SlimComm: Doppler-Guided Sparse Queries for Bandwidth-Efficient Cooperative 3-D Perception
Supplementary Material

1. Dataset Details

Vehicles in Ego Perception Range

Figure 1 shows the number of vehicles within the perception range of the ego vehicle. On average, OPV2V-R contains 16.74 ± 7.90 vehicles per frame, while Adver-City-R has 20.75 ± 7.28. This highlights that Adver-City-R exhibits higher traffic density, which increases the likelihood of occlusions and makes perception more challenging.

Scenario Statistics in Adver-City-R

As Table 1 shows, Adver-City-R covers five distinct scenario types that reflect common driving situations in both rural and urban environments. Scenario lengths range from about 23 s to 39 s, providing clips of manageable duration while still capturing diverse traffic interactions. The traffic density differs strongly across road types: rural scenarios typically involve around 10 vehicles within the ego perception range, whereas urban intersections exceed 30 on average. Speed distributions also vary: vehicles and CAVs travel faster in rural settings, while urban intersections are characterized by slower and denser traffic flows. These variations emphasize the heterogeneity of Adver-City-R and demonstrate its suitability for evaluating perception systems under diverse and challenging traffic conditions.

Road Type Splits in Adver-City-R and OPV2V-R

Adver-City-R enforces a strict road-type split: *urban intersections* occur only in the test set, *rural curved non-junction* roads only in validation, and *rural intersections*, *rural straight non-junction* roads, and *urban non-junction* roads only in training. This means that test scenarios involve road geometries that are unseen during training, leading to a more demanding evaluation. Combined with the generally higher traffic density and occlusion levels of Adver-City-R, this results in more challenging conditions compared to OPV2V-R. In contrast, OPV2V-R does not enforce such separation, and all road types appear across its splits.

Figure 2 shows the trajectories of all vehicles in representative scenarios from OPV2V-R and Adver-City-R. The routes are color-coded with a time gradient from the beginning to

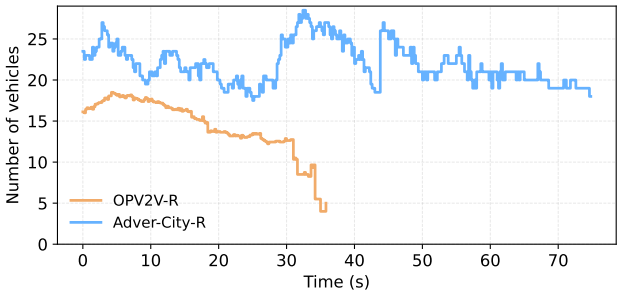
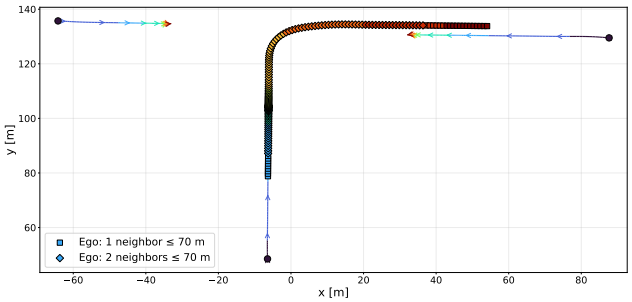
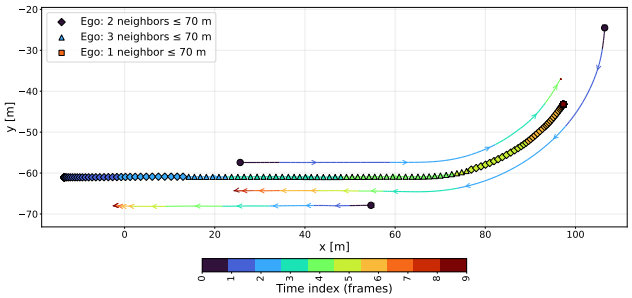


Figure 1. Number of vehicles within the perception range of the ego vehicle in the OPV2V-R and Adver-City-R test sets.



(a) Adver-City-R



(b) OPV2V-R

Figure 2. Ego vehicle routes and communication events in OPV2V-R and Adver-City-R. Trajectories are color-coded by time; markers indicate the number of neighbors within 70 m.

Scenario	Percentage(%)	Length(s) mean / std	Traffic density mean / std	Traffic Speed (km/h) mean / std	CAV speed (km/h) mean / std
rural_curved_non_junction	17.36	24.50 / 3.30	15.00 / 6.00	14.48 / 11.45	28.45 / 10.29
rural_intersection	27.28	22.80 / 0.00	10.00 / 4.00	15.64 / 11.37	10.36 / 12.95
rural_straight_non_junction	16.16	38.50 / 2.60	10.00 / 4.00	21.86 / 11.24	21.31 / 13.11
urban_intersection	18.14	29.70 / 7.80	33.00 / 15.00	12.92 / 12.72	10.95 / 11.32
urban_non_junction	21.05	25.60 / 1.10	27.00 / 12.00	13.69 / 11.96	25.58 / 12.67
overall	100.00	28.20 / 6.90	19.00 / 13.00	15.28 / 12.37	19.37 / 14.10

Table 1. Summary of Adver-City-R dataset statistics by road type. Traffic density means the number of vehicles spawned around the ego vehicle within a 140 m radius

the end of each scenario. Markers highlight the frames in which the ego vehicle detects neighboring vehicles within the **70 m communication range**. The marker shape indicates the number of neighbors: a square represents one, a diamond two, a triangle three, and a star four or more.

A clear structural difference can be observed between the two datasets. In OPV2V-R, vehicles frequently move along the same road segments, resulting in long periods of continuous communication. In contrast, Adver-City-R features vehicles approaching from different directions and meeting mainly at intersections, which concentrates communication events around junction areas.

2. Ego Query Generator Architecture

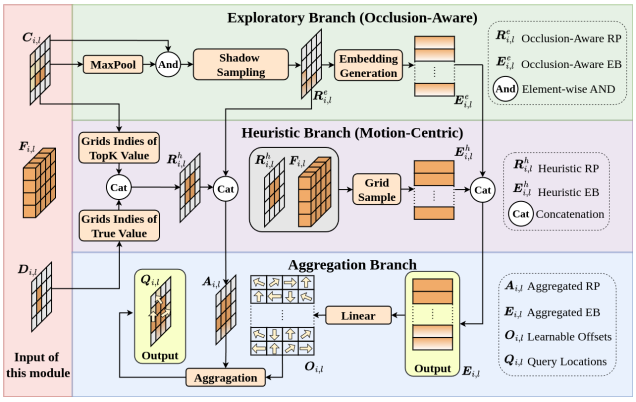


Figure 3. Detailed architecture of the Ego Query Generator, consisting of a motion-centric heuristic branch, an occlusion-aware exploratory branch, and an aggregation branch that combines reference points (RP) and embeddings (EB) into the final query set.

Figure 3 gives a compact overview of Ego Query Generator. Per scale, the module (i) selects HRP from dynamic cells and top-confidence cells, (ii) places ERP behind occluder peaks in the confidence map via shadow sampling, and (iii) concatenates both and applies a coarse offset followed by a 3×3 deformable halo to obtain the final sampling locations.

3. Qualitative Evaluation

3.1. Quality of RP and Query Locations

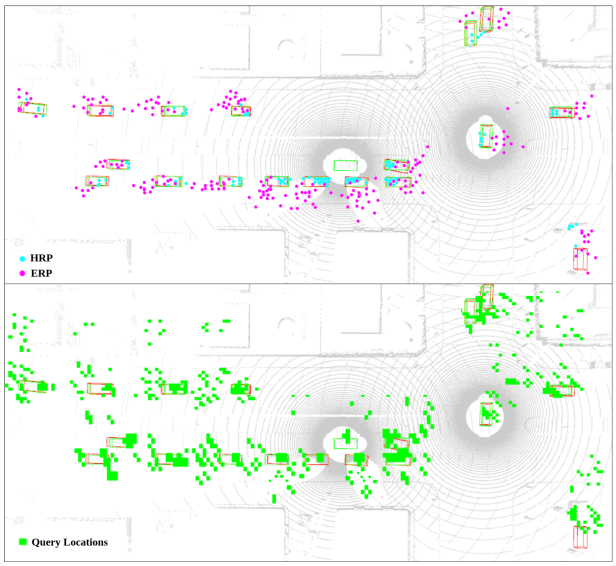


Figure 4. Qualitative distribution of HRP (cyan), ERP (magenta), and final query locations (green) with 3×3 halos in a dense scene.

Figure 4 shows complementary spatial patterns of the two reference points types and final query locations. As shown in the top panel, HRP concentrate on visible vehicles and high-confidence edges, reflecting the motion/score-driven selection used for object refinement. ERP are cast behind strong returns along the ego-to-occluder rays, spreading into partially or fully occluded regions. This complementarity expands coverage into occlusions while keeping precision on visible objects, which translates to fewer misses in cluttered areas without introducing extra false positives.

The bottom panel visualizes the final query locations. They are obtained by applying the two-stage offset to the union of HRP/ERP anchors: a coarse nudge to the anchor center followed by a 3×3 deformable halo. As a result, queries appear as compact 3×3 clusters around visible vehicles and extend into the occluder shadow regions suggested by ERP. Background road areas remain sparsely sampled.

This distribution concentrates sampling where evidence is strongest or likely hidden, while keeping the overall query budget low.

3.2. Qualitative Comparison with SOTA Methods

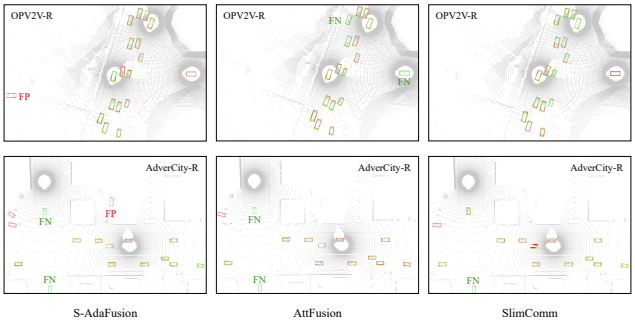


Figure 5. Visualization of detection results in two very dense scenarios. Green: ground truth bounding box. Red: prediction result.

Fig. 5 compares S-AdaFusion, AttFusion, and SlimComm on two occlusion-heavy scenes (OPV2V-R and AdverCity-R). As annotated in the figure, in OPV2V-R, S-AdaFusion covers most vehicles but produces a clear FP in clutter. AttFusion misses multiple vehicles near occluders (several FNs). SlimComm successfully detects all vehicles with precise and well-aligned bounding boxes. In AdverCity-R, under stronger occlusion, S-AdaFusion shows both FP and FN, and AttFusion accumulates additional FNs. SlimComm maintains the best alignment and the fewest total errors among the three, reflecting a stronger precision–recall balance under heavy occlusion.

4. Robustness against Asynchronous

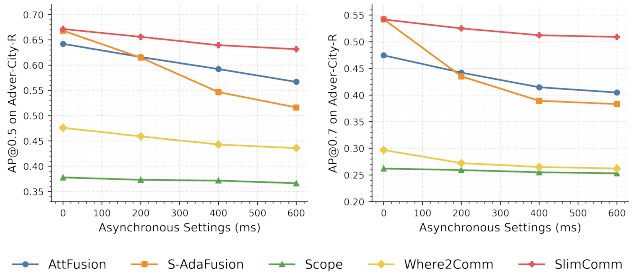


Figure 6. Robustness of different frameworks under localization asynchrony. Left: AP evaluated at IoU = 0.5. Right: AP evaluated at IoU = 0.7.

Figure 6 evaluates methods under injected delays up to 600 ms on Adver-City-R. Two patterns emerge: (i) models that maximize dense cross-attention at 0 ms achieve higher initial AP but are brittle to delay (steep negative slope); (ii) confidence/selection-driven schemes in Where2Comm and Scope transmit only high-confidence regions and therefore exhibit slower degradation. SlimComm inherits the benefits of both: reference/exploratory queries are placed from stable priors (Doppler motion and foreground confidence), while offset-based deformable attention and halo aggregation reduce reliance on exact temporal alignment. Consequently, SlimComm maintains a substantially flatter AP–delay curve compared to attention-heavy baselines while retaining high absolute AP.