SyncMap: Predictive State Bridging for Consistent BEV Sharing in Multi-Vehicle V2X Collaboration

Wei Duan[1], Xiao Wu[1], Yibo Wang[1], Minghao Yu[2], Kai Liu[2], and Jian Zhou[2,*]

***Abstract*—Infrastructure-assisted cooperative perception enables vehicles to achieve beyond-line-of-sight awareness through roadside units (RSUs). However, its application in multi-vehicle collaboration is severely hindered by non-uniform communication delays and the lack of spatial consistency in real-world scenarios. To address these challenges, we propose SyncMap, a lightweight framework that integrates high-definition map-aligned dynamic bird's-eye view (BEV) representation with predictive state bridging (PSB) to achieve spatiotemporal alignment of multi-source perception data under latency. SyncMap constructs a map-referenced, structured BEV representation based on roadside multi-object detection and compensates for heterogeneous communication delays across targets via historical trajectory prediction, thereby generating temporally aligned global environmental snapshots. These snapshots are dynamically generated with respect to the ego vehicle, providing a unified perception input for all connected vehicles within the region. We evaluate SyncMap in CARLA, demonstrating its superior performance in perception consistency. Furthermore, we integrate SyncMap into a multi-vehicle VLM-RL decision-making system. Experimental results show that temporally aligned inputs significantly improve task success rates in cooperative driving, validating the critical role of spatiotemporal consistency in multi-vehicle semantic coordination. This work presents an efficient and scalable shared perception solution for real-world deployment of vehicle-infrastructure cooperative systems.**

I. Introduction

In recent years, autonomous vehicles (AVs) have shown great promise in improving traffic safety and efficiency [1], [2]. [3]However, in complex urban environments—such as dense intersections or mixed-traffic scenarios—occlusions and limited field-of-view hinder reliable perception, especially for vulnerable road users in blind spots. This often leads to conservative behaviors or safety risks when relying solely on ego-vehicle sensing.

Vehicle-to-everything (V2X) cooperation, particularly through roadside units (RSUs) equipped with multimodal sensors, offers a promising solution by providing a "God's-eye view" of the environment [4], [5]. Such infrastructure-assisted perception enables beyond-line-of-sight awareness and supports a shared environmental understanding among connected vehicles, enhancing both safety and coordination efficiency.

[1]Key Laboratory of Metallurgical Equipment and Control Technology, Ministry of Education, Wuhan University of Science and Technology, Wuhan 430081, China (e-mail:duanwei@whu.edu.cn; wuxiao@wust.edu.cn; wangyibo@wust.edu.cn)

[2]Minghao Yu,Kai Liu,and Jian Zhou* are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, 430072, China (e-mail:yuminghao@whu.edu.cn; kailiu@whu.edu.cn; jianzhou@whu.edu.cn)

Despite its potential, practical deployment faces a critical challenge: time delays in perception-communication pipelines. [6], [7] RSU-generated detections often experience non-uniform delays [8] (spanning hundreds of ms to seconds [9]) due to processing and transmission latency, causing spatiotemporal misalignment [10], [11]. This leads to inconsistent state perceptions between vehicles—termed multi-vehicle cognitive clock asynchrony [12]—that break cooperation safety requirements [13], [14].

This issue critically affects semantic decision frameworks like VLM-RL [15], [16], which require coherent structured inputs [17] to align language instructions (e.g., "safely cross while yielding") with perception. Temporal misalignment disrupts this alignment, risking unsafe actions.

To address this, we propose SyncMap, a lightweight, delay-aware RSU-based framework for multi-vehicle cooperative perception (Fig. 1). It delivers on-demand, temporally aligned environmental snapshots to connected vehicles via predictive state bridging (PSB), aligning delayed observations to ensure spatiotemporal consistency. Key contributions include:

- A cooperative alignment framework based on high-definition maps supports vehicle-centric BEV generation and cross-vehicle spatial consistency.
- Design a Predictive State Bridging (PSB) mechanism to compensate for non-uniform V2X delays through lightweight state prediction.
- Experimental validation in Carla shows that the proposed method significantly enhances driving safety in the VLM system.

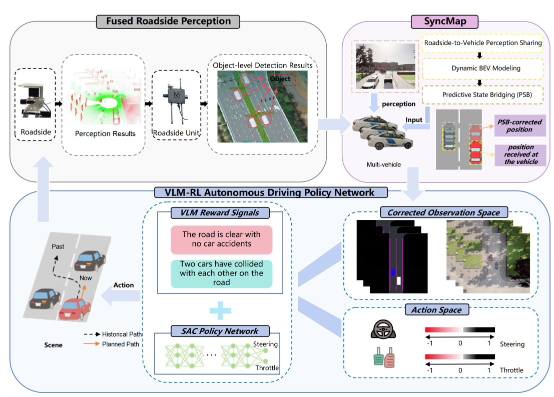

Fig. 1: System overview diagram.

II. Methodology

This paper proposes SyncMap, a lightweight framework for spatiotemporally consistent BEVs via HD map fusion [18], combining HD-map alignment and PSB for semantic decision pipelines [19].

A. HD-Map-Aligned Dynamic BEV Modeling

In vehicle-infrastructure systems, SyncMap ensures spatial consistency for shared perception by fusing RSU perception, HD maps, and vehicle pose to create a vehicle-centric BEV.

Let the vehicle's global pose in $\mathcal{W}$ be $\mathbf{T}_v^{\mathcal{W}} = (\text{lat}_0, \text{lon}_0, \psi_0)$, where $(\text{lat}_0, \text{lon}_0)$ denotes heading. RSU detections of traffic participants are given as $\mathbf{z}_i = (\text{lat}_i, \text{lon}_i, v_i, \psi_i)$ in $\mathcal{W}$. A globally referenced HD map $\mathcal{M}^{\mathcal{W}}$ (OpenDRIVE) [20], [21] provides road topology, lane boundaries, and static semantics.

To generate a vehicle-centric BEV, dynamic states are transformed from $\mathcal{W}$ to body-fixed frame $\mathcal{B}$ via a two-step process: projecting to local Cartesian frame C via tangent plane approximation [22], [23], then aligning with the vehicle's orientation [24].

Let R denote Earth's mean radius. The local coordinates of target i are computed as:

$$\begin{bmatrix} x_i^l \\ y_i^l \\ 1 \end{bmatrix} = \mathbf{T}_{\text{proj}} \cdot \begin{bmatrix} \Delta\text{lon}_i \\ \Delta\text{lat}_i \\ 1 \end{bmatrix}, \quad \mathbf{T}_{\text{proj}} = \begin{bmatrix} R\cos(\text{lat}_0) & 0 & 0 \\ 0 & R & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

Let $\Delta\text{lon}_i = \text{lon}_i - \text{lon}_0$, $\Delta\text{lat}_i = \text{lat}_i - \text{lat}_0$ to minimize high-latitude longitude distortion and preserve geometric accuracy [25].

Next, the local coordinates (x_i^l, y_i^l) are rotated into $\mathcal{B}$ using the SE(2) transformation:

$$\mathbf{T}_{C\to\mathcal{B}} = \begin{bmatrix} \cos\psi_0 & -\sin\psi_0 & 0 \\ \sin\psi_0 & \cos\psi_0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The full transformation into $\mathcal{B}$ is then:

$$\mathbf{p}_i^b = \mathbf{T}_{C\to\mathcal{B}} \cdot \mathbf{T}_{\text{proj}} \cdot \begin{bmatrix} \Delta\text{lon}_i \\ \Delta\text{lat}_i \\ 1 \end{bmatrix}, \quad \psi_i^b = \psi_i - \psi_0.$$

Transformed dynamic states are discretized as a BEV tensor $\mathcal{B} \in \mathbb{R}^{H\times W\times C}$ with occupancy, velocity, heading, and category. Static map elements (e.g., lane markings) from $\mathcal{M}^{\mathcal{W}}$ are projected into $\mathcal{B}$ via the same transformation [26], forming a unified dynamic-static representation [19], [27].

This sensor-agnostic BEV uses RSU perception and HD maps to provide a standardized model for cooperative driving.

B. Predictive State Bridging

HD-map-aligned BEVs ensure spatial consistency, but non-uniform V2X delays (500–1000 ms) cause outdated roadside perception, leading to misalignment and fragmented trajectories that degrade planning and performance [28]. This discrepancy is illustrated in Fig. 2, which compares the delayed semantic bird's-eye view (BEV) with the original BEV.

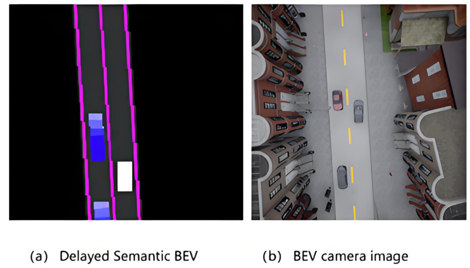

Fig. 2: Comparison between the delayed semantic BEV and the original BEV image.

We introduce PSB, a lightweight method that aligns delayed perception with current decisions via forward prediction, using motion history to generate temporally aligned snapshots .

PSB avoids complex black-box predictors [11] by leveraging vehicle motion continuity and physical predictability for short-term delays ($\Delta t < 2$ s) [29].

PSB uses a Kalman Filter (KF) with a Constant Velocity (CV) model, providing optimal low-cost state estimation ideal [16], [30] for edge deployment on RSUs under linear-Gaussian assumptions.

Let the state of target O_i at detection time t_{det} be:

$$\mathbf{x}_i = [p_x, p_y, v_x, v_y, \theta]^T \tag{1}$$

where (p_x, p_y) is position, (v_x, v_y) velocity, and θ heading. This state is received at t_{now}, with delay $\Delta t = t_{\text{now}} - t_{\text{det}}$.

PSB propagates the state forward using the CV model:

$$\mathbf{x}_i(t_{\text{now}}) = \mathbf{F}\mathbf{x}_i(t_{\text{det}}) + \mathbf{w}, \quad \mathbf{w} \sim \mathcal{N}(0, \mathbf{Q}) \tag{2}$$

and $\mathbf{Q}$ models process noise (e.g., acceleration disturbances).

In practice, PSB performs only the prediction step:

$$\hat{\mathbf{x}}_i^-(t_{\text{now}}) = \mathbf{F}\mathbf{x}_i(t_{\text{det}}) \tag{3}$$

Without new measurements, the delayed state is extrapolated to t_{now}, aligning past observations to the present.

The PSB-corrected state $\hat{\mathbf{x}}_i(t_{\text{now}})$ is integrated into SyncMap's dynamic BEV, reflecting current estimates and ensuring spatiotemporal coherence [30], [31].

III. Experimental Evaluation

To assess perception latency's impact on language-guided driving policies and validate the PSB compensation mechanism, we developed a CARLA-based cooperative framework and conducted ablation studies under non-uniform latency to evaluate policy performance.

A. Experimental Setup

Experiments used CARLA's Town02 with simulated roadside BEV perception mimicking real RSU outputs. A 500–1000 ms stochastic delay was applied to simulate V2X latency, causing temporal misalignment. Policies trained without delay were evaluated under latency using one of three inference-time strategies:

TABLE I: Performance comparison of CV, Misuse, and PSB-Compensated policies during the inference phase under non-uniform perception delays of 500-1000 ms (mean ± standard deviation, after 1M-step training).

Policy	Steps	RC ↑	TD ↑	CS ↓	CPS ↓	SR ↑
CV Policy		0.71±0.06	1510.4±177.5	1.73±1.18	0.002562±0.000714	0.43±0.06
Misuse Policy	100000	0.71±0.06	1510.4±177.52	1.73±1.18	0.002562±0.000714	0.43±0.06
PSB-Compensated Policy		**0.85 ± 0.05**	**1811.4 ± 127.23**	1.77±2.53	**0.001241 ± 0.000755**	**0.7 ± 0.1**
CV Policy		0.75±0.02	1563.3±76.15	**1.88 ± 0.82**	0.001961±0.000295	0.57±0.06
Misuse Policy	200000	0.72±0.04	1487.2±140.89	3.10±1.45	0.002637±0.00048	0.5±0.1
PSB-Compensated Policy		**0.80 ± 0.09**	**1696.2 ± 234.2**	2.68±0.31	**0.001729 ± 0.000318**	**0.6 ± 0.0**
CV Policy		0.76±0.09	1607.8±250.84	1.31±1.33	0.001826±0.000764	0.53±0.06
Misuse Policy	300000	0.74±0.06	1588.9±85.78	2.20±0.42	0.001875±0.000695	0.47±0.06
PSB-Compensated Policy		**0.88 ± 0.03**	**1911.7 ± 79.53**	**0.84 ± 0.79**	**0.00118 ± 0.000447**	**0.8 ± 0.0**
CV Policy		0.76±0.09	1861.3±219.46	1.61±1.54	0.000945±0.000753	0.73±0.12
Misuse Policy	400000	0.80±0.06	1708.1±139.45	2.46±2.17	0.001528±0.000595	0.67±0.06
PSB-Compensated Policy		**0.93 ± 0.04**	**1957.4 ± 120.19**	**0.13 ± 0.18**	**0.000662 ± 0.000525**	**0.8 ± 0.1**
CV Policy		**0.93 ± 0.06**	**1973.0 ± 129.49**	0.42±0.36	0.000896±0.000693	0.77±0.12
Misuse Policy	500000	0.81±0.01	1751.3±49.71	3.61±1.27	0.001612±0.0002	0.57±0.06
PSB-Compensated Policy		0.91±0.07	1967.2±132.69	**0.30 ± 0.33**	**0.000702 ± 0.000544**	**0.87 ± 0.12**

- Misuse Policy: Directly uses delayed BEV inputs without compensation.
- CV Policy: Applies a Constant Velocity model to extrapolate target states linearly.
- PSB-Compensated Policy: Uses a Predictive State Bridging module for motion-aware forward prediction.

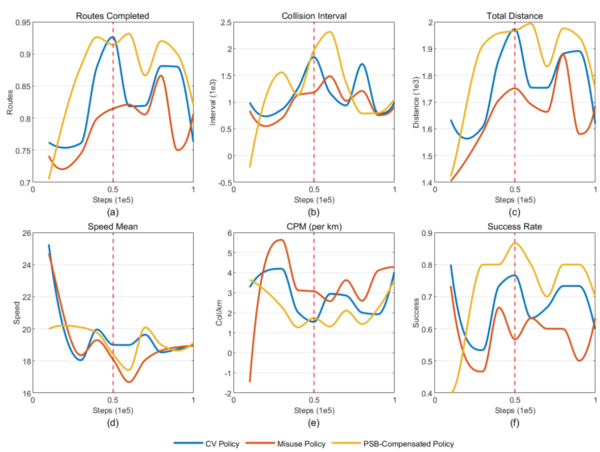

Fig. 3: Comparison of periodic task performance across different strategies during the inference process.

PSB showed greater adaptability across traffic densities, maintaining higher success and robustness in dense traffic, while Misuse and CV policies faltered. This validates PSB's effectiveness under latency (Fig. 4).

Models were trained for 1×10^6 steps using VLM-RL-SAC [13], evaluated every 1×10^5 steps. Performance was assessed via efficiency (RC, TD, SR) and safety (TCF, CS, ICT) metrics over three trials on 10 routes for statistical reliability.

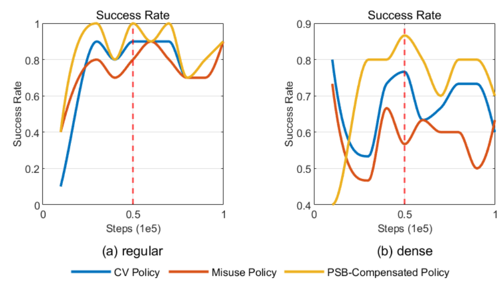

Fig. 4: Comparison of success rates among the three policies under varying traffic densities.

B. Simulation Results and Analysis

All policies converged by 500k steps. The Misuse Policy showed high instability, with SR dropping to 0.57 and poor safety. The CV Policy improved SR to 0.77 but exhibited oscillations, highlighting limitations of linear prediction.

The PSB-Compensated Policy achieved superior, stable performance: SR reached 0.87, RC 0.91, and collision probability dropped significantly(Fig. 3), demonstrating consistent effectiveness against non-uniform delays.

IV. Conclusion

This paper addresses spatiotemporal misalignment from non-uniform communication delays in vehicle-infrastructure cooperation, proposing SyncMap—a lightweight framework combining HD-map-aligned dynamic BEVs with Predictive State Bridging (PSB) for consistent environmental modeling. By generating temporally aligned global snapshots, SyncMap bridges delayed perception to semantic decision-making. Validated with VLM-RL policies, it improves task success rates and enables deployable, language-guided multi-vehicle cooperation. Future work integrates VLM-RL with real-world deployment and SyncMap's roadside services to evaluate practical performance and robustness.

References

[1] Li Z, Wang W, Li H, et al. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

[2] Z. Wu, G. Chen, Y. Gan, L. Wang, and J. Pu, "MVFusion: Multi-view 3D object detection with semantic-aligned radar and camera fusion," in Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), London, United Kingdom, 2023, pp. 2766–2773.

[3] J. Zhang, J. Huang, S. Jin and S. Lu, "Vision-Language Models for Vision Tasks: A Survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 8, pp. 5625-5644, Aug. 2024, doi: 10.1109/TPAMI.2024.3369699.

[4] Xu R, Xiang H, Tu Z, et al. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 107-124.

[5] Yu H, Yang W, Ruan H, et al. V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 5486-5495.

[6] Chen, Q., Ma, X., Tang, S., Guo, J., Yang, Q., Fu, S.: F-cooper: Feature based co operative perception for autonomous vehicle edge computing system using 3d point clouds. In: Proceedings of the 4th ACM/IEEE Symposium on Edge Computing. pp. 88–100 (2019).

[7] Y. Yuan, H. Cheng, M. Y. Yang, and M. Sester, "Generating evidential BEV maps in continuous driving space," 2023, arXiv:2302.02928.

[8] Y. He, B. Wu, Z. Dong, J. Wan and W. Shi, "Towards C-V2X Enabled Collaborative Autonomous Driving," in IEEE Transactions on Vehicular Technology, vol. 72, no. 12, pp. 15450-15462, Dec. 2023, doi: 10.1109/TVT.2023.3299844.

[9] B. Gao, J. Liu, H. Zou, J. Chen, L. He, and K. Li, "Vehicle-road-cloud collaborative perception framework and key technologies: A review," IEEE Transactions on Intelligent Transportation Systems, 2024.

[10] Huang T, Liu J, Zhou X, et al. V2X cooperative perception for autonomous driving: Recent advances and challenges[J]. arXiv preprint arXiv:2310.03525, 2023.

[11] Shi J, Zhao J, Zhuo L, et al. V2V Cooperative Perception With Adaptive Communication Loss for Autonomous Driving[J]. IEEE Transactions on Intelligent Transportation Systems, 2025.

[12] Y. Fu, C. Li, F. R. Yu, T. H. Luan, and Y. Zhang, "A selective federated reinforcement learning strategy for autonomous driving," IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 2, pp. 1655–1668, February 2023.

[13] C. Xiang, C. Feng, X. Xie, B. Shi, H. Lu, Y. Lv, M. Yang, and Z. Niu, "Multi-sensor fusion and cooperative perception for autonomous driving: A review," IEEE Intelligent Transportation Systems Magazine, vol. 15, no. 5, pp. 36–58, 2023.

[14] Y. Liu, Q. Huang, R. Li, X. Chen, Z. Zhao, S. Zhao, Y. Zhu, and H. Zhang, "Rethinking collaborative perception from the spatial-temporal importance of semantic information," 2023, arXiv:2307.16517.

[15] Zhou X, Liu M, Yurtsever E, et al. Vision language models in autonomous driving: A survey and outlook[J]. IEEE Transactions on Intelligent Vehicles, 2024.

[16] Yu H, Yang W, Hao R, Wang C, Zhong J, Luo P, et al. DriveE2E: Closed-Loop Benchmark for End-to-End Autonomous Driving through Real-to-Simulation. arXiv preprint arXiv:2509.23922v1. 2025.

[17] Li Y, Ren S, Wu P, et al. Learning distilled collaboration graph for multi-agent perception[J]. Advances in Neural Information Processing Systems, 2021, 34: 29541-29552.

[18] Li Y, Ge Z, Yu G, et al. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection[C]//Proceedings of the AAAI conference on artificial intelligence. 2023, 37(2): 1477-1485.

[19] Sontakke, S. A., Zhang, J., Arnold, S. M. R, et al. RoboCLIP: One Demonstration is Enough to Learn Robot Policies. Neural Information Processing Systems. https://doi.org/10.48550/arxiv.2310.07899.

[20] Guo Y, Zhou J, Dong Q, et al. Refined high-definition map model for roadside rest area[J]. Transportation Research Part A: Policy and Practice, 2025, 195: 104463.

[21] Zhou J, Yu M, Guo Y, et al. A high-definition map architecture for transportation digital twin system construction[J]. International Journal of Applied Earth Observation and Geoinformation, 2025, 144: 104822.

[22] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger and H. Li, "End-to-End Autonomous Driving: Challenges and Frontiers," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 12, pp. 10164-10183,Dec. 2024, doi: 10.1109/TPAMI.2024.3435937.

[23] G. Guo and S. Zhao, "3D multi-object tracking with adaptive cubature Kalman filter for autonomous driving," IEEE Transactions on Intelligent Vehicles, vol. 8, no. 1, pp. 512–519, February 2023.

[24] Huang Z, Sheng Z, Qu Y, et al. Vlm-rl: A unified vision language models and reinforcement learning framework for safe autonomous driving[J]. Transportation Research Part C: Emerging Technologies, 2025, 180: 105321.

[25] Zhou, Z., Cai, T., Zhao, S. Z, et al. AutoVLA: A Vision-Language-Action Model for End-to-End Autonomous Driving with Adaptive Reasoning and Reinforcement Fine-Tuning. http://arxiv.org/abs/2506.13757v1.

[26] Y. Fuchida, T. Taniguchi, T. Takano, T. Mori, K. Takenaka and T. Bando, "Driving word2vec: Distributed semantic vector representation for symbolized naturalistic driving data," 2016 IEEE Intelligent Vehicles Symposium (IV), Gothenburg, Sweden, 2016, pp. 1313-1320.

[27] Sima, C., Renz, K., Chitta, K, et al. DriveLM: Driving with Graph Visual Question Answering [Online post]. European Conference on Computer Vision. https://doi.org/10.48550/arxiv.2312.14150.

[28] Jiang Y, Zhan G, Lan Z, Liu C, Cheng B, Li SE. A Reinforcement Learning Benchmark for Autonomous Driving in General Urban Scenarios. IEEE Transactions on Intelligent Transportation Systems. 2023;25(5):4335-45.

[29] Salzmann T, Ivanovic B, Chakravarty P, et al. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data[C]//European Conference on Computer Vision. Cham: Springer International Publishing, 2020: 683-700.

[30] Hao R, Jing B, Yu H, Nie Z. StyleDrive: Towards Driving-Style Aware Benchmarking of End-To-End Autonomous Driving. arXiv preprint arXiv:2506.23982v2. 2025.

[31] Gao H, He W, Liu T, Xiong J, Shuai B, Chen C, et al. Explicit Nonlinear Control for Optimal Trajectory Tracking of Autonomous Vehicles. In: Proceedings of the IEEE International Conference on Robotics and Automation. 2025. p. 1094-100.

[32] Feng D, Harakeh A, Waslander S L, et al. A review and comparative study on probabilistic object detection in autonomous driving[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(8): 9961-9980.