V2X-Seq: A Large-Scale Sequential Dataset for Vehicle-Infrastructure Cooperative Perception and Forecasting

Haibao Yu[1,2], Wenxian Yang[1], Hongzhi Ruan[1,5], Zhenwei Yang[1,6], Yingjuan Tang[1,7], Xu Gao[3], Xin Hao[3], Yifeng Shi[3], Yifeng Pan[3], Ning Sun[4], Juan Song[4], Jirui Yuan[1], Ping Luo[2], Zaiqing Nie[1] *
[1]Institute for AI Industry Research (AIR), Tsinghua University [2]The University of Hong Kong
[3]Baidu Inc. [4]Beijing Connected and Autonomous Vehicles Technology Co., Ltd
[5]University of Chinese Academy of Science
[6]University of Science and Technology Beijing [7]Beijing Institute of Technology

Abstract

Utilizing infrastructure and vehicle-side information to track and forecast the behaviors of surrounding traffic participants can significantly improve decision-making and safety in autonomous driving. However, the lack of real-world sequential datasets limits research in this area. To address this issue, we introduce V2X-Seq, the first large-scale sequential V2X dataset, which includes data frames, trajectories, vector maps, and traffic lights captured from natural scenery. V2X-Seq comprises two parts: the sequential perception dataset, which includes more than 15,000 frames captured from 95 scenarios, and the trajectory forecasting dataset, which contains about 80,000 infrastructure-view scenarios, 80,000 vehicle-view scenarios, and 50,000 cooperative-view scenarios captured from 28 intersections' areas, covering 672 hours of data. Based on V2X-Seq, we introduce three new tasks for vehicle-infrastructure cooperative (VIC) autonomous driving: VIC3D Tracking, Online-VIC Forecasting, and Offline-VIC Forecasting. We also provide benchmarks for the introduced tasks. Find data, code, and more up-to-date information at https://github.com/AIR-THU/DAIR-V2X-Seq.

1. Introduction

Although single-vehicle autonomous driving has made significant advancements in recent years, it still faces significant safety challenges due to its limited perceptual field and inability to accurately forecast the behaviors of traffic participants. These challenges hinder autonomous vehicles from making well-informed decisions and driving safer. A promising solution to address these challenges is to leverage infrastructure information via Vehicle-to-Everything (V2X) communication, which has been shown to significantly expand perception range and enhance autonomous driving safety [1, 38]. However, current research primarily focuses on utilizing infrastructure data to improve the perception ability of autonomous driving, particularly in the context of frame-by-frame 3D detection. To enable well-informed decision-making for autonomous vehicles, it is critical to also incorporate infrastructure data to track and predict the behavior of surrounding traffic participants.

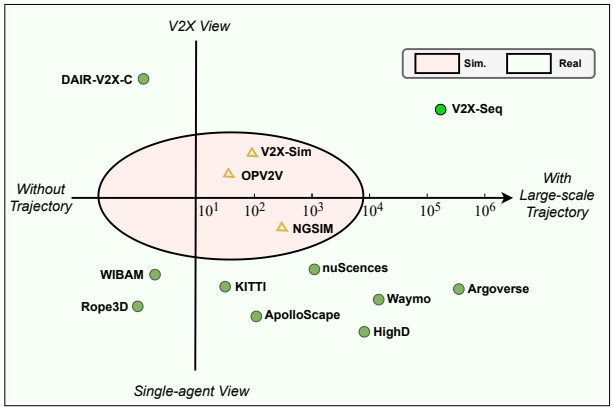

Figure 1. Autonomous driving datasets. V2X-Seq is the first large-scale, real-world, and sequential V2X dataset. The green circle denotes the real-world dataset, and the pink triangle denotes the simulated dataset. The abscissa represents the number of sequences.

To accelerate the research on cooperative sequential perception and forecasting, we release a large-scale sequential V2X dataset, V2X-Seq. All elements of this dataset were captured and generated from real-world scenarios. Compared with DAIR-V2X [38], which focuses on 3D object detection tasks, V2X-Seq is specifically designed for tracking and trajectory forecasting tasks. The V2X-Seq dataset is divided into two parts: the sequential perception dataset

*Corresponding author. Work done while at AIR. For any questions or discussions, please email dair@air.tsinghua.edu.cn.

Table 1. Comparison with the public autonomous driving dataset. '-' denotes that the information is not provided. 'Real/Sim.' indicates whether the data was collected from the real world or a simulator. V2X view includes multi-vehicle cooperative view and vehicle-infrastructure cooperative view. V2X-Seq is the first large-scale sequential V2X dataset and focuses on vehicle-infrastructure cooperative view. All data elements, including the traffic light signals, are captured and generated from the real world.

Dataset	Year	Real/Sim.	View	With Trajectory	With 3D Boxes	With Maps	With Traffic Light	Tracked Objects/Scene	Total Time (hour)	Scenes
KITTI [12]	2012	Real	Single-vehicle	✓	✓	×	×	43.67	1.5	50
nuScenes [3]	2019	Real	Single-vehicle	✓	✓	✓	×	75.75	5.5	1,000
Waymo Motion [10, 28]	2021	Real	Single-vehicle	✓	✓	✓	✓	-	574	103,354
Argoverse [5]	2019	Real	Single-vehicle	✓	×	✓	×	50.03	320	324,557
ApolloScape [16, 25]	2019	Real	Single-vehicle	✓	×	×	×	50.6	2.5	103
HighD [17]	2018	Real	Drone	✓	×	✓	×	-	16.5	5,940
WIBAM [14]	2021	Real	Infrastructure	×	×	×	×	0	0.25	0
NGSIM [29]	2016	Sim.	Infrastructure	✓	×	×	×	-	1.5	540
V2X-Sim 2.0 [21]	2022	Sim.	V2X	✓	✓	×	×	-	0.3	100
OPV2V [35]	2021	Sim.	V2X	✓	✓	×	×	26.5	0.2	73
Cooper(inf) [1]	2019	Sim.	V2X	✓	✓	×	×	30	-	<100
DAIR-V2X-C [38]	2021	Real	V2X	×	✓	✓	×	0	0.5	100
V2X-Seq/Perception	2023	Real	V2X	✓	✓	✓	×	110	0..43	95
V2X-Seq/Forecasting	2023	Real	V2X	✓	✓	✓	✓	101	583	210,000

and the trajectory forecasting dataset. The sequential perception dataset comprises 15,000 frames captured from 95 scenarios, which include infrastructure images, infrastructure point clouds, vehicle-side images, vehicle-side point clouds, 3D detection/tracking annotations, and vector maps. The trajectory forecasting dataset comprises 210,000 scenarios, including 50,000 cooperative-view scenarios, that were mined from 672 hours of data collected from 28 intersection areas. To our knowledge, V2X-Seq is the first sequential V2X dataset that includes such a large-scale scenarios, making it an ideal resource for developing and testing cooperative perception and forecasting algorithms.

Based on the V2X-Seq dataset, we introduce three novel tasks for vehicle-infrastructure cooperative perception and forecasting. The first task is VIC3D Tracking, which aims to cooperatively locate, identify, and track 3D objects using sequential sensor inputs from both the vehicle and infrastructure. The second task is Online-VIC trajectory forecasting, which focuses on accurately predicting future behavior of target agents by utilizing past infrastructure trajectories, ego-vehicle trajectories, real-time traffic lights, and vector maps. The third task is Offline-VIC trajectory forecasting, which involves extracting relevant knowledge from previously collected infrastructure data to facilitate vehicle-side forecasting. These proposed tasks are accompanied by rich benchmarks. Additionally, we propose an intermediate-level framework, FF-Tracking, to effectively solve the VIC3D Tracking task.

The main contributions are organized as follows:

- We release the V2X-Seq dataset, which constitutes the first large-scale sequential V2X dataset. All data are captured and generated from the real world.
- Based on the V2X-Seq dataset, we introduce three tasks for the vehicle-infrastructure cooperative autonomous driving community. To enable a fair evaluation of these tasks, we have carefully designed a set of benchmarks.
- We propose a middle fusion method, named FF-Tracking, for solving VIC3D Tracking and our proposed method can efficiently overcome the latency challenge.

2. Related Work

Autonomous Driving Datasets. Public datasets have greatly facilitated the development of autonomous driving. Kitti [12] is the pioneering dataset for autonomous driving. nuScenes [3], Waymo Open [10, 28], ApolloScape [16], and ONCE [26] are large-scale and real-world datasets that support 3D object detection, tracking and prediction tasks. Argoverse [5], Argoverse 2.0 [34], Lyft [13], and nuPlan [4] release large-scale trajectories generated from the raw sensor data to support motion prediction and planning tasks. These datasets are all captured with single-vehicle sensors. Repo3D [37], WIBAM [14], and A9-Dataset [7] release the infrastructure-only 3D detection dataset. HighD [17] and NGSIM [29] release the drone or infrastructure-only trajectories dataset. OpenV2V [35], V2X-Sim 2.0 [21], and Cooper(inf) [1] release small-scale sequential and simulated datasets for multi-vehicle cooperative perception. DAIR-V2X-C [38] is the first real-world V2X dataset that supports VIC3D object detection; however, it does not provide the trajectory information. Compared with these existing public autonomous driving datasets, our V2X-Seq is the first large-scale sequential V2X dataset. All data are captured and generated from the real world. The dataset also includes vector maps and real-time traffic light signal data. It will be suitable for studying the Vehicle-Infrastructure Cooperative sequential perception and trajectory forecasting tasks.

Cooperative Autonomous Driving. Utilizing data from the road environment to enhance the safety of autonomous driving has attracted significant research attention in recent years. Some research works have focused on multi-vehicle cooperative perception, where lightweight feature-level data is transmitted and shared for improved perception of other vehicles [6, 22, 32]. To address communication delays in multi-vehicle 3D object detection, [20] proposes a time-compensation module for latency. On the other hand, some works have explored the use of infrastructure data to improve autonomous driving. For instance, [38] formalizes the vehicle-infrastructure cooperative 3D object detection task and highlights the latency challenges in cooperative perception. [39] further proposed to use feature flow prediction to overcome the uncertain latency. Other works such as [1, 14, 23, 24] also consider transmitting feature-level data from infrastructure to the vehicle side. To empower only sharing sparse yet perceptually critical information, [15] utilizes a spatial confidence map. Moreover, [21] applies the Transformer [31] to fuse the features. Works such as [8,27,30] integrate the infrastructure data for control in autonomous driving. However, most current works on cooperative autonomous driving focus on perceptual completion, overlooking the importance of temporal perception and forecasting. In this paper, we contribute to this field by releasing the V2X-Seq dataset, which is suitable for exploring sequential perception and forecasting tasks in vehicle-infrastructure cooperative settings.

3. V2X-Seq Dataset

To enable the exploration of the role of infrastructure in sequential perception and trajectory forecasting, we introduce the V2X-Seq dataset. This large-scale, real-world dataset contains sequential vehicle-to-everything (V2X) data. The sequential perception component of the dataset is presented in Section 3.1, while the trajectory forecasting component is detailed in Section 3.2. Additionally, we provide an overview of the vector maps and traffic lights used in the dataset in Section 3.3.

3.1. The Sequential Perception Dataset.

3D tracking is a critical component in autonomous driving, as it provides sequential perception information that facilitates 3D detection and prediction. To enable exploration of the role of infrastructure in 3D tracking, we release the Sequential Perception Dataset (SPD). The SPD builds on the DAIR-V2X-C 3D detection dataset [38] and consists of more than 15,000 frames captured from 95 representative scenes with 10∼20s duration sequences, comprising both vehicle sequential frames (images and point clouds) and infrastructure sequential frames (images and point clouds) sampled at 10 Hz. We provide 3D tracking annotations for each object of interest in each sequence, with unique tracking IDs shared by the same objects in each sequence, even if they are fully occluded in some frames. Additionally, for each scene, we provide an extra local vector map.

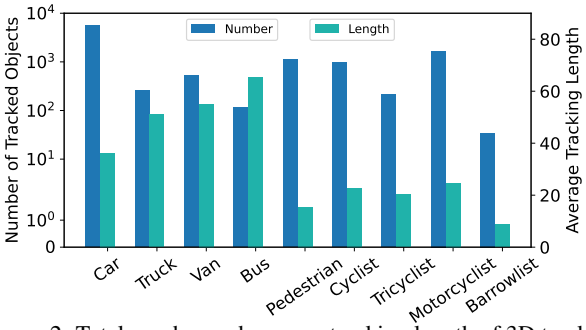

Figure 2. Total number and average tracking length of 3D tracked objects per category for the sequential perception dataset (SPD). The distribution of tracked objects is relatively balanced.

Data Collection and Annotation. The SPD builds on the DAIR-V2X-C [38]. We select 95 representative scenes from this dataset, where an autonomous driving vehicle drives through intersections equipped with sensors. SPD provides high-quality 3D annotations for ten object classes in every image and point cloud frame, including category attributes, occlusion state, truncated state, and a 7-dimensional cuboid modelled as x, y, z, width, length, height, and yaw angle. The object categories include various vehicles, pedestrians, and cyclists. Building upon the DAIR-V2X-C dataset, our annotators assigned a unique tracking ID to each annotated object, except for static traffic cone objects. The same object in one sequence is assigned a unique tracking ID, even when it is completely occluded in some frames. Moreover, we provide cooperative tracking annotations for the cooperative-view sequences based on spatial and temporal matching. Specifically, for each frame in each ego-vehicle sequence, we generate an infrastructure frame with the same timestamp as the corresponding ego-vehicle frame. This frame contains the 3D boxes interpolated and estimated from the infrastructure trajectories. Next, we convert these 3D boxes into an ego-vehicle coordinate system and match and fuse the two-side 3D boxes based on the Euclidean distance measurement and the Hungarian method [18]. To account for possible calibration and interpolation precision errors that may cause spatial matching errors, we compute the similarity of the two-side trajectories corresponding to the two matched 3D boxes. We filter out the matching with low scores and manually refine them to obtain accurate cooperative tracking annotations.

3.2. The Trajectory Forecasting Dataset

We are also interested in studying trajectory forecasting to predict the future locations of tracked objects. Accurately predicting the behavior of surrounding traffic participants can facilitate more rational decision-making and improve the safety of autonomous driving. However, the ego-vehicle

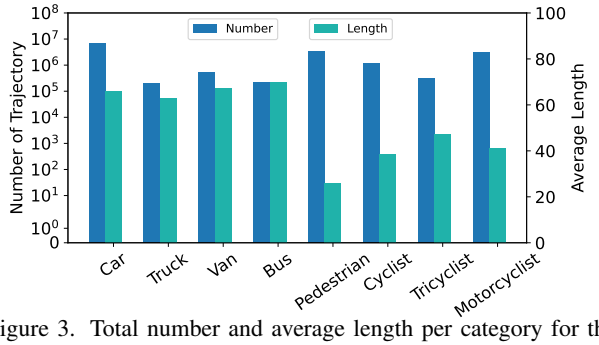

Figure 3. Total number and average length per category for the trajectory forecasting dataset in a relatively uniform distribution of trajectory categories.

prediction capabilities are significantly limited by the lack of sufficient perceptual information and the lack of interaction between different traffic participants. It is valuable to study the Vehicle-Infrastructure Cooperation (VIC) trajectory forecasting to fully utilize the infrastructure data and improve the forecasting ability. Although the Sequential Perception Dataset (SPD) can be used to study the VIC trajectory forecasting, the data scale needs to be larger, and the richness of the trajectories needs to be higher to explore various behaviors. Therefore, we mined interesting trajectories from 336 driving and 336 infrastructure hours at 28 urban intersections in the Beijing Yizhuang Area to form a large-scale trajectory dataset. Details about the data collection and trajectory mining are presented in the Appendix.

The Trajectory Forecasting Dataset (TFD) is composed of about 50,000 cooperative-view, 80,000 infrastructure-view and 80,000 ego-vehicle-view scenarios. Each scenario includes a sequence of tracked object data for 10 seconds at 10HZ, a local vector map, and real-time traffic light signals (only provided for cooperative-view and infrastructure-view scenarios). Among them, 50,000 cooperative-view scenarios were collected at the same time and intersection, where the ego vehicle drove through the equipped intersections. The tracked objects contain 3D boxes modeled with 7 dimensions, an object type attribute from 8 classes, and a trajectory ID. Additionally, we provide cooperative trajectory annotations for cooperative-view scenarios. The cooperative trajectory is generated in a similar way to cooperative tracking annotation but without manual refinement. Each cooperative trajectory is marked with which trajectories it originated from. The released dataset is diverse in terms of different classes and locations. The distribution of classes is presented in Figure 3. We provide the detailed data collection and generation in the Appendix.

3.3. Vector Maps and Traffic Lights.

We provide vector maps for the areas covering the selected 28 intersections, organized similarly to Argoverse [5]. The vector maps contain lane centerlines, crosswalks, and stoplines, represented by line segments with starting and ending points. To meet data security requirements, we add a constant offset to the coordinates of the points located in the world coordinate system. For each lane centerline, we provide attributes such as turning left or right, and we also provide the actual lane width so that we can calculate the boundaries of each lane. As traffic vehicles must follow the lane, including the centerline and boundary, to obey traffic rules, building the spatial context between trajectories and vector maps can provide valuable hints for trajectory tracking and forecasting.

Additionally, we provide real-time traffic light signals for the infrastructure portion of the Trajectory Forecasting Dataset (TFD). During the collection and storage of infrastructure sensor data, we also record traffic light data at 10 Hz. The traffic light signals include the timestamp, location, color status, shape status, and time remaining. This information can significantly influence the behavior of traffic participants. It is worth noting that although nuPlan [4] also provides traffic light data, their data is estimated offline based on traffic flow statistics, whereas our data is obtained directly from the traffic lights themselves.

4. VIC3D Tracking Task

In this section, we detail the formalization of the Vehicle-Infrastructure Cooperative 3D (VIC3D) Tracking task, along with the corresponding evaluation metrics. Furthermore, we propose the FF-Tracking framework, which builds upon FFNet [39], to address the issue of degraded tracking performance caused by latency, thereby improving the overall efficiency of VIC3D Tracking.

Task Description. VIC3D Tracking aims to cooperatively locate, identify, and track 3D objects using both infrastructure and ego-vehicle sequential data while operating under limited communication bandwidth. The input for VIC3D Tracking consists of sequential frames from both ego-vehicle and infrastructure sources:

- Ego-vehicle sequential frames $I_v(t_v^{'})|t_v^{'} \leq t_v$ as well as its relative pose $M_v(t_v^{'})|t_v^{'} \leq t_v$: captured at and before time t_v, where $I_v(\cdot)$ denotes the capturing function of ego-vehicle sensors.
- Infrastructure sequential frames $I_i(t_i^{'})|t_i^{'} \leq t_i$ as well as its relative pose $M_i(t_i^{'})|t_i^{'} \leq t_i$: captured at and before time t_i, where $I_i(\cdot)$ denotes the capturing function of infrastructure sensors. Here, t_i should be earlier than t_v (i.e., $t_i < t_v$) due to the communication delay.

The outputs of VIC3D Tracking include the category, location, orientation, and unique tracking ID of each object in the area of interest surrounding the ego vehicle over time t_v. The corresponding ground truth is the set of 3D tracked objects appearing in one of the cooperative-view sensors over time t_v, which can be formulated as:

$$GT = (GT_v \cup GT_i) \cap R, \tag{1}$$

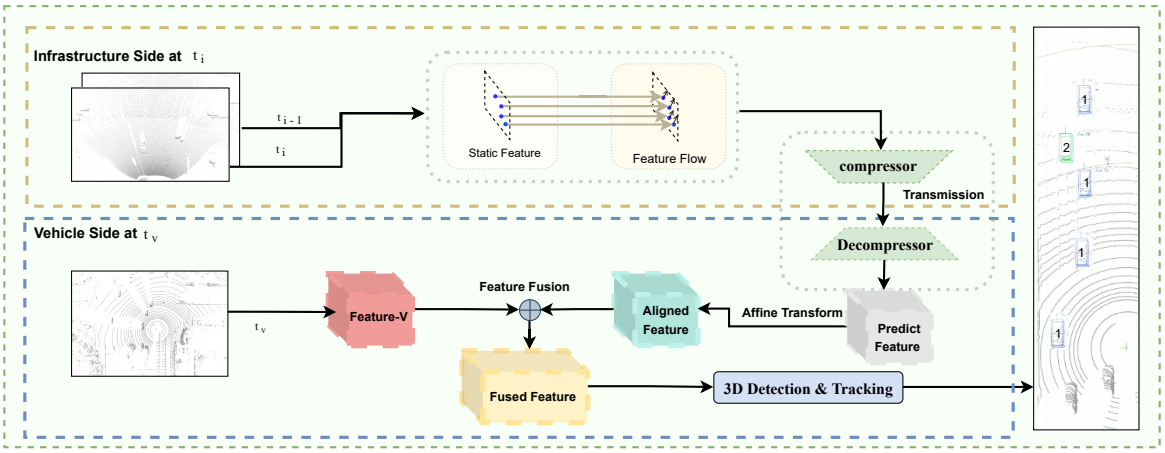

Figure 4. Overview of the FF-Tracking framework. The framework transmits compressed features and feature flows, which can effectively reduce transmission costs while removing fusion errors caused by communication delays.

where GT_v is the ground truth for ego-vehicle sensor perception, GT_i is the ground truth for infrastructure sensor perception, and R is the ego-vehicle interest region.

Evaluation Metrics and Analysis. VIC3D Tracking has two primary objectives: achieving better tracking performance while minimizing transmission costs to reduce bandwidth consumption. To assess these objectives, we use the following metrics:

- MOTA, MOTP and IDS: Multi-Object Tracking Accuracy (MOTA), Multi-Object Tracking Precision (MOTP), and ID Switch (IDS) are three commonly used evaluation metrics for 3D tracking [3, 11]. We use these metrics to measure the performance of VIC3D Tracking approach.
- BPS: Byte Per Second (BPS) measures the amount of data transmitted from the infrastructure to the ego vehicle per second, taking into account the transmission frequency.

However, achieving these objectives presents several challenges. Firstly, we need to reduce the amount of data transmitted to meet the limited communication bandwidth requirement, while ensuring that the transmitted data are valuable enough to improve the tracking performance. The intermediate form is the most likely to achieve a balance between performance and transmission among the three possible data transmission forms (raw, intermediate, and perceived data). Secondly, latency can cause significant damage to cooperative fusion due to scene changes and dynamic object movements over time. Hence, we should consider the use of prediction alignment to remove fusion errors.

FF-Tracking Framework. To address the challenges of VIC3D Tracking, we propose a middle fusion framework called FF-Tracking, which is based on feature flow prediction in FFNet [39]. FF-Tracking transmits both feature and feature flow instead of the single static feature from the infrastructure to the ego vehicle. We predict the future feature to align with the ego-vehicle timestamp using the following linear estimation:

$$F_{future}(t) = F_0 + t * F_1, \tag{2}$$

where F_0 denotes the static feature and F_1 denotes the feature flow. With the predicted feature, we can effectively address the fusion error and solve the latency challenges. To further reduce the transmission cost, we compress the features and feature flows before transmitting them. This approach enables us to achieve the goals of better tracking performance and lower transmission cost while meeting the limited communication bandwidth requirement.

The FF-Tracking framework consists of following parts. 1) Extracting the feature and feature flow from past sequential infrastructure frames. 2) Compressing, transmitting, and decompressing the static feature and feature flow. 3) Predicting the infrastructure feature using Eq. 2. 4) Fusing the features. We transform the predicted feature into a local ego-vehicle coordinate system and then fuse it with the ego-vehicle feature. We extract the ego-vehicle feature from the ego-vehicle point clouds. 5) Generating the tracking results. We use a Single Shot Detector (SSD) [36] to generate the 3D object outputs and then use the AB3DMOT [33] to track the objects and assign a unique tracking ID for each object. The whole process is also illustrated in Fig. 4. Please refer [39] to more feature flow prediction configurations.

5. VIC Trajectory Forecasting Tasks

In this section, we present two trajectory forecasting tasks based on the trajectory forecasting dataset: Online-VIC Forecasting and Offline-VIC Forecasting. These tasks aim to investigate how to effectively leverage real-time infrastructure information and offline behavior knowledge transfer from the infrastructure to the vehicle side.

5.1. Online-VIC Forecasting Task

Task Formulation. Online-VIC Forecasting can be formulated as the problem of predicting future trajectories using real-time infrastructure and vehicle-side data. The inputs for Online-VIC Forecasting are:

- A set of infrastructure trajectories $\{T_i^{(l)}(t_i)\}$ and traffic light signals, where the trajectory $T_i^{(l)}(t_i)$ contains the sequential coordinates of agent $A_i^{(l)}$ at and before time t_i.
- Local vector maps.
- A set of ego-vehicle trajectories $\{T_v^{(k)}(t_v)\}$, where the trajectory $T_v^{(k)}(t_v)$ contains the sequential coordinates of agent $A_v^{(k)}$ at and before time t_v. Note that t_i should be earlier than t_v due to the latency. However, in this paper, we ignore the latency to explore how to integrate infrastructure information better and consider t_i equal to t_v.

The output is the specified target agent's future coordinates for time steps $t = t_v + 1, \cdots, t_{pred}$. To make the forecasting task more challenging, we predict longer trajectories and define the forecasting task as observing the past 50 frames ($5s$) and then predicting the future 50 frames ($5s$).

Evaluation Metrics and Analysis. In autonomous driving, there are often diverse possible future behaviors of traffic participants. Therefore, we output multiple possible future trajectories for each target agent for evaluation. Similar to Argoverse [5], we use the minimum Average Displacement Error (minADE), minimum Final Displacement Error (minFDE), and Missing Rate (MR) as the metrics to measure the prediction performance. We evaluate the model with Top-K predictions as our metrics, where $K = 6$.

Our approach is based on the Trajectory Forecasting Dataset (TFD), which involves receiving and fusing infrastructure data from an intersection environment with complicated traffic situations. There are several challenges to achieving better prediction performance. One of these challenges is to effectively utilize valuable infrastructure information to enhance the incomplete perception results of the vehicle side, which is limited due to the single-vehicle view. Another challenge to establish a proper social context by incorporating infrastructure-perceived agents for better reasoning about the future behaviors of the target agent. Finally, it is crucial to improve the encoding of vector maps and traffic light signals to better assist in prediction.

5.2. Offline-VIC Forecasting Task

The Offline-VIC Forecasting task aims to transfer knowledge extracted from various infrastructure sequences to predict ego-vehicle trajectories. During inference, the model can only utilize the ego-vehicle data and cannot access real-time infrastructure data, similar to the traditional trajectory forecasting task [5]. Similar to Online-VIC Forecasting, we define the prediction task as observing the past 50 frames (5 seconds) and predicting the future 50 frames (5 seconds). We measure the prediction results using minADE, minFDE, and MR metrics and evaluate the model with Top-K predictions, where K=6. The main challenge in solving this task is extracting appropriate knowledge from heterogeneous infrastructure data for transfer.

6. Experiments

6.1. VIC3D Tracking Benchmarks

In this section, we present the results of our extensive experiments, which include different fusion approaches, input modalities, and latency settings. The experiments are conducted on the sequential perception dataset (SPD), and the train/valid/test split ratio is set to 5:2:3. We only consider four classes of Car, Van, Bus and Truck and the objects located in a rectangular of [0, -39.68, 100, 39.68]. The results are summarized in Table 2 and visualized in Figure 5.

6.1.1 Baselines

The VIC3D Tracking problem can be tackled using three solutions: early fusion, middle fusion, and late fusion. Early fusion involves fusing infrastructure raw data, middle fusion fuses intermediate-level infrastructure data like feature maps, and late fusion fuses infrastructure perception results. Raw data contains all information but requires the highest transmission cost, while perception results consume the least amount of transmission cost but lose valuable information. We conducted experiments to evaluate the performance of these fusion solutions for VIC3D Tracking.

Solution with Middle Fusion. We implemented the FF-Tracking model and a simple middle fusion model to explore middle fusion with intermediate data. We first explain how to train the FF-Tracking model. We pre-trained the FF-Tracking model on the training part of the sequential perception dataset for 40 epochs without considering latency. The learning rate was set to 0.001, and the weight decay was set to 0.01. We fine-tuned the FF-Tracking model on the training part of thesequential perception dataset for 20 epochs by adding random latency. The learning rate was set to 0.001, and the weight decay was set to 0.01. We then applied V2VNet [32] as a simple middle fusion model to solve VIC3D Tracking and compared it with FF-Tracking. Compared to FF-Tracking, the V2VNet [32] only transmits a single feature and keeps the other configurations the same as FF-Tracking. We trained the model for 40 epochs with a learning rate of 0.001 and a weight decay of 0.01. Note that FF-Tracking incurs a higher transmission cost per second compared to simple middle fusion due to the requirement of transmitting additional feature flow. Furthermore, in $0ms$ latency, FF-Tracking degenerates into V2VNet, which suggests that FF-Tracking and V2VNet manifest equivalent tracking performance under $0ms$ latency conditions.

Table 2. **Evaluation Results for VIC3D Tracking on SPD at Different Latency Levels.** The "Vehicle Only" approach utilizes only ego-vehicle data, while "Concat fusion" combines pseudo images generated from point clouds. The evaluation of tracking performance employs three metrics: MOTA, MOTP, and IDS. Additionally, the transmission cost per second is assessed using the BPS metric. Notably, **in this experiment we only compare MOTA scores for the evaluation** and do not consider the MOTP and IDS scores for comparison.

Modality	Latency (ms)	Fusion Type	Fusion Method	MOTA ↑	MOTP	IDS	BPS (Byte/s) ↓
Image	0	Vehicle Only	-	10.96	58.69	2	0
	0	Late Fusion	Hungarian [18]	22.27	57.25	194	3.3×10^3
PointCloud	0	Vehicle Only	-	39.31	67.28	109	0
	0	Early Fusion	Concat	**56.03**	70.17	296	1.3×10^7
	0	Late Fusion	Hungarian [18]	53.18	72.35	273	3.3×10^3
	0	Middle Fusion	V2VNet [32]	54.75	69.76	222	6.2×10^5
PointCloud	0	Middle Fusion	**FF-Tracking**	54.75	69.76	222	6.2×10^5
PointCloud	200	Early Fusion	Concat	51.27	69.67	234	1.3×10^7
	200	Late Fusion	Hungarian [18]	50.32	71.58	260	3.3×10^3
	200	Middle Fusion	V2VNet [32]	48.38	68.99	231	6.2×10^5
PointCloud	200	Middle Fusion	**FF-Tracking**	**52.26**	69.64	225	1.2×10^6

Solution with Early Fusion. We implement early fusion with point cloud inputs. First, we convert the infrastructure point cloud into the ego-vehicle coordinate system. Then, we convert both infrastructure and ego-vehicle point clouds into pseudo-images and fused them. We used PointPillars [19] as a detector to generate 3D outputs and AB3DMOT [33] to track each object. We directly train and evaluate the detector with the fused point cloud. Additionally, we also evaluate the model with different latencies.

Solution with Late Fusion. To investigate the fusion effect with perception results, we implement late fusion using point cloud and image inputs. Specifically, we employ PointPillars [19] to locate and identify objects from both infrastructure sequential frames and ego-vehicle sequential frames. Additionally, we use ImvoxelNet [9] to perceive 2D objects from the infrastructure and ego-vehicle sequential images. Next, we transmit the infrastructure objects to the ego vehicle and fuse them with the ego-vehicle objects based on Euclidean distance measurements. Then we use AB3DMOT [33] to track the fused objects. Finally, we evaluate the model's performance with different latencies.

6.1.2 Analysis

V2X view vs. Single-vehicle view. In Table 2, we present the evaluation results for both fusion and no-fusion methods. When using point clouds as input, all fusion methods outperform the no-fusion strategy, even when there is a performance drop due to communication delay. For instance, with point cloud as input and 200ms latency, the early fusion method improves the MOTA (multiple object tracking accuracy) of vehicles by 11.96% (from 39.31% to 51.27%). Thus, vehicle-infrastructure cooperative perception can effectively enhance 3D tracking performance.

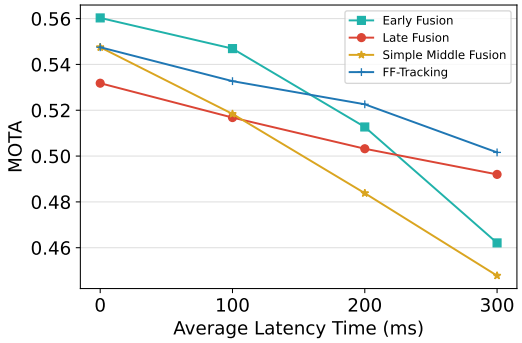

Figure 5. Comparison of VIC3D Tracking Baseline Models with Varying Latencies. Our proposed FF-Tracking model demonstrates greater robustness to latency when compared to the early fusion, late fusion, and simple middle fusion models.

Middle Fusion vs. Early Fusion&Late Fusion. We compared the performance of middle fusion, early fusion, and late fusion techniques using point cloud as input and with 0ms latency. Our results indicate that early fusion achieves higher tracking performance than middle fusion (56.03% vs. 54.75% MOTA), while middle fusion requires less transmission cost (6.2×10^5 Byte/s and 1.2×10^6 Byte/s vs. 1.3×10^7 Byte/s). Although late fusion requires the least transmission cost with 3.3×10^3 Byte/s, it still achieves lower tracking performance than middle fusion (53.18% vs. 54.75% MOTA). Our findings suggest that the middle fusion technique can achieve a better balance between transmission cost and tracking performance.

FF-Tracking can overcome the latency challenge. We present evaluation results of different fusion methods with a 200ms latency, as shown in Table 2. All the fusion methods show a performance drop as the latency increases. For

example, the early fusion has a 4.76% MOTA drop, and the simple middle fusion has a 6.37% MOTA drop when the latency is increased from $0ms$ to $200ms$. In comparison, our FF-Tracking model only has a 2.49% MOTA drop. We also present additional evaluation results at different latencies in Fig. 5. Our FF-Tracking model remains robust to all latencies, and importantly, outperforms early fusion by up to 4% MOTA at $300ms$ latency. Additionally, our FF-Tracking model achieves the best tracking performance when the latency reaches $200ms$.

6.2. Trajectory Forecasting Benchmarks

This section provides the baselines for solving the Online-VIC and Offline-VIC forecasting tasks on the trajectory forecasting dataset (TFD) with a train/val/test split of 5:2:3. The evaluation results are presented in Table 3.

6.2.1 Baselines

We choose TNT [40] and HiVT [41] as base models and train them with different configurations. We encode only the trajectories and vector maps that are within 50m of the ego vehicle. We evaluate the models on val part of 50,000 cooperative-view dataset. Specifically:

- **Baseline 1:** We only use ego-vehicle data and vector maps from the 50,000 cooperative-view data. We train the TNT [40] and HiVT [41] models for 30 epochs, and the other settings remain the same as the original.
- **Baseline 2:** We use vector maps and both ego-vehicle and infrastructure trajectories. We propose the PP-VIC framework, a simple yet effective hierarchical perception-prediction method for solving the Online-VIC Forecasting task. Firstly, we first use CBMOT [2] to fuse the infrastructure and ego-vehicle trajectories. We only fuse or add infrastructure trajectories that are relatively complete or have very high detection scores. Then, we apply TNT [40] and HiVT [41] to encode the trajectories and vector maps to generate future trajectories, respectively. We train the PP-VIC model for 30 epochs, and the other settings remain the same as the original.
- **Baseline 3:** We use ego-vehicle data and vector maps from the 50,000 cooperative-view data and additionally use 80,000 infrastructure-view trajectories. We pre-train the TNT [40] on the 80,000 infrastructure trajectories and then fine-tune the TNT [40] initialized with the pre-trained models. We train HiVT [41] in the same way.

6.2.2 Analysis

Online infrastructure trajectories are useful. Compared to baselines that do not use any infrastructure information, PP-VIC achieves lower minADE, minPDE, and MR. PP-VIC with TNT [40] achieves a minADE that is 3.74 lower than the TNT [40] model that does not use infrastructure trajectory information, and PP-VIC with [41] achieves a minADE that is 0.28 lower than the [41] model that does not use trajectory information. These results suggest that online utilization of infrastructure trajectories can improve forecasting performance.

Table 3. Evaluation results for different baselines. Using infrastructure trajectories can improve forecasting performance.

Using Infrastructure Trajectories	Prediction Model	K = 6		
		minADE ↓	minFDE ↓	MR ↓
×	TNT [40]	12.01	24.15	0.84
Online	TNT [40]	8.27	17.25	0.76
Offline	TNT [40]	4.36	9.23	0.62
×	HiVT [41]	1.55	2.59	0.36
Online	HiVT [41]	1.27	2.36	0.35
Offline	HiVT [41]	1.52	2.27	0.30

Offline infrastructure trajectories are useful. TNT [40] pretrained on extra infrastructure trajectories achieves a 7.65 minADE reduction compared to TNT [40] without the use of any infrastructure data. HiVT [41] pretrained on extra infrastructure trajectories achieves a 0.03 minADE reduction compared to HiVT [41] without the use of any infrastructure data. The experimental results demonstrate that extracting knowledge from infrastructure trajectories can effectively improve forecasting performance.

7. Conclusion

This paper presents a large-scale sequential V2X dataset, where all the data elements, including data frames, trajectories, vector maps, and traffic lights, are captured and generated from natural scenery. The paper introduces three new tasks for the vehicle-infrastructure cooperative autonomous driving community to better study how to utilize infrastructure information to improve sequential perception and trajectory forecasting ability. Several benchmarks are carefully designed for the fair evaluation of the introduced tasks. The experimental results demonstrate that infrastructure data can improve tracking and trajectory forecasting ability. Moreover, this paper proposes a novel FF-Tracking approach to solve the VIC3D Tracking problem.

Acknowledgements

This work was supported by Baidu Inc. through the Apollo-AIR Joint Research Center, and partially supported by the General Research Fund of HK under Grants No. 27208720 and No. 17200622. The authors would like to express their gratitude to the Beijing High-level Autonomous Driving Demonstration Area and Beijing Academy of Artificial Intelligence for their support throughout the dataset construction and release process.

References

[1] Eduardo Arnold, Mehrdad Dianati, Robert de Temple, and Saber Fallah. Cooperative perception for 3d object detection in driving scenarios using infrastructure sensors. *IEEE Transactions on Intelligent Transportation Systems*, 2020. 1, 2, 3

[2] Nuri Benbarka, Jona Schröder, and Andreas Zell. Score refinement for confidence-based 3d multi-object tracking. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8083–8090. IEEE, 2021. 8

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2, 5

[4] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021. 2, 4

[5] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8748–8757, 2019. 2, 4, 6

[6] Qi Chen, Sihai Tang, Qing Yang, and Song Fu. Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 514–524. IEEE, 2019. 3

[7] Christian Creß, Walter Zimmer, Leah Strand, Maximilian Fortkord, Siyi Dai, Venkatnarayanan Lakshminarasimhan, and Alois Knoll. A9-dataset: Multi-sensor infrastructure-based dataset for mobility research. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 965–970. IEEE, 2022. 2

[8] Jiaxun Cui, Hang Qiu, Dian Chen, Peter Stone, and Yuke Zhu. Coopernaut: End-to-end driving with cooperative perception for networked vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17252–17262, 2022. 3

[9] Anton Konushin Danila Rukhovich, Anna Vorontsova. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. *arXiv preprint arXiv:2106.01178*, 2021. 7

[10] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9710–9719, 2021. 2

[11] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 5

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012. 2

[13] John Houston, Guido Zuidhof, Luca Bergamini, Yawei Ye, Long Chen, Ashesh Jain, Sammy Omari, Vladimir Iglovikov, and Peter Ondruska. One thousand and one hours: Self-driving motion prediction dataset. In *Conference on Robot Learning*, pages 409–418. PMLR, 2021. 2

[14] Matthew Howe, Ian Reid, and Jamie Mackenzie. Weakly supervised training of monocular 3d object detectors using wide baseline multi-view traffic camera data. In *The British Machine Vision Conference (BMVC)*, 2021. 2, 3

[15] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *Advances in Neural Information Processing Systems*, 2022. 3

[16] Xinyu Huang, Peng Wang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The apolloscape open dataset for autonomous driving and its application. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2702–2719, 2019. 2

[17] Robert Krajewski, Julian Bock, Laurent Kloeker, and Lutz Eckstein. The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2118–2125. IEEE, 2018. 2

[18] Harold W. Kuhn. The hungarian method for the assignment problem. In *50 Years of Integer Programming*, 2010. 3, 7

[19] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 7

[20] Zixing Lei, Shunli Ren, Yue Hu, Wenjun Zhang, and Siheng Chen. Latency-aware collaborative perception. In *ECCV*, 2022. 3

[21] Yiming Li, Dekun Ma, Ziyan An, Zixun Wang, Yiqi Zhong, Siheng Chen, and Chen Feng. V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. *IEEE Robotics and Automation Letters*, 7(4):10914–10921, 2022. 2, 3

[22] Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, and Wenjun Zhang. Learning distilled collaboration graph for multi-agent perception. *Advances in Neural Information Processing Systems*, 34:29541–29552, 2021. 3

[23] Yen-Cheng Liu, Junjiao Tian, Nathaniel Glaser, and Zsolt Kira. When2com: Multi-agent perception via communication graph grouping. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 4106–4115, 2020. 3

[24] Yen-Cheng Liu, Junjiao Tian, Chih-Yao Ma, Nathan Glaser, Chia-Wen Kuo, and Zsolt Kira. Who2com: Collaborative perception via learnable handshake communication. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6876–6883. IEEE, 2020. 3

[25] Yuexin Ma, Xinge Zhu, Sibo Zhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6120–6127, 2019. 2

[26] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, et al. One million scenes for autonomous driving: Once dataset. *arXiv preprint arXiv:2106.11037*, 2021. 2

[27] Hang Qiu, Po-Han Huang, Namo Asavisanu, Xiaochen Liu, Konstantinos Psounis, and Ramesh Govindan. Autocast: scalable infrastructure-less cooperative perception for distributed collaborative driving. *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*, 2022. 3

[28] Pei Sun, Henrik Kretzschmar, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 2

[29] Federal Highway Administration US Department of Transportation. Next generation simulation (ngsim) vehicle trajectories and supporting data, 2016. 2

[30] Rodolfo Valiente, Mahdi Zaman, Sedat Ozer, and Yaser P Fallah. Controlling steering angle for cooperative self-driving vehicles utilizing cnn and lstm-based deep networks. In *2019 IEEE intelligent vehicles symposium (IV)*, pages 2423–2428. IEEE, 2019. 3

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[32] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Binh Yang, Wenyuan Zeng, James Tu, and Raquel Urtasun. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *ECCV*, 2020. 3, 6, 7

[33] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. Ab3dmot: A baseline for 3d multi-object tracking and new evaluation metrics. *arXiv preprint arXiv:2008.08063*, 2020. 5, 7

[34] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 2

[35] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2583–2589. IEEE, 2022. 2

[36] Zetong Yang, Yanan Sun, Shu Liu, and Jiaya Jia. 3dssd: Point-based 3d single stage object detector. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 5

[37] Xiaoqing Ye, Mao Shu, Hanyu Li, Yifeng Shi, Yingying Li, Guangjie Wang, Xiao Tan, and Errui Ding. Rope3d: The roadside perception dataset for autonomous driving and monocular 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 2

[38] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, and Zaiqing Nie. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21361–21370, 2022. 1, 2, 3

[39] Haibao Yu, Yingjuan Tang, Enze Xie, Jilei Mao, Jirui Yuan, Ping Luo, and Zaiqing Nie. Vehicle-infrastructure cooperative 3d object detection via feature flow prediction. *arXiv preprint arXiv:2303.10552*, 2023. 3, 4, 5

[40] Hang Zhao, Jiyang Gao, Tian Lan, Chen Sun, Ben Sapp, Balakrishnan Varadarajan, Yue Shen, Yi Shen, Yuning Chai, Cordelia Schmid, et al. Tnt: Target-driven trajectory prediction. In *Conference on Robot Learning*, pages 895–904. PMLR, 2021. 8

[41] Zikang Zhou, Luyao Ye, Jianping Wang, Kui Wu, and Kejie Lu. Hivt: Hierarchical vector transformer for multi-agent motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8823–8833, 2022. 8