# V2X-BEVDet4D: A V2X-Enhanced 360-Degree 4D Perception for Occlusion-Free Autonomous Driving

Faisal Hawlader, Gamal Elghazaly, and Raphaël Frank
Interdisciplinary Center for Security, Reliability and Trust (SnT)
University of Luxembourg, 29 Avenue J.F Kennedy, L-1855 Luxembourg
firstname.lastname@uni.lu

*Abstract*—**Autonomous vehicles face critical perception challenges, when objects such as pedestrians or vehicles are hidden behind obstacles. Transformer-based models, such as BEVDet4D, achieve robust 360-Degree 4D object detection by leveraging multi-camera inputs. However, their performance significantly deteriorates for occluded objects. To overcome this limitation, we propose V2X-BEVDet4D, a cooperative perception framework built on BEVDet4D. It enhances the base model by fusing its outputs with object detections from roadside infrastructure, transmitted via ETSI-compliant Cooperative Perception Messages (CPMs) over ITS-G5. Results show up to a 183% improvement in nuScenes Detection Score (NDS) for objects located 100 m away from the vehicle, precisely where standalone BEVDet4D detection is most limited. Preliminary results also demonstrate a CPM transmission latency of 3.44 ms (±1.3 ms std), confirming the real-time feasibility. To our knowledge, this is the first framework to fuse ETSI-compliant V2X messages into a BEV-based 360-Degree 4D object detection pipeline, enabling temporal consistency across frames.**

*Index Terms*—**Cooperative Perception; V2X; ITS-G5.**

## I. Introduction

Reliable perception is fundamental to the safe operation of autonomous vehicles. However, current systems struggle with occlusions, where pedestrians, cyclists, or vehicles are hidden behind static or dynamic obstacles, particularly in dense urban environments. Recent transformer-based models such as BEVFormer [1] use six-camera surround views to generate 360-degree bird's-eye view (BEV) representations for 3D object detection. BEVDet4D [2] extends this by enabling 4D detection, i.e., temporally consistent 3D localization across frames. Despite their effectiveness, these models rely solely on cameras mounted on the vehicle, making them inherently vulnerable to occlusions. A promising solution is Vehicle-to-Everything (V2X) cooperative perception, in which vehicles and infrastructure share object detections. Previous work has explored V2X-based data sharing [3]. However, integrating Cooperative Perception Messages (CPMs) into BEV-based transformer models remains an open challenge. Existing methods often require complex retraining or fine-tuning of base models, lack compliance with ETSI CPM standards, or introduce substantial computational overhead. To address these limitations, we introduce V2X-BEVDet4D, the first framework to integrate ETSI-compliant CPMs into a BEV-based 4D detection without requiring model retraining. Our approach performs lightweight post-inference fusion using standardized V2X messages, enabling real-time, occlusion-resilient perception with minimal system overhead.

## II. Methodology

### A. System Overview

V2X-BEVDet4D is a modular framework that enhances BEV-based 4D object detection by fusing CPMs received from roadside infrastructure. An overview of the system is illustrated in Figure 1. The end-to-end pipeline operates in six stages: **(1) Multi-view images:** the ego vehicle captures the surrounding environment using six synchronized cameras. **(2) Feature extraction:** a backbone network encodes the images into perspective-aligned feature maps. **(3) BEV encoder:** the extracted features are projected into a unified BEV representation. **(4) Ego-view detection:** BEVDet4D performs 4D object detection on the BEV feature map. **(5) CPM transmission:** roadside infrastructure transmits ETSI-compliant CPMs over ITS-G5. **(6) Post-inference fusion:** the ego vehicle fuses BEVDet4D outputs with received CPM detections to produce a complete 360-Degree 4D perception output. This architecture enables occlusion-resilient perception without modifying the BEVDet4D model, thereby ensuring compatibility with existing pipelines.

### B. Hardware Configuration

To reflect real-world deployment constraints, the vehicle executes stages 1, 2, 3, 4 and 6, as detailed in Subsection II-A, on a GeForce GTX 1650 GPU. The roadside unit is dedicated solely to stage 5 (CPM transmission). Both the vehicle and roadside unit are equipped with YoGoKo ITS-G5 transceivers operating at 5.9 GHz, with all parameters configured as described in our previous work [3].

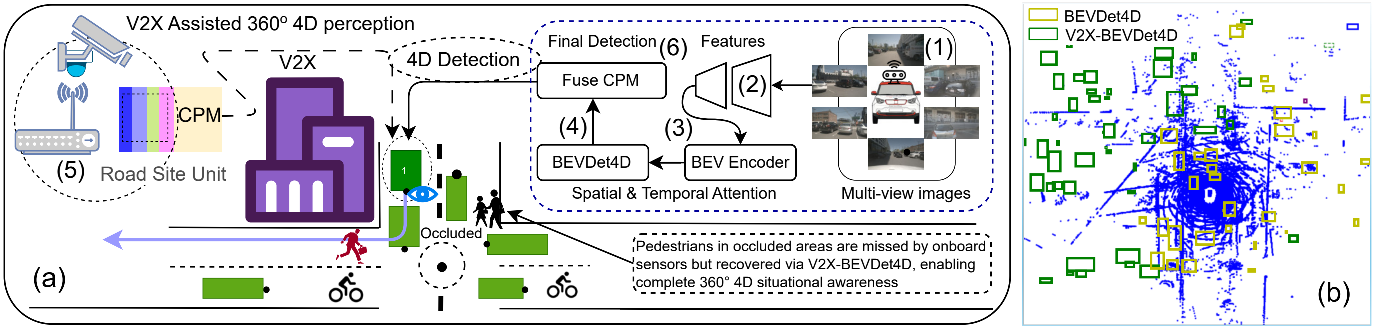
### C. Dataset and Model

We use the nuScenes dataset [4], a standard benchmark for object detection that provides 1,000 annotated urban driving scenes with synchronized sensor data recorded at 2 Hz. We use BEVDet4D [2] as the base perception model for the vehicle. For this work, we use the pretrained model without any additional fine-tuning.

### D. CPM Transmission and Fusion Strategy

To emulate ETSI-compliant CPMs, we encode nuScenes ground-truth annotations for perception evaluation. Separately, we evaluate the network transmission performance of CPMs using real ITS-G5 devices along a fixed route in Luxembourg[1].

[1] **Test route:** http://g-o.lu/3/Rf1L



Pedestrians in occluded areas are missed by onboard sensors but recovered via V2X-BEVDet4D, enabling complete 360° 4D situational awareness

Fig. 1: (a) System overview of V2X-BEVDet4D with CPM fusion received from a roadside unit over ITS-G5. (b) Qualitative results: BEVDet4D misses occluded objects, while V2X-BEVDet4D recovers them through CPM fusion.

Each CPM includes object class, 3D position, heading and timestamp, structured according to ETSI TS 103 324. Post-inference fusion is performed without modifying BEVDet4D. CPM detections are transformed into the vehicle's coordinate frame using synchronized poses and matched to BEVDet4D outputs using (i) spatial proximity, (ii) class label consistency. Unmatched objects are appended as additional detections.

## III. Experiments & Results

### A. *CPM Transmission Latency (Real-World Test)*

We evaluated the performance of CPM delivery on ITS-G5. As shown in Figure 2, the system maintains low latency, with an average of 3.44 ms, a maximum of 11.66 ms, and a standard deviation of 1.23 ms. The latency remains stable up to 120 m, with packet loss remaining below 3%. However, packet loss increases notably beyond 150 m, and jitter becomes more pronounced past 100 m. These results confirm that CPM-based cooperative perception can operate within urban driving scenarios, making our framework deployable without additional network optimization.

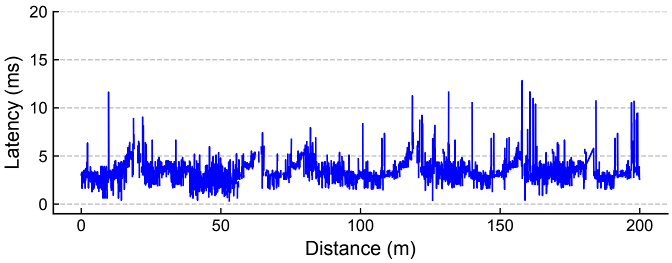


Fig. 2: CPM transmission latency vs. distance between the vehicle and roadside unit. Latency remains low and stable under 120 m, with increased losses observed beyond 150 m.

### B. *Perception Quality and Detection Range (nuScenes)*

We compare the baseline BEVDet4D model with our proposed V2X-BEVDet4D system. Fusion is applied selectively, only when objects are located beyond 25 m and are not detected by the baseline model.

Table I shows that V2X-BEVDet4D significantly improves detection performance at increasing distances. At 25 m, the NDS improves by 52.08%, reaching 183.00% at 100 m. These results highlight the critical value of V2X fusion in mitigating occlusion and extending detection range, particularly in complex urban environments.

| **Distance (m)** | **25** | **50** | **75** | **100** |
|---|---|---|---|---|
| BEVDet4D | 0.48 | 0.31 | 0.21 | 0.15 |
| V2X-BEVDet4D | **0.73** | **0.62** | **0.57** | **0.43** |
| Improvement (%) | ↑52.08% | ↑100.00% | ↑171.42% | ↑183.00% |

TABLE I: NDS comparison for BEVDet4D and V2X-BEVDet4D across object distances from the ego vehicle.

## IV. Conclusion

This preliminary work demonstrates how standardized V2X communication can meaningfully extend the capabilities of modern BEV-based perception. By selectively fusing cooperative detections via ETSI-compliant CPMs, V2X-BEVDet4D bridges a critical gap in current AVs, the inability to see beyond field of view. Our approach is unique in that it does not require model retraining, and operates under real-world communication constraints using ITS-G5. The system reliably recovers occluded objects, making real-world deployment feasible. More broadly, this study shows that intelligent post-inference V2X fusion can elevate perception performance even with limited infrastructure. Looking ahead, our method offers a scalable foundation for cooperative perception, infrastructure-assisted driving, and city-scale situational awareness. As ongoing work, we plan to extend the system to multiple roadside units and evaluate it in more realistic scenarios, including complex urban intersections.

## References

[1] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers," *arXiv preprint arXiv: 2203.17270*, Mar. 2022. [Online]. Available: http://arxiv.org/abs/2203.17270

[2] J. Huang and G. Huang, "Bevdet4d: Exploit temporal cues in multi-camera 3d object detection," *arXiv preprint arXiv:2203.17054*, 2022.

[3] F. Hawlader, F. Robinet, G. Elghazaly, and R. Frank, "Cloud-assisted 360-degree 3d perception for autonomous vehicles using v2x communication and hybrid computing," in *2025 20th Wireless On-Demand Network Systems and Services Conference (WONS)*, 2025, pp. 1–8.

[4] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," *arXiv preprint arXiv:1903.11027*, 2019.