SparseCoop: Cooperative Perception with Kinematic-Grounded Queries

Jiahao Wang[1], Zhongwei Jiang[2], Wenchao Sun[1], Jiaru Zhong[3], Haibao Yu[4], Yuner Zhang[5], Chenyang Lu[1], Chuang Zhang[1], Lei He[1], Shaobing Xu[1] *, Jianqiang Wang[1] *

[1]Tsinghua University
[2]Nanyang Technological University
[3]The Hong Kong Polytechnic University
[4]The University of Hong Kong
[5]University of Pennsylvania

Abstract

Cooperative perception is critical for autonomous driving, overcoming the inherent limitations of a single vehicle, such as occlusions and constrained fields-of-view. However, current approaches sharing dense Bird's-Eye-View (BEV) features are constrained by quadratically-scaling communication costs and the lack of flexibility and interpretability for precise alignment across asynchronous or disparate viewpoints. While emerging sparse query-based methods offer an alternative, they often suffer from inadequate geometric representations, suboptimal fusion strategies, and training instability. In this paper, we propose SparseCoop, a fully sparse cooperative perception framework for 3D detection and tracking that completely discards intermediate BEV representations. Our framework features a trio of innovations: a kinematic-grounded instance query that uses an explicit state vector with 3D geometry and velocity for precise spatio-temporal alignment; a coarse-to-fine aggregation module that effectively integrates information from both matched and unmatched instances; and a cooperative instance denoising task that provides stable, abundant supervision to accelerate and stabilize training. Experiments on V2X-Seq and Griffin datasets show SparseCoop achieves state-of-the-art performance. Notably, it delivers this performance with superior computational efficiency and a highly competitive transmission cost, while showing remarkable robustness to real-world challenges like communication latency.

Code — https://github.com/wang-jh18-SVM/SparseCoop

1 Introduction

A robust perception system is fundamental for autonomous driving, but systems confined to a single vehicle are inherently limited by sensor field-of-view constraints, long-range sensing fall-off, and severe occlusions. These challenges create a critical bottleneck, preventing autonomous systems from achieving the comprehensive awareness necessary for safe deployment in complex scenarios. To overcome these individual limitations, cooperative perception has emerged as a key paradigm. By enabling information exchange between multiple agents—such as vehicles (V2V), infrastructure (V2I), and drones (V2D)—it creates a collective sensing system with capabilities far beyond a single vehicle's. Among various strategies, feature-level fusion is widely researched for its effective balance between preserving rich information and maintaining manageable communication bandwidth (Gao et al. 2024; Han et al. 2023).

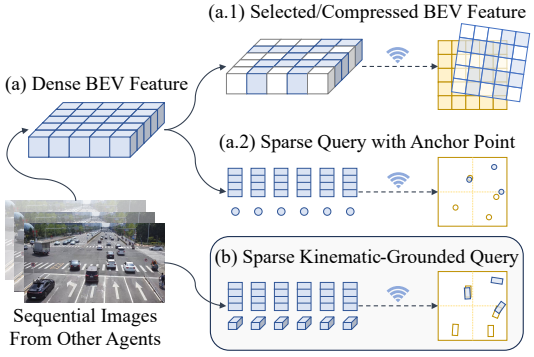


Figure 1: Comparison of cooperative perception pipelines. Existing methods (a) are bottlenecked by the intermediate dense BEV features, even when these features are selected, compressed (a.1) or encoded into sparse queries anchored to reference points (a.2). In contrast, SparseCoop (b) is a fully sparse paradigm that bypasses BEV step, directly extracting queries grounded by rich state vectors from image features.

The central challenge in feature-level fusion lies in finding a representation that is both efficient for transmission and expressive for robust fusion. The dominant approach has been to share dense Bird's-Eye-View (BEV) feature maps, which provide a unified spatial grid (Xu et al. 2022b, 2023; Hu et al. 2023), as shown in Figure 1(a.1). However, this paradigm suffers from fundamental drawbacks: it creates prohibitive communication and computational costs that scale quadratically with perception range, and its abstract scene-level features are difficult to align precisely across agents, especially under temporal asynchrony. To address these inefficiencies, recent research has shifted towards sparse, query-based methods that represent the scene

*Corresponding authors: Jianqiang Wang and Shaobing Xu (wjqlws@tsinghua.edu.cn, shaobxu@tsinghua.edu.cn)

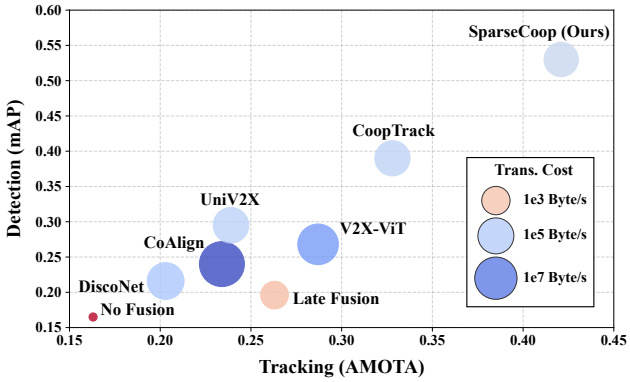


Figure 2: Performance comparison on V2X-Seq dataset. The X-axis and Y-axis represent perception metrics, while the bubble size and color encode the transmission cost on a logarithmic scale.

with a compact set of object-centric queries.

While this emerging sparse paradigm promises greater efficiency and interpretability, it introduces its own set of unresolved challenges. A primary limitation lies in the geometric representation of the queries themselves. Often anchored to just a single reference point (Fan et al. 2024; Yu et al. 2025), these queries lack the expressiveness required to handle the significant viewpoint rotations and temporal shifts inherent in real-world cooperative scenarios. Moreover, their instance fusion strategies are often suboptimal, struggling to effectively integrate information from different agents without losing crucial, unique observations. The training process is also frequently unstable and inefficient bacause of the limited overlapping viewpoints between agents and sparse supervision signals. Finally, many of these methods still depend on a dense BEV component (Liu et al. 2025; Wang et al. 2025b; Yuan et al. 2025; Wang et al. 2025c; Zhong et al. 2025), as shown in Figure 1(a.2), thereby inheriting its computational scaling limitations.

In this paper, we propose SparseCoop, a fully sparse cooperative detection and tracking framework that directly confronts these challenges. Our framework is built on a suite of innovations designed for a robust, instance-centric approach: To address the representation problem, we introduce the kinematic-grounded query, which uses an explicit state vector encoding 3D geometry and velocity, as shown in Figure 1(b), for robust spatio-temporal alignment. To solve the fusion challenge, we design a coarse-to-fine aggregation module that effectively balances information from both matched and unmatched instances. Finally, to overcome training instability, we introduce a cooperative instance denoising task that provides a stable and abundant source of supervision.

Our work makes several key contributions:

- We propose SparseCoop, a novel, fully sparse cooperative perception framework that operates directly on temporal instance-level representations, eliminating the computational bottlenecks of dense BEV maps.
- We introduce a trio of innovations to enable this: a Kinematic-Grounded Association module for precise alignment, a Coarse-to-Fine Aggregation module for effective fusion, and a Cooperative Denoising strategy that stabilizes training.
- Our method achieves state-of-the-art (SOTA) detection and tracking performance on both the V2I V2X-Seq and V2D Griffin datasets with low communication and computation costs and strong latency robustness.

2 Related Work

Cooperative perception methods are generally categorized into three fusion strategies (Caillot et al. 2022; Han et al. 2023; Gao et al. 2024): early, intermediate, and late fusion. Early fusion (Valiente et al. 2019; Chen et al. 2019; Arnold et al. 2022) directly integrates raw sensor data, offering the potential for the highest performance by retaining rich information, but at the cost of significant bandwidth overhead. In contrast, late fusion minimizes communication by exchanging final detection results, making cooperation more interpretable but highly dependent on the accuracy of individual perception and coordinate transformation.

Intermediate fusion offers a compromise by transmitting certain network features, with BEV feature maps being the most common representation (Wang et al. 2020; Li et al. 2021; Hu et al. 2023; Lu et al. 2023). Due to the high bandwidth required for dense BEV maps, several strategies have been proposed to reduce transmission costs, such as selecting salient regions (Hu et al. 2022, 2024; Yuan et al. 2023) or applying feature compression techniques (Wang et al. 2020; Xu et al. 2022a; Yin et al. 2024; Wang et al. 2024). However, these scene-level features still lack the flexibility to handle significant spatio-temporal misalignments. For instance, in V2D scenarios with large viewpoint differences, the flattened BEV representation struggles with scale changes and distortion (Wang et al. 2026). What's more, addressing large temporal asynchrony often requires transmitting an additional dense flow map (Yu et al. 2023a), further increasing communication overhead.

To overcome the limitations of dense representations, recent studies have begun to explore the transmission of sparse, instance-level queries (Chen, Shi, and Jia 2023; Fan et al. 2024). These methods, often built on Transformer architectures (Vaswani et al. 2017; Wang et al. 2022; Liu et al. 2022, 2023), significantly reduce communication costs by only sharing a compact set of object-centric queries. However, this paradigm faces several challenges. First, the prevalent query representation—often just a latent feature anchored to a reference point (Fan et al. 2024; Zhong et al. 2024, 2025; Liu et al. 2025; Yu et al. 2025; Wang et al. 2025a)—lacks the explicit structure needed for robust spatio-temporal alignment against significant viewpoint and temporal disparities. Second, existing instance fusion strategies are often suboptimal. Simpler approaches use linear networks to fuse matched instance pairs (Yu et al. 2025; Zhong et al. 2025), which is efficient but has limited expressive power. Other methods employ global (Chen, Shi, and Jia 2023; Fan et al. 2024; Zhong et al. 2024) or se-

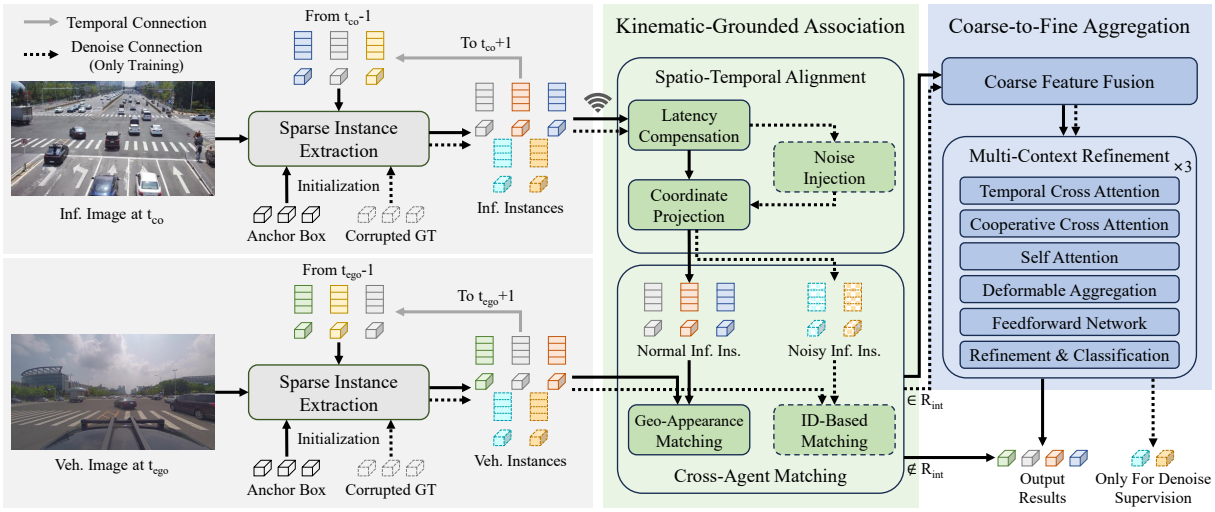


Figure 3: An overview of the SparseCoop framework. Each agent independently performs Sparse Instance Extraction. The ego-vehicle then uses the proposed Kinematic-Grounded Association and Coarse-to-Fine Aggregation modules to fuse transmitted instances with its own. Cooperative Instance Denoising (dashed lines) is only active during training to stabilize convergence.

lective masked attention (Wang et al. 2025a,c) mechanisms across cooperative instances, while powerful, neglecting to fully leverage additional, fine-grained context from the ego-vehicle's own sensor data during the fusion process. Finally, these methods often struggle with training instability. Due to differing agent viewpoints and occlusions, the number of co-observed objects available for supervision is inherently low. This scarcity of positive training samples is compounded by the strict one-to-one matching process required during training, especially in early training stages, which can hinder model convergence.

3 Method

3.1 Preliminaries: Task and Query Definition

We adopt the standard task formulation from prominent cooperative perception benchmarks (Yu et al. 2023b; Wang et al. 2026). The objective is to generate a temporally consistent set of 3D tracked objects $\{\hat{o}_i\}$. An ego-agent at its current timestamp t_{ego} utilizes its own sensor data alongside information shared by a cooperative agent from a potentially asynchronous timestamp t_{co}, where $t_{\text{co}} \leq t_{\text{ego}}$ due to communication latency. The final output is generated within the ego-agent's predefined region of interest (ROI), R_{ego}.

The fundamental unit of representation, transmission, and fusion in our framework is the **Kinematic-Grounded Query (KGQ)**. A KGQ is an *instance* defined as a pair $\{\mathcal{F}, \mathcal{S}\}$, where $\mathcal{F}$ is a latent feature vector encoding semantic information and $\mathcal{S}$ is its explicit 11-dimensional state vector, defined as:

$$\mathcal{S} = (x, y, z, l, w, h, \sin(\theta), \cos(\theta), v_x, v_y, v_z) \quad (1)$$

This state vector describes the object's 3D position, dimensions, heading angle, and velocity. This rich, explicit representation is a key distinction from prior works that rely on simpler geometric inputs like a single reference point, and it is central to our robust alignment and fusion strategy.

With a road side unit (RSU) as an example of a cooperative agent, the overall data flow of our framework is illustrated in Figure 3. Each agent first independently generates a set of KGQs from its own sensor data. High-confidence instances from the cooperative agent are then transmitted to the ego-vehicle. Upon reception, the ego-vehicle employs the Kinematic-Grounded Association (KGA) and Coarse-to-Fine Aggregation (CFA) modules to process these incoming KGQs, fusing them with its own to produce the final set of tracked objects. During training, additional KGQs are initialized from corrupted ground truth (GT) boxes for Cooperative Instance Denoising (CID).

3.2 Sparse Instance Extraction

As depicted in Figure 3, the first stage of our pipeline, deployed independently on each agent, is Sparse Instance Extraction. For this, we adapt the Sparse4D framework (Lin et al. 2023a,b,c). The process begins by initializing a set of KGQs with predefined anchor boxes derived from dataset priors. These KGQs are then refined by directly aggregating information from multi-scale image features using deformable attention. This fully-sparse paradigm is a core design choice, as it bypasses computationally expensive dense BEV representations and their prohibitive scaling costs, making it well-suited for long-range cooperative perception. To ensure temporal consistency for tracking, the framework employs a recurrent mechanism where high-confidence KGQs are assigned a tracking ID and propagated to the subsequent frame. This stage ultimately provides a continuous stream of refined, temporally-aware KGQs for the cooperative fusion modules.

3.3 Kinematic-Grounded Association

Once the ego-vehicle receives KGQs from a cooperative agent, it performs association. This critical step addresses the first challenge outlined in our introduction: robustly

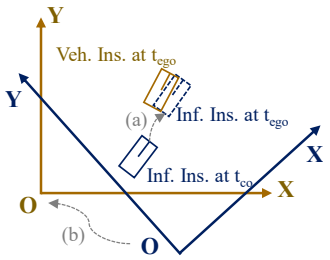


Figure 4: Spatio-Temporal Alignment for KGQ state vectors

matching instances across different viewpoints and asynchronous timestamps. Existing methods (Yu et al. 2025; Zhong et al. 2024) often rely on the Euclidean distance between simple reference points, a strategy that is fragile in complex scenarios. Our association mechanism overcomes this by leveraging the rich geometric and kinematic information encoded in each KGQ's state vector $\mathcal{S}$, enabling precise alignment and matching.

Spatio-Temporal Alignment. The first step of association is aligning the transmitted cooperative instances, $\{\mathcal{F}_{\text{co}}(t_{\text{co}}), \mathcal{S}_{\text{co}}(t_{\text{co}})\}$, to the ego-vehicle's current spatio-temporal frame. We address latency compensation and coordinate projection separately.

To handle latency from asynchronous communication, we utilize the velocity (v_x, v_y, v_z) encoded within the state vector $\mathcal{S}_{\text{co}}(t_{\text{co}})$. Applying a constant velocity motion model, we predict the object's state at the ego-vehicle's timestamp, yielding an updated state vector $\mathcal{S}_{\text{co}}(t_{\text{ego}})$, as shown in Figure 4(a). For the feature vector $\mathcal{F}_{\text{co}}$, we find that the recurrent nature of the extraction module provides sufficient time invariance, so we use it directly: $\mathcal{F}_{\text{co}}(t_{\text{ego}}) = \mathcal{F}_{\text{co}}(t_{\text{co}})$.

Next, we project each instance into the ego-vehicle's coordinate system. The transformation matrix is calculated as $\mathbf{T}_{\text{co}\to\text{ego}}(t_{\text{ego}}) = \mathbf{T}_{\text{ego}\to\text{glb}}(t_{\text{ego}})^{-1} \cdot \mathbf{T}_{\text{co}\to\text{glb}}(t_{\text{co}})$. We apply this to the state vector to get the fully aligned $\widetilde{\mathcal{S}_{\text{co}}}(t_{\text{ego}})$, as shown in Figure 4(b). For the feature vector, we follow prior work (Fan et al. 2024; Yuan et al. 2025) and use a rotation-aware multi-layer perceptron (MLP) to update it:

$$\widetilde{\mathcal{F}_{\text{co}}}(t_{\text{ego}}) = \text{MLP}([\mathcal{F}_{\text{co}}(t_{\text{ego}}); \mathbf{r}_{\text{co}\to\text{ego}}(t_{\text{ego}})])$$

where $\mathbf{r}_{\text{co}\to\text{ego}}(t_{\text{ego}}) \in \mathbb{R}^{1\times 9}$ is the flattened rotation matrix from $\mathbf{T}_{\text{co}\to\text{ego}}(t_{\text{ego}})$.

Cross Agent Matching After alignment, we associate the ego-vehicle's instances $\{\mathcal{F}_{\text{ego}}, \mathcal{S}_{\text{ego}}\}$ with the aligned cooperative instances $\{\widetilde{\mathcal{F}_{\text{co}}}, \widetilde{\mathcal{S}_{\text{co}}}\}$. For simplicity, we omit the timestamp t_{ego} in the following sections.

First, cooperative instances outside the ego-vehicle's ROI R_{ego} are filtered out, as they are not involved in the task. Within R_{ego}, we further define a smaller interaction range, R_{int}, to determine which instances undergo fusion. This distinction is based on the principle that for distant or occluded objects, the ego-vehicle's own sensory data is often unreliable. Forcing fusion in such cases can corrupt high-quality data from cooperative agents with noisy local estimates. Therefore, we only perform matching and subsequent fusion for instances within this trusted near-field region, R_{int}.

For instances inside R_{int}, we introduce a *Geo-Appearance Matching* (GAM) strategy to find optimal instance pairs. This approach directly addresses the fragility of prior methods (Zhong et al. 2024; Yu et al. 2025) that rely on a simple Euclidean distance between single reference points, which can be ambiguous in dense traffic scenarios. To improve robustness, our GAM strategy constructs a pairwise cost matrix C using two distinct components. For an ego-instance i and a cooperative-instance j, a *Geometric Similarity* is computed as a weighted L1 distance between their state vectors ($\mathcal{S}_{\text{ego},i}$ and $\widetilde{\mathcal{S}_{\text{co},j}}$), and an *Appearance Similarity* is computed as the cosine distance between their feature vectors ($\mathcal{F}_{\text{ego},i}$ and $\widetilde{\mathcal{F}_{\text{co},j}}$). The final cost combines these two scores, creating a more discriminative matching criterion that ensures reliable associations, especially for closely-located objects.

The association process results in three distinct groups of instances: 1) the successfully matched pairs, 2) the unmatched ego-vehicle instances, and 3) the unmatched cooperative instances. This entire collection is then passed to the next module for comprehensive refinement.

3.4 Coarse-to-Fine Aggregation

This module consists of a coarse fusion step followed by a more intensive refinement process, as detailed below.

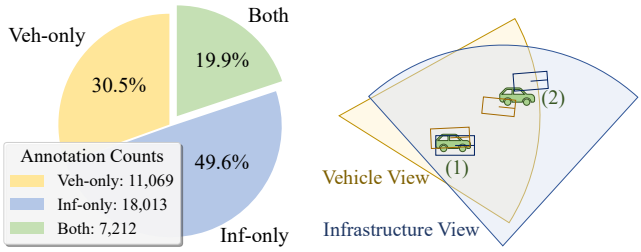
Coarse Fusion. For the successfully matched instance pairs, we first perform a coarse fusion. Following prior works (Yu et al. 2025; Zhong et al. 2025), we employ a lightweight linear network to fuse their respective feature vectors, creating a single, consolidated feature representation for each matched object.

$$\mathcal{F}_{\text{fused}} = \text{MLP}([\mathcal{F}_{\text{ego}}; \widetilde{\mathcal{F}_{\text{co}}}])$$

Multi-Context Refinement. The resulting fused KGQs, along with all unmatched ones from both the ego-vehicle and the cooperative agent, then proceed through an iterative refinement process. This process is inspired by the Sparse4D decoder architecture (Lin et al. 2023a,b,c) but is specifically adapted for the cooperative perception task. Unlike prior cooperative methods that often refine instances using only temporal (Zhong et al. 2024) or cooperative contexts (Wang et al. 2025a,c), we argue that leveraging the ego-vehicle's rich image features is also critical for cooperation.

Each refinement stage is composed of several key operations to leverage multiple contexts. First, *Multi-Head Attention* mechanisms enable rich instance-level interactions. Temporal cross-attention links current instances with those from the previous frame, enabling the model to understand object motion and maintain tracking consistency. Cooperative cross-attention allows interaction with the full set of aligned cooperative KGQs, which is essential for incorporating information from areas occluded to the ego-vehicle. Self-attention captures relationships among all instances within the current frame, helping the model reason about the scene's layout and avoid duplicate detections. For these operations, the state vectors $\mathcal{S}$ are transformed into high-dimensional embeddings to serve as positional encodings.

Following the attention layers, a *Deformable Aggregation* module further refines each instance by sampling from the



(a) Distribution of annotation visibility in V2X-Seq dataset

(b) KGQs from different agents for the same GTs (as car icons)

Figure 5: Motivation for CID. (a) A significant portion of ground-truth objects are visible to only one agent, limiting opportunities for cooperative supervision. (b) Even when an object is visible to both agents, predictions for the same GT (2) can be too far apart to be matched, further reducing positive samples for the fusion module.

ego-vehicle's multi-scale image features. This step is crucial for grounding the abstract instance representations in the raw visual data, leading to more precise localization. Finally, a *Prediction Head*, consisting of a feedforward network and an output layer, predicts the classification scores and state vector refinements.

3.5 Cooperative Instance Denoising

A key challenge in sparse cooperative perception is the scarcity of positive supervision signals, particularly during the early stages of training when few instances are successfully matched across agents, as illustrated in Figure 5. To address this, we draw inspiration from denoising techniques in object detection (Lin et al. 2023c; Li et al. 2023, 2024) and introduce a Cooperative Instance Denoising task. The denoising instances are initialized by adding small perturbations to GT objects and then fed into the network alongside the normal ones. Since the denoising queries for each agent originate from the same GT set, their correspondence is known a priori. This design provides a stable and abundant supervisory signal that guides the fusion module, promoting robust model convergence from the outset of training.

Noise Injection. To simulate realistic uncertainties, we inject two types of noise into the GT state vectors during training. The first, *Observation Noise*, simulates intra-agent uncertainties like sensor measurement errors by adding small, random perturbations to the attributes of GT state vectors in their respective local coordinate systems. Specifically, we add noise sampled from a uniform distribution over (-2.0m, 2.0m) for positional attributes and (-0.5, 0.5) for all other dimensions. The second, a novel *Transformation Noise*, simulates inter-agent uncertainties such as extrinsic calibration errors or timestamp asynchrony. This is achieved by applying minor random rotations and translations to the coordinate transformation matrix $\mathbf{T}_{\text{co}\to\text{ego}}$. The translations are sampled from a normal distribution with a mean of 0 and a standard deviation of 1.0m, while rotations are sampled with a standard deviation of 2.0 degrees.

Denoising Pipeline and Supervision. The denoising instances are processed through a pipeline that mirrors the main network but incorporates critical modifications to ensure effective and clean supervision. First, during cross-agent matching, denoising queries are associated using their tracking IDs inherited from the original GT objects. This provides the network with a stable and plentiful stream of perfectly matched pairs, which is essential for learning the feature alignment and fusion process. Second, to prevent information leakage, we enforce a strict separation between the normal and denoising pipelines within the attention mechanisms of the refinement module. A custom attention mask is implemented to ensure normal KGQs and denoising queries operate in distinct groups. This separation is vital as it prevents the model from accessing GT information during the standard perception task, forcing it to learn meaningful fusion strategies instead of trivial shortcuts. The entire process provides strong, stable supervisory signals throughout all training stages, leading to more robust convergence.

4 Experiments

4.1 Datasets and Metrics

We evaluate our approach on two well-established datasets.

V2X-Seq. This is a large-scale, sequential dataset designed for V2I cooperative 3D object detection and tracking (Yu et al. 2023b). It features sensor data from both an ego-vehicle and a connected RSU, capturing complex urban traffic scenarios with significant occlusions. We follow the official protocol in its CVPR 2025 challenge (Hao et al. 2025) and report performance on the 2Hz validation split.

Griffin. This is a pioneering simulated dataset for aerial-ground cooperative perception (Wang et al. 2026). It presents unique challenges due to the large viewpoint disparity and dynamic transformations in V2D scenarios, and provides a code interface to inject real-world imperfections like latency during training and inference.

Evaluation Metrics. Both benchmarks adopt established metrics from the NuScenes benchmark (Caesar et al. 2020) for 3D object detection and tracking, including Average Precision (AP) to assess detection quality and Average Multi-Object Tracking Accuracy (AMOTA) for tracking performance. We also assess the transmission cost in Bytes per second (BPS) and computational efficiency by inference Frames per second (FPS).

4.2 Implementation Details

All experiments were conducted on NVIDIA 3090 GPUs with ResNet50 (He et al. 2016) backbone, following a two-stage training strategy like CoopTrack (Zhong et al. 2025). The single-agent models were first trained for 72 epochs, then the cooperative model was initialized with weights of single-vehicle model and fine-tuned for 48 epochs.

4.3 Comparison with Existing Works

The detection and tracking performance of SparseCoop across the two datasets are presented in Table 1, which details the AP and AMOTA scores alongside the associated communication costs. Our method achieves a new SOTA detection and tracking performance on both datasets, with

Method	V2X-Seq			Griffin-25m			
	AP ↑	**AMOTA ↑**	**TC (BPS) ↓**	**AP ↑**	**AMOTA ↑**	**TC (BPS) ↓**	**CE (FPS) ↑**
No Fusion Baseline	0.166	0.130	0	0.375	0.365	0	8.10
Late Fusion Baseline	0.196	0.263	6.60×10^2	0.378	0.377	1.56×10^3	6.83
Early Fusion Baseline	0.243	0.209	8.19×10^7	0.607	0.670	3.11×10^8	5.17
V2X-ViT (ECCV 2022)	0.268	0.287	2.56×10^6	0.465	<u>0.508</u>	8.00×10^5	7.56
Where2Comm (NIPS 2022)	0.162	0.106	5.40×10^5	0.396	0.406	3.30×10^5	<u>7.60</u>
UniV2X (AAAI 2025)	0.295	0.239	6.96×10^4	0.419	0.456	$\mathbf{5.58 \times 10^4}$	7.06
CoopTrack (ICCV 2025)	<u>0.390</u>	<u>0.328</u>	<u>5.64×10^4</u>	<u>0.479</u>	0.488	1.17×10^5	6.23
SparseCoop (Ours)	**0.530**	**0.421**	$\mathbf{3.17 \times 10^4}$	**0.559**	**0.509**	<u>9.73×10^4</u>	**11.64**

Table 1: Model performance, transmission cost (TC), and computational efficiency (CE) comparison on the V2X-Seq and Griffin-25m datasets. Baseline performance metrics are sourced from (Zhong et al. 2025; Wang et al. 2026). Bold and underlined values denote the best and second-best performance among intermediate fusion methods.

Kinematic-Grounded Association		Coarse-to-Fine Aggregation		Cooperative Denoising		Metrics	
LC	GAM	CFF	MCR	ON	TN	AP↑	AMOTA↑
✓	✓	✓	✓	✓	✓	<u>0.530</u>	**0.421**
✗	✓	✓	✓	✓	✓	0.505	0.414
✓	✗	✓	✓	✓	✓	0.502	0.414
✗	✗	✓	✓	✓	✓	0.505	0.408
✓	✓	✗	✓	✓	✓	0.489	0.375
✓	✓	✓	✗	✓	✓	0.512	0.379
✓	✓	✓	✓	✗	✓	0.521	<u>0.416</u>
✓	✓	✓	✓	✓	✗	**0.531**	0.394
✓	✓	✓	✓	✗	✗	0.521	0.352

Table 2: Ablation study of our proposed modules on the V2X-Seq dataset. The top row shows the full model. Subsequent rows show performance when components are removed (✗). Experiments are grouped by the major submodule being ablated. Abbreviations are: LC (Latency Compensation), GAM (Geo-Appearance Matching), CFF (Coarse Feature Fusion), MCR (Multi-Context Refinement), ON (Observation Noise), and TN (Transformation Noise).

a competitive transmission cost. On the V2X-Seq dataset, SparseCoop achieves an AP of 0.530 and an AMOTA of 0.421, outperforming all other methods. Notably, this performance is achieved with the lowest transmission cost among all learning-based methods, at only 3.17×10^4 BPS. This represents a significant improvement over the next best-performing methods, CoopTrack and UniV2X, while requiring substantially less bandwidth. On the Griffin-25m dataset, SparseCoop continues to demonstrate its superiority, achieving the highest AP (0.559), AMOTA (0.509) and computational efficiency (11.64 FPS). While its transmission cost of 9.73×10^4 BPS is second only to UniV2X, the performance gains are substantial, with 8% in AP and over 50% in FPS against the next-best competitor.

4.4 Ablation Study

To validate the effectiveness of our proposed components, we conduct a comprehensive ablation study on the V2X-Seq dataset, systematically dissecting the contributions of the Kinematic-Grounded Association, Coarse-to-Fine Aggregation, and Cooperative Instance Denoising modules. The detailed results are presented in Table 2.

Effect of Kinematic-Grounded Association. The exclusion of Latency Compensation (LC), which rectifies temporal asynchrony using kinematic priors, results in a performance drop to 0.505 AP and 0.414 AMOTA. A similar degradation is observed when we remove KGQ's effect on Geo-Appearance Matching (GAM) and revert to a simpler point-based matching metric, with scores falling to 0.502 AP and 0.414 AMOTA. This underscores that leveraging the full geometric and kinematic information encoded in KGQ state vectors is critical for accurate cross-agent instance association. When both components are disabled, the model's performance deteriorates further, confirming their synergistic contribution to achieving robust alignment in complex cooperative scenarios.

Effect of Coarse-to-Fine Aggregation. The Coarse-to-Fine Aggregation module is designed to effectively fuse information from matched instances and refine their representations. Removing the initial Coarse Feature Fusion (CFF) stage, where matched pairs are first integrated, causes a substantial decline in performance to 0.489 AP and 0.375 AMOTA.

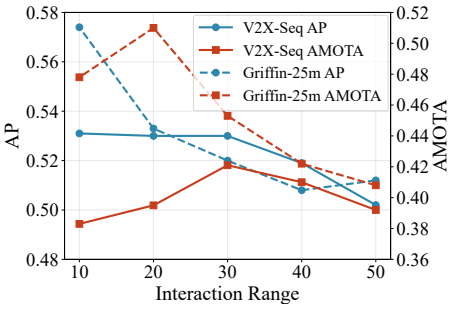


Figure 6: Effect of interaction range on two datasets.

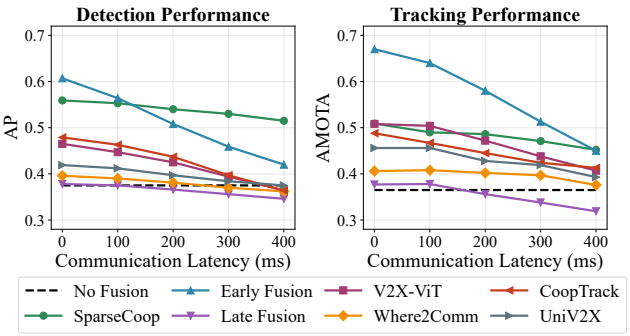


Figure 7: Impact of latency on Griffin-25m dataset.

This highlights the necessity of this direct fusion step for consolidating redundant observations. Furthermore, ablating the subsequent Multi-Context Refinement (MCR) process leads to a significant performance drop to 0.512 AP and 0.379 AMOTA. This comprehensive stage facilitates interaction with temporal and cooperative instances while also using deformable aggregation to ground queries in the ego-vehicle's image features. The performance decline underscores that both multi-agent contextualization and refinement against raw visual data are crucial for comprehensive scene understanding.

Effect of Cooperative Denoising. This training-only task is introduced to provide stable and abundant supervisory signals, mitigating the challenges of sparse positive samples in cooperative settings. Removing Observation Noise (ON), which simulates intra-agent sensor uncertainty, slightly degrades performance. However, the removal of Transformation Noise (TN), which models inter-agent calibration and asynchrony errors, leads to a more significant drop in tracking performance (AMOTA from 0.421 to 0.394), confirming its importance for learning robust alignment. When the entire denoising pipeline is deactivated, the model suffers a severe degradation, particularly in tracking, with AMOTA plummeting to 0.352. This result decisively demonstrates that the cooperative denoising task is crucial for stabilizing the training process and enabling the model to learn effective fusion strategies.

Effect of Interaction Range. A crucial hyperparameter in our framework is the interaction range (R_{int}), which determines the threshold for fusing cooperative instances with the ego-vehicle's perception. As shown in Figure 6, our experiments reveal a clear and significant trend across both datasets. As R_{int} is decreased, we observe a general increase in detection accuracy (AP). Concurrently, tracking performance (AMOTA) first improves and then declines after reaching an optimal point. Based on our tests, the peak performance on the V2X-Seq dataset was achieved at $R_{int} = 30m$, while the optimal range for the Griffin-25m dataset was found to be $15m$.

The observed increase in AP with a smaller R_{int} suggests that for distant or occluded targets where the ego-vehicle's perception is inherently unreliable, forcing an interaction can be counterproductive. The model appears to get confused by the low-quality local data, which can suppress the high-quality cooperative detection. By directly outputting these instances, their integrity is preserved. The behavior of AMOTA, however, indicates a trade-off. While focusing fusion on nearby objects is beneficial, an overly restrictive R_{int} can cause the system to output duplicate detections for the same object—one from the ego-vehicle and another from the cooperative agent—thereby degrading tracking performance. Thus, the optimal R_{int} represents a balance between maximizing information gain from fusion in the reliable near-field while preventing data corruption and redundancy in the far-field.

Robustness to Communication Latency. We evaluated the framework's resilience to communication latency, a critical factor for real-world deployment, using the Griffin-25m benchmark. Results are shown in Figure 7. While the performance of all tested cooperative methods degraded with increasing latency, SparseCoop demonstrated a markedly superior level of robustness. At zero latency, early fusion methods held a performance advantage, as expected. However, as latency was increased to 200ms and beyond, SparseCoop's AP scores surpassed those of all other methods, including early fusion. This indicates a significantly more graceful performance degradation compared to competing approaches.

This exceptional robustness can be directly attributed to our framework's architectural design. The Kinematic-Grounded Association module incorporates an explicit Latency Compensation mechanism. By leveraging the velocity information encoded within each instance's state vector, the model applies kinematic priors to accurately predict an object's state at the ego-vehicle's current timestamp, effectively neutralizing the temporal asynchrony caused by communication delays.

5 Conclusion

In this work, we have presented SparseCoop, a fully sparse framework that solves key challenges in cooperative perception by avoiding dense BEV maps. Our method uses kinematic-grounded queries for robust spatio-temporal alignment, a coarse-to-fine aggregation module to effectively fuse instance data, and a cooperative denoising task to stabilize training. Experiments show SparseCoop achieves SOTA performance on the V2X-Seq and Griffin datasets.

Acknowledgments

This work was supported by the National Natural Science Foundation of China for the Science Fund for Creative Research Groups (No. 52221005) and the Key Project (No. 52131201).

References

Arnold, E.; Dianati, M.; De Temple, R.; and Fallah, S. 2022. Cooperative Perception for 3D Object Detection in Driving Scenarios Using Infrastructure Sensors. *IEEE Transactions on Intelligent Transportation Systems*, 23(3): 1852–1864.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11618–11628.

Caillot, A.; Ouerghi, S.; Vasseur, P.; Boutteau, R.; and Dupuis, Y. 2022. Survey on Cooperative Perception in an Automotive Context. *IEEE Transactions on Intelligent Transportation Systems*, 23(9): 14204–14223.

Chen, Q.; Tang, S.; Yang, Q.; and Fu, S. 2019. Cooper: Cooperative Perception for Connected Autonomous Vehicles Based on 3D Point Clouds. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, 514–524.

Chen, Z.; Shi, Y.; and Jia, J. 2023. TransIFF: An Instance-Level Feature Fusion Framework for Vehicle-Infrastructure Cooperative 3D Detection with Transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 18159–18168. Paris, France: IEEE. ISBN 979-8-3503-0718-4.

Fan, S.; Yu, H.; Yang, W.; Yuan, J.; and Nie, Z. 2024. QUEST: Query Stream for Practical Cooperative Perception. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 18436–18442.

Gao, B.; Liu, J.; Zou, H.; Chen, J.; He, L.; and Li, K. 2024. Vehicle-Road-Cloud Collaborative Perception Framework and Key Technologies: A Review. *IEEE Transactions on Intelligent Transportation Systems*, 25(12): 19295–19318.

Han, Y.; Zhang, H.; Li, H.; Jin, Y.; Lang, C.; and Li, Y. 2023. Collaborative Perception in Autonomous Driving: Methods, Datasets, and Challenges. *IEEE Intelligent Transportation Systems Magazine*, 15(6): 131–151.

Hao, R.; Yu, H.; Zhong, J.; Wang, C.; Wang, J.; Kan, Y.; Yang, W.; Fan, S.; Yin, H.; Qiu, J.; Mu, Y.; Sun, J.; Chen, L.; Zimmer, W.; Zhang, D.; Zhang, S.; Schwager, M.; Luo, P.; and Nie, Z. 2025. Research Challenges and Progress in the End-to-End V2X Cooperative Autonomous Driving Competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1828–1839.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hu, Y.; Fang, S.; Lei, Z.; Zhong, Y.; and Chen, S. 2022. Where2comm: Communication-Efficient Collaborative Perception via Spatial Confidence Maps. In *Advances in Neural Information Processing Systems*, volume 35, 4874–4886.

Hu, Y.; Lu, Y.; Xu, R.; Xie, W.; Chen, S.; and Wang, Y. 2023. Collaboration Helps Camera Overtake LiDAR in 3D Detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9243–9252.

Hu, Y.; Peng, J.; Liu, S.; Ge, J.; Liu, S.; and Chen, S. 2024. Communication-Efficient Collaborative Perception via Information Filling with Codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15481–15490.

Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L. M.; and Zhang, L. 2024. DN-DETR: Accelerate DETR Training by Introducing Query DeNoising. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4): 2239–2251.

Li, F.; Zhang, H.; Xu, H.; Liu, S.; Zhang, L.; Ni, L. M.; and Shum, H.-Y. 2023. Mask DINO: Towards A Unified Transformer-based Framework for Object Detection and Segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3041–3050.

Li, Y.; Ren, S.; Wu, P.; Chen, S.; Feng, C.; and Zhang, W. 2021. Learning Distilled Collaboration Graph for Multi-Agent Perception. In *Advances in Neural Information Processing Systems*, volume 34, 29541–29552. Curran Associates, Inc.

Lin, X.; Lin, T.; Pei, Z.; Huang, L.; and Su, Z. 2023a. Sparse4D: Multi-View 3D Object Detection with Sparse Spatial-Temporal Fusion. arXiv:2211.10581.

Lin, X.; Lin, T.; Pei, Z.; Huang, L.; and Su, Z. 2023b. Sparse4D v2: Recurrent Temporal Fusion with Sparse Model. arXiv:2305.14018.

Lin, X.; Pei, Z.; Lin, T.; Huang, L.; and Su, Z. 2023c. Sparse4D v3: Advancing End-to-End 3D Detection and Tracking. arXiv:2311.11722.

Liu, H.; Chu, H.; Zhuo, J.; Zou, B.; Chen, J.; and Ma, H. 2025. SparseComm: An Efficient Sparse Communication Framework for Vehicle-Infrastructure Cooperative 3D Detection. *Pattern Recognition*, 158: 110961.

Liu, Y.; Wang, T.; Zhang, X.; and Sun, J. 2022. PETR: Position Embedding Transformation for Multi-View 3D Object Detection. In *Lecture Notes in Computer Science*, Lecture Notes in Computer Science, 531–548. Cham: Springer Nature Switzerland. ISBN 978-3-031-19812-0.

Liu, Y.; Yan, J.; Jia, F.; Li, S.; Gao, A.; Wang, T.; and Zhang, X. 2023. PETRv2: A Unified Framework for 3D Perception from Multi-Camera Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3262–3272.

Lu, Y.; Li, Q.; Liu, B.; Dianati, M.; Feng, C.; Chen, S.; and Wang, Y. 2023. Robust Collaborative 3D Object Detection in Presence of Pose Errors. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 4812–4818.

Valiente, R.; Zaman, M.; Ozer, S.; and Fallah, Y. P. 2019. Controlling Steering Angle for Cooperative Self-driving Vehicles Utilizing CNN and LSTM-based Deep Networks. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, 2423–2428.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, J.; Cao, X.; Zhong, J.; Zhang, Y.; Han, Z.; Yu, H.; Zhang, C.; He, L.; Xu, S.; and Wang, J. 2026. Griffin: Aerial-Ground Cooperative Detection and Tracking Dataset and Benchmark. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wang, R.; Gao, X.; Xiang, H.; Xu, R.; and Tu, Z. 2025a. CoCMT: Communication-Efficient Cross-Modal Transformer for Collaborative Perception. arXiv:2503.13504.

Wang, S.; Bin, L.; Xiao, X.; Xiang, Z.; Shan, H.; and Liu, E. 2025b. IFTR: An Instance-Level Fusion Transformer for Visual Collaborative Perception. In *Computer Vision – ECCV 2024*, 124–141. Cham: Springer Nature Switzerland. ISBN 978-3-031-73021-4.

Wang, T.-H.; Manivasagam, S.; Liang, M.; Yang, B.; Zeng, W.; and Urtasun, R. 2020. V2VNet: Vehicle-to-Vehicle Communication for Joint Perception and Prediction. In *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, 605–621. Cham: Springer International Publishing. ISBN 978-3-030-58536-5.

Wang, Y.; Guizilini, V. C.; Zhang, T.; Wang, Y.; Zhao, H.; and Solomon, J. 2022. DETR3D: 3D Object Detection from Multi-View Images via 3D-to-2D Queries. In *Proceedings of the 5th Conference on Robot Learning*, 180–191. PMLR.

Wang, Z.; Fan, S.; Huo, X.; Xu, T.; Wang, Y.; Liu, J.; Chen, Y.; and Zhang, Y.-Q. 2024. EMIFF: Enhanced Multi-scale Image Feature Fusion for Vehicle-Infrastructure Cooperative 3D Object Detection. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 16388–16394. Yokohama, Japan: IEEE. ISBN 979-8-3503-8457-4.

Wang, Z.; Xu, S.; Zhuang, X.; Xu, T.; Wang, Y.; Liu, J.; Chen, Y.; and Zhang, Y.-Q. 2025c. CoopDETR: A Unified Cooperative Perception Framework for 3D Detection via Object Query. In *IEEE International Conference on Robotics and Automation, ICRA 2025, Atlanta, GA, USA, May 19-23, 2025*, 2732–2739. IEEE.

Xu, R.; Tu, Z.; Xiang, H.; Shao, W.; Zhou, B.; and Ma, J. 2023. CoBEVT: Cooperative Bird's Eye View Semantic Segmentation with Sparse Transformers. In *Proceedings of The 6th Conference on Robot Learning*, 989–1000. PMLR.

Xu, R.; Xiang, H.; Tu, Z.; Xia, X.; Yang, M.-H.; and Ma, J. 2022a. V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer. In *Lecture Notes in Computer Science*, Lecture Notes in Computer Science, 107–124. Cham: Springer Nature Switzerland. ISBN 978-3-031-19842-7.

Xu, R.; Xiang, H.; Xia, X.; Han, X.; Li, J.; and Ma, J. 2022b. OPV2V: An Open Benchmark Dataset and Fusion Pipeline for Perception with Vehicle-to-Vehicle Communication. In *2022 International Conference on Robotics and Automation (ICRA)*, 2583–2589.

Yin, H.; Tian, D.; Lin, C.; Duan, X.; Zhou, J.; Zhao, D.; and Cao, D. 2024. V2VFormer++: Multi-Modal Vehicle-to-Vehicle Cooperative Perception via Global-Local Transformer. *IEEE Transactions on Intelligent Transportation Systems*, 25(2): 2153–2166.

Yu, H.; Tang, Y.; Xie, E.; Mao, J.; Luo, P.; and Nie, Z. 2023a. Flow-Based Feature Fusion for Vehicle-Infrastructure Cooperative 3D Object Detection. *Advances in Neural Information Processing Systems*, 36: 34493–34503.

Yu, H.; Yang, W.; Ruan, H.; Yang, Z.; Tang, Y.; Gao, X.; Hao, X.; Shi, Y.; Pan, Y.; Sun, N.; Song, J.; Yuan, J.; Luo, P.; and Nie, Z. 2023b. V2X-seq: A Large-Scale Sequential Dataset for Vehicle-Infrastructure Cooperative Perception and Forecasting. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5486–5495.

Yu, H.; Yang, W.; Zhong, J.; Yang, Z.; Fan, S.; Luo, P.; and Nie, Z. 2025. End-to-End Autonomous Driving Through V2X Cooperation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(9): 9598–9606.

Yuan, Y.; Cheng, H.; Yang, M. Y.; and Sester, M. 2023. Generating Evidential BEV Maps in Continuous Driving Space. *ISPRS Journal of Photogrammetry and Remote Sensing*, 204: 27–41.

Yuan, Y.; Xia, Y.; Cremers, D.; and Sester, M. 2025. SparseAlign: A Fully Sparse Framework for Cooperative Object Detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 22296–22305.

Zhong, J.; Wang, J.; Xu, J.; Li, X.; Nie, Z.; and Yu, H. 2025. CoopTrack: Exploring End-to-End Learning for Efficient Cooperative Sequential Perception. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, 26954–26965.

Zhong, J.; Yu, H.; Zhu, T.; Xu, J.; Yang, W.; Nie, Z.; and Sun, C. 2024. Leveraging Temporal Contexts to Enhance Vehicle-Infrastructure Cooperative Perception. In *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, 915–922.