sybil attack detection in vehicular ad hoc networks using RSSI," *IEEE Trans. Mobile Comput.*, vol. 18, no. 2, pp. 362–375, Feb. 2019.

[39] Q. Zhu, C.-X. Wang, B. Hua, M. Kai, S. Jiang, and M. Yao, *3GPP TR 38.901 Channel Model*. Hoboken, NJ, USA: Wiley, 2021, pp. 1–35.

[40] S. Gyawali, S. Xu, Y. Qian, and R. Q. Hu, "Challenges and solutions for cellular based V2X communications," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 1, pp. 222–255, 1st Quart., 2021.

[41] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, Jun. 2017, pp. 5998–6008.

[42] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, 2006, pp. 850–855.

[43] X. Jia, Z. Yang, Q. Li, Z. Zhang, and J. Yan, "Bench2Drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving," 2024, *arXiv:2406.03877*.

[44] P. Wu, X. Jia, L. Chen, J. Yan, H. Li, and Y. Qiao, "Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2022, pp. 6119–6132. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/286a371d8a0a559281f682f8fbf89834-Paper-Conference.pdf

[45] I. Kotseruba and J. K. Tsotsos, "Understanding and modeling the effects of task and context on drivers' gaze allocation," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jul. 2024, pp. 1337–1344.

[46] H. Yu, Y. Tang, E. Xie, J. Mao, P. Luo, and Z. Nie, "Flow-based feature fusion for vehicle-infrastructure cooperative 3D object detection," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Jan. 2023, pp. 34493–34503.

[47] X. Hu, Z. Huang, A. Huang, J. Xu, and S. Zhou, "A dynamic multi-scale voxel flow network for video prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 6121–6131.

[48] Y. Wang, Z. Gao, M. Long, J. Wang, and P. S. Yu, "PredRNN++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning," in *Proc. 35th Int. Conf. Mach. Learn.*, Jan. 2018, pp. 5123–5132.

[49] C. Tan et al., "Temporal attention unit: Towards efficient spatiotemporal predictive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18770–18782.

[50] Z. Chang et al., "MAU: A motion-aware unit for video prediction and beyond," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, Dec. 2021, pp. 26950–26962. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/e25cfa90f04351958216f97e3efdabe9-Paper.pdf
[51] V. Le Guen and N. Thome, "Disentangling physical dynamics from unknown factors for unsupervised video prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11471–11481.
[52] A. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Dec. 2018, pp. 12689–12697.
[53] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3D object detection and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11784–11793.
[54] P. Wu, S. Chen, and D. N. Metaxas, "MotionNet: Joint perception and motion prediction for autonomous driving based on bird's eye view maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11382–11392.
[55] NVIDIA.(2022). *NVIDIA DRIVE Thor*. Accessed: Jun. 16, 2025. [Online]. Available: https://blogs.nvidia.com/blog/drive-thor/
[56] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video frame synthesis using deep voxel flow," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4473–4481.
[57] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, Dec. 2015, pp. 2017–2025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2015/file/33ceb07bf4eeb3da587e268d663aba1a-Paper.pdf
[58] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with Gumbel-Softmax," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2016, pp. 1–7.
[59] Y. Bengio, N. Léonard, and A. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," 2013, *arXiv:1308.3432*.
[60] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[61] B. Jaeger, K. Chitta, and A. Geiger, "Hidden biases of end-to-end driving models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 8206–8215.
[62] Y. Hu, S. Chen, Y. Zhang, and X. Gu, "Collaborative motion prediction via neural motion message passing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6318–6327.
[63] ASAM. *ASAM OpenDRIVE Standard*. Accessed: Jan. 16, 2025. [Online]. Available: https://www.asam.net/standards/detail/opendrive/
[64] R.-H. Hwang, M. M. Islam, M. A. Tanvir, M. S. Hossain, and Y.-D. Lin, "Communication and computation offloading for 5G V2X: Modeling and optimization," in *Proc. IEEE Global Commun. Conf.*, Dec. 2020, pp. 1–6.
[65] Y. Wang, H. Chen, G. Yin, Y. Mo, N. De Boer, and C. Lv, "Motion state estimation of preceding vehicles with packet loss and unknown model parameters," *IEEE/ASME Trans. Mechatronics*, vol. 29, no. 5, pp. 3461–3472, Oct. 2024.
[66] H. Yu et al., "DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 21361–21370.
[67] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 961–971.

Jiahao Huang (Student Member, IEEE) received the B.E. degree in information engineering from Zhejiang University, Hangzhou, China, in 2023, where he is currently pursuing the Ph.D. degree with the College of Information Science and Electronic Engineering. His research interests include autonomous driving, pragmatic communication, and deep reinforcement learning.

Jianhang Zhu (Member, IEEE) received the B.S. degree in communication engineering from Jilin University, Changchun, China, in 2020. He is currently pursuing the Eng.D. degree with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou. His research interests include graph neural networks, multi-agent reinforcement learning, and edge computing.

Rongpeng Li (Senior Member, IEEE) received the B.E. degree from Xidian University, Xi'an, China, in June 2010, and the Ph.D. degree from Zhejiang University, Hangzhou, China, in June 2015. From August 2015 to September 2016, he was a Research Engineer with the Wireless Communication Laboratory, Huawei Technologies Company Ltd., Shanghai, China. He was a Visiting Scholar with the Department of Computer Science and Technology, University of Cambridge, Cambridge, U.K., from February 2020 to August 2020. He is currently an Associate Professor with the College of Information Science and Electronic Engineering, Zhejiang University. His current research interests include networked intelligence for comprehensive efficiency (NICE).

Zhifeng Zhao (Member, IEEE) received the B.E. degree in computer science, the M.E. degree in communication and information systems, and the Ph.D. degree in communication and information systems from the PLA University of Science and Technology, Nanjing, China, in 1996, 1999, and 2002, respectively. From 2002 to 2004, he was a Post-Doctoral Researcher with Zhejiang University, Hangzhou, China, where his researches were focused on multimedia next-generation networks (NGNs) and softswitch technology for energy efficiency. Currently, he is with the Zhejiang Laboratory, Hangzhou, as the Chief Engineering Officer. His research interests include software-defined networks (SDNs), wireless networks in 6G, computing networks, and collective intelligence. He is the Symposium Co-Chair of ChinaCom in 2009 and 2010. He is the Technical Program Committee (TPC) Co-Chair of the 10th IEEE International Symposium on Communication and Information Technology (ISCIT 2010).

Honggang Zhang was a Professor with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. He was an Honorary Visiting Professor with the University of York, York, U.K., and an International Chair Professor of excellence with the Université Européenne de Bretagne, Supélec, France. He is currently a Professor with the School of Computer Science and Engineering, Macau University of Science and Technology, Macau, China. His research interests include cognitive radio networks, semantic communications, green communications, machine learning, artificial intelligence, intelligent computing, and the Internet of Intelligence.

Dr. Zhang was a co-recipient of the 2021 IEEE Communications Society Outstanding Paper Award and the 2021 IEEE Internet of Things Journal Best Paper Award. He served as the Chair for the Technical Committee on Cognitive Networks of the IEEE Communications Society from 2011 to 2012. He was the Founding Chief Managing Editor of *Intelligent Computing*, a science partner journal. He was the Leading Guest Editor of the Special Issues on Green Communications for IEEE *Communications Magazine*. He served as a Series Editor for *IEEE Communications Magazine* (Green Communications and Computing Networks Series) from 2015 to 2018. He is the Associate Editor-in-Chief of *China Communications*.