RAALIGN: MULTI-AGENT FEATURE ALIGNMENT AND RELIABILITY-AWARE FUSION FOR VEHICLE–INFRASTRUCTURE COOPERATIVE OBJECT LOCALIZATION AND PERCEPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing V2X cooperative perception methods mainly rely on LiDAR-only or LiDAR-camera inputs, and remain vulnerable to adverse weather and cross-agent alignment errors. We introduce RAAlign, a reliability-aware LiDAR–4D Radar fusion framework for cooperative 3D object detection. RAAlign exchanges quality-gated complementary messages, promotes object-level consistency across agents and modalities using an InfoNCE objective, and compensates for residual offsets before weighting neighbor features using spatial and temporal reliability for attention aggregation. Experiments on V2X-R show that RAAlign improves average AP@0.7 from 69.05 to 72.99 on AttFuse and from 70.24 to 73.64 with MDD, with the largest weather gain reaching 2.24 percentage points in snow. These results provide evidence that reliability-aware cross-agent LiDAR–4D Radar fusion improves cooperative perception under adverse weather and imperfect alignment.

Keywords: V2X cooperative perception, LiDAR–4D Radar fusion, adverse-weather 3D object detection, cross-agent cross-modal consistency, reliability-aware alignment.

1 INTRODUCTION

Vehicle-to-everything (V2X) cooperative perception extends the ego vehicle's observations by aggregating information from nearby vehicles and roadside units (Zimmer et al., 2024; Zhang et al., 2024). Fog and snow can degrade LiDAR observations (Huang et al., 2025), while pose uncertainty can misalign cross-agent features and impair cooperative detection (Zhang et al., 2024).

Existing V2X 3D detectors commonly use LiDAR or LiDAR-camera inputs (Zimmer et al., 2024; Zhang et al., 2024). LiDAR provides detailed geometry for 3D localization, but fog, snow, and other strongly scattering conditions can reduce point density, introduce noise, and attenuate returns (Park et al., 2025). Cameras add semantic texture, although illumination changes, adverse weather, and occlusion can limit their observations (Yue et al., 2025; Zimmer et al., 2024). 4D Radar offers greater resilience to adverse weather, velocity measurements, and long-range observations (Chae et al., 2024; 2025). Its spatial resolution remains lower than that of LiDAR, limiting precise 3D boundary estimation when used alone (Chae et al., 2024). Cooperative LiDAR–4D Radar detection therefore requires deciding when, where, and how strongly radar evidence should compensate for degraded LiDAR geometry.

Coordinate transformation alone cannot ensure reliable alignment between intermediate features from different agents. GPS/IMU localization errors, communication latency, and synchronization offsets can leave these features misaligned (Zhang et al., 2024; Song et al., 2025). Extrinsic calibration drift in roadside cameras adds geometric uncertainty to vehicle–infrastructure fusion (Zhu et al., 2025). Direct fusion can then displace object responses, causing false positives, missed detections, or errors in 3D object localization (Zhang et al., 2024; Song et al., 2025). Neighbor features therefore require reliability assessment before aggregation, particularly when weather degradation and misalignment coexist.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We propose RAAlign, an alignment-error-robust and reliability-aware fusion method for cooperative LiDAR–4D Radar 3D object detection. We use V2X-R because its cooperative data and fusion baseline support evaluation under simulated adverse weather (Huang et al., 2025). This setting allows us to study complementary message transmission, consistent object representations across agents and modalities, and residual alignment uncertainty. RAAlign modifies the cross-agent cooperation stage as a plug-in component while retaining the detection backbone. The V2X-R pipeline comprises single-vehicle encoding, agent fusion, modality fusion, and detection prediction. RAAlign gates complementary messages, regularizes object embeddings during training, and scales aligned neighbor features before aggregation. The main contributions are summarized as follows:

- RAAlign supports cooperative LiDAR–4D Radar detection across fusion backbones.
- Reliability-gated messaging selects complementary evidence for communication.
- An object-level consistency loss constrains representations across agents and modalities.
- Alignment-aware reliability fusion alone improves Avg AP@0.7 on AttFuse w/MDD by 2.24 percentage points.

2 Related Work

Recent cooperative benchmarks provide real-world vehicle-to-vehicle data in V2V4Real (Xu et al., 2023) and vehicle–infrastructure observations in TUMTraf V2X (Zimmer et al., 2024). V2X-R adds cooperative LiDAR, camera, and 4D Radar sensing with simulated fog and snow (Huang et al., 2025). Single-agent research provides unified BEV representations for multi-sensor fusion (Liu et al., 2023), while MetaBEV addresses corrupted or missing LiDAR and camera inputs (Ge et al., 2023). These single-agent designs do not address the reliability of complementary LiDAR–4D Radar evidence exchanged between agents under adverse weather.

LiDAR–4D Radar detectors adapt radar contributions to weather (Chae et al., 2024) and exploit Doppler information for robust fusion (Chae et al., 2025). V2X-R uses radar-guided diffusion to denoise degraded LiDAR features (Huang et al., 2025). For communication, CodeFilling selects complementary messages according to agents' information demands and represents them with a codebook (Hu et al., 2024). HM-ViT models interactions among agents with heterogeneous LiDAR or camera inputs (Xiang et al., 2023). Selecting messages under adverse weather still requires separating useful LiDAR–4D Radar complementarity from sensor degradation while maintaining same-object consistency across agents and modalities.

Spatial and temporal misalignment also affect the reliability of received features. SparseAlign combines temporal context modeling, relative pose correction, and spatial fusion of sparse object queries (Yuan et al., 2025). TraF-Align uses predicted trajectories to align asynchronous features and maintain cross-agent semantic consistency (Song et al., 2025). Residual correspondence uncertainty and varying sensing quality can still make an aligned neighbor feature unreliable, leaving its appropriate contribution weight to fusion unclear.

3 Method

We study cooperative LiDAR–4D Radar 3D object detection in V2X scenes with N agents. The ego vehicle localizes and recognizes vehicles using its own observations and those of neighboring CAVs or RSUs. We retain the V2X-R pipeline of single-vehicle encoding, agent fusion, modality fusion, and detection prediction (Huang et al., 2025). RAAlign introduces reliability-gated complementary messaging, object-level consistency constraints, and reliability modulation within the cross-agent cooperation stage. This preserves the detector architecture so that performance changes can be attributed to the cooperation mechanism. Each agent's LiDAR and 4D Radar point clouds are encoded into BEV features, with the projected features defined as

$$\begin{gathered} F_i^L = P_L\left(X_i^L\right), \quad F_i^R = P_R\left(X_i^R\right), \\ F_i^L, F_i^R \in \mathbb{R}^{C\times H\times W}. \end{gathered} \tag{1}$$

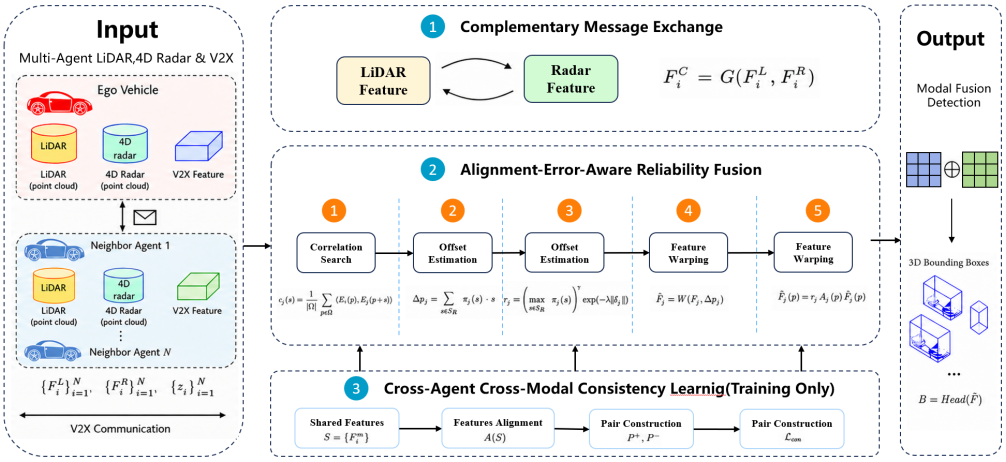


Figure 1: Overall framework of multi-agent LiDAR–4D Radar fusion for alignment-error-robust cooperative perception.

In (1), F_i^L and F_i^R denote the LiDAR and 4D Radar BEV features of agent i. The 1×1 convolutional projections P_L and P_R map both modalities into a shared channel space without adding detection branches. This space supports joint modeling of LiDAR geometric boundaries and weather-robust 4D Radar observations in cooperative messages.

Modality degradation and cross-agent alignment uncertainty jointly motivate the cooperation design. Fog and snow reduce LiDAR point density and introduce noise. 4D Radar provides more stable long-range and Doppler observations, but its spatial resolution limits boundary precision. Localization error, communication latency, and extrinsic drift can also leave neighbor BEV features misaligned after coordinate transformation. As shown in Fig. 1, RAAlign selects complementary observations for transmission, constrains object representations during training, and weights received features before fusion.

3.1 Reliability-Gated Complementary Messaging

Cross-agent exchange transmits LiDAR-to-LiDAR and radar-to-radar main features together with gated complementary responses between the two modalities. Differences between projected features can reflect useful target cues, weather noise, sparse sampling, or local mismatches. Their value for communication therefore depends on modality-specific quality responses, confidence-peak intervals, and local neighborhood statistics. A learnable gate uses these cues to select complementary evidence before transmission, retaining target information that one modality represents weakly. The gate and complementary message are defined by the following equations.

$$\begin{aligned} G_i &= \sigma\left(g\left(\left[F_i^L, F_i^R, D_i, S_i, q_i^L, q_i^R\right]\right)\right), \\ C_i &= G_i \odot h\left(D_i\right). \end{aligned} \tag{2}$$

In (2), D_i is a candidate complementary descriptor formed from projected responses, confidence-peak intervals, and local statistics of both modalities. The quality cues q_i^L and q_i^R come from the LiDAR and 4D Radar main features, respectively. Here, S_i denotes spatial neighborhood statistics, G_i is the sigmoid gating map, and C_i is the complementary message transmitted to other agents. The Complement branch gates cross-modal responses according to their compensation value, rather than simply concatenating features or transmitting feature differences unconditionally. In occluded, long-range, and adverse-weather regions, C_i can supplement weak ego observations. Strong noise, unstable confidence peaks, or inconsistent modality responses reduce the gate's transmission strength.

3.2 OBJECT-LEVEL CROSS-AGENT CROSS-MODAL CONSISTENCY

The exchanged evidence must also support consistent representations of the same object. Across agents, a vehicle can appear under different viewpoints, occlusion patterns, and sensor responses. LiDAR and 4D Radar further differ in spatial density, boundary sharpness, and velocity sensitivity. These differences can cause shared representations to drift after cross-agent aggregation. Such drift can affect AP@0.7, which is sensitive to errors in box centers, sizes, and headings. We impose object-level cross-agent cross-modal consistency during training to constrain these representations. For the BEV projection region R_k of object k, we pool and normalize the shared feature Z_i^m of modality m to obtain

$$z_{i,k}^m = \mathrm{Normalize}\left(\mathrm{RoIPool}\left(Z_i^m, R_k\right)\right). \tag{3}$$

Representations of the same object across agents and modalities form positive pairs. Representations of different objects form negative pairs in the shared embedding space. We constrain this space with the following object-level InfoNCE loss.

$$\begin{aligned} \mathcal{L}_{con} &= -\frac{1}{|\mathcal{P}|} \sum_{(a,p)\in\mathcal{P}} \log R_{a,p}, \\ R_{a,p} &= \frac{\exp\left(\mathrm{sim}\left(z_a, z_p\right)/\tau\right)}{\sum_{n\in\mathcal{A}(a)} \exp\left(\mathrm{sim}\left(z_a, z_n\right)/\tau\right)}. \end{aligned} \tag{4}$$

In (4), $\mathcal{P}$ contains positive pairs, and $\mathcal{A}(a)$ is the candidate sample set for anchor a. The ratio $R_{a,p}$ normalizes positive-pair similarity over all candidates; $\mathrm{sim}(\cdot)$ denotes cosine similarity, and τ is the temperature coefficient. The constraint encourages close semantic representations of the same physical object under cross-view, cross-modal, and cross-weather perturbations while allowing modality-specific low-level responses. This training objective supports a consistent object-level feature basis for reliability estimation and attention aggregation. The consistency loss is computed only during training and adds no communication or forward computation at inference. Appendix A gives the complete training objective in (11).

3.3 ALIGNMENT-AWARE RELIABILITY FUSION

Consistent object representations can still be spatially misaligned after the initial pose transformation. RAAlign accounts for residual offsets associated with localization noise, time delay, and local extrinsic errors by modulating neighbor features before fusion. Ego and neighbor features are mapped into a low-dimensional embedding space, and candidate translations form a set S_R within search radius R. For each displacement s, the average correlation between ego embeddings and translated neighbor embeddings is

$$c_i(s) = \frac{1}{|\Omega|} \sum_{p\in\Omega} \left\langle E_e(p), E_i(p+s) \right\rangle. \tag{5}$$

Softmax converts the candidate correlations into a displacement probability distribution. A soft-argmax yields a continuous residual displacement for each neighboring agent.

$$\begin{aligned} \pi_i(s) &= \mathrm{softmax}\left(c_i(s)/\tau_s\right), \\ \Delta_i &= \sum_{s\in S_R} \pi_i(s)\cdot s. \end{aligned} \tag{6}$$

Equations (5) and (6) make residual alignment estimation differentiable and avoid discrete jumps from hard matching. A sharp correlation peak indicates a more concentrated spatial correspondence, whereas a flat or multimodal distribution suggests alignment uncertainty. RAAlign uses spatial peak confidence and a time-delay prior to determine a global reliability weight for each neighbor.

$$
\begin{aligned}
r_i^{sp} &= \left(\max_{s \in S_R} \pi_i(s) \right)^{\gamma}, \\
r_i^{time} &= \exp\left(-\lambda |\delta_i|\right), \\
r_i &= r_i^{sp} \cdot r_i^{time}.
\end{aligned}
\tag{7}
$$

In (7), r_i^{sp} measures spatial reliability from correlation-peak strength, and r_i^{time} measures temporal reliability. The time offset δ_i is defined between neighbor i and the ego agent. Their product r_i sets the global reliability of that neighbor. Alignment quality also varies across BEV locations within a neighbor feature map. After continuous displacement compensation, we compute point-wise similarity to obtain local reliability modulation.

$$
\begin{aligned}
\widehat{Y}_i &= \mathrm{Warp}\left(Y_i, \Delta_i\right), \\
A_i(p) &= \sigma\left(\alpha \cdot s_i(p) - \beta\right), \\
s_i(p) &= \cos\left(\phi\left(Y_e(p)\right), \phi\left(\widehat{Y}_i(p)\right)\right).
\end{aligned}
\tag{8}
$$

$$
Y_i^{rel}(p) = r_i \cdot A_i(p) \cdot \widehat{Y}_i(p). \tag{9}
$$

In (8), $\mathrm{Warp}(\cdot)$ compensates neighbor features using the estimated continuous displacement. The similarity $s_i(p)$ compares the ego feature with the compensated neighbor feature at location p, and $A_i(p)$ gives the corresponding local reliability. Equation (9) scales neighbor features by both global agent-level reliability and local position-level reliability. Fusion intensity thus varies with global alignment confidence and local spatial similarity. The resulting multi-agent feature sequence is passed to standard scaled dot-product attention.

$$
\begin{aligned}
\mathrm{Attn}(Q, K, V) &= \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V, \\
H(p) &= \mathrm{Attn}\left(W_Q X(p), W_K X(p), W_V X(p)\right).
\end{aligned}
\tag{10}
$$

RAAlign supplies reliability-modulated BEV features to the attention operator in (10), preserving compatibility with intermediate-fusion backbones such as AttFuse and InterFusion. Appendix A provides the complete fusion procedure in Algorithm 1.

4 Experiments

4.1 Experimental Details and Metrics

We evaluate cooperative 3D detection on V2X-R under normal, fog, and snow conditions using LiDAR and 4D Radar inputs. The experiments follow the original data split, evaluation range, and weather settings (Huang et al., 2025). The baseline pipeline encodes both modalities into BEV features, and camera images are excluded. Appendix A.3 provides the sensor configuration and data formats used in this evaluation.

Unless otherwise specified, models are trained for 20 epochs on two NVIDIA RTX 3090 GPUs with Adam and a batch size of 2. All variants share data loading, coordinate transformation, detection heads, and inference post-processing in V2X-R/OpenCooD (Xu et al., 2022; Huang et al., 2025).

We report detection AP at IoU thresholds of 0.3, 0.5, and 0.7. Avg AP denotes the mean across the three weather conditions. Parameter counts, FLOPs, model inference latency, and end-to-end latency characterize computational cost. Inference and latency measurements use a batch size of 1.

4.2 Baseline Comparisons and Weather Robustness

The comparisons include single-modality AttFuse, multimodal AttFuse, InterFusion, and AttFuse w/MDD, together with variants incorporating the proposed modules. Complement gates communication messages, Consistency adds object-level supervision during training, and the RAAlign fusion

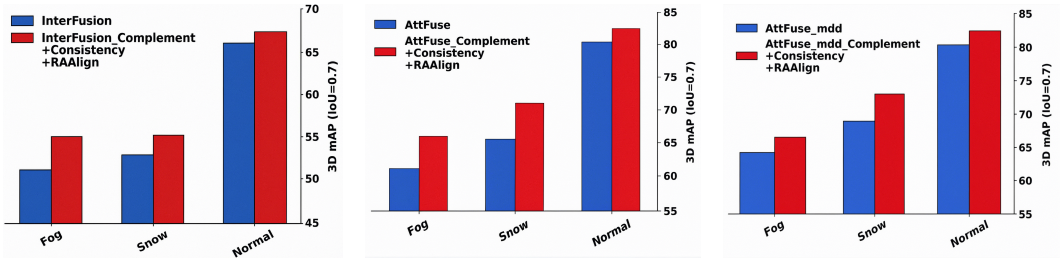
module modulates neighbor features. Table 1 reports their multi-weather average AP under the shared evaluation protocol.

Table 1: Overall comparison of different methods under the multi-weather average setting.

Method	Modality	Avg AP@0.3	Avg AP@0.5	Avg AP@0.7
AttFuse	L	66.34	64.57	54.52
AttFuse	4DR	80.44	73.18	47.60
AttFuse w/MDD	L+4DR	85.08	83.46	70.24
AttFuse	L+4DR	85.20	83.10	69.05
AttFuse + RAAlign	L+4DR	87.55	85.89	71.57
AttFuse + Comp. + Cons. + RAAlign (Ours)	L+4DR	88.59	86.42	72.99
InterFusion	L+4DR	82.91	79.69	55.86
InterFusion + RAAlign	L+4DR	85.05	81.52	57.23
InterFusion + Comp. + Cons. + RAAlign (Ours)	L+4DR	86.26	82.57	59.46
AttFuse w/MDD + Comp. + Cons. + RAAlign (Ours)	L+4DR	87.62	85.41	73.64

LiDAR-only and 4D Radar-only AttFuse achieve Avg AP@0.7 of 54.52 and 47.60, respectively, compared with 69.05 for multimodal AttFuse. Adding all three modules raises multimodal AttFuse to 72.99, a gain of 3.94 percentage points. The full framework also improves InterFusion from 55.86 to 59.46 and AttFuse w/MDD from 70.24 to 73.64. These gains of 3.60 and 3.40 points extend the improvement to a different fusion backbone and a baseline with modality denoising.

The weather-specific comparisons in Fig. 2 show higher AP@0.7 with the full framework under normal, fog, and snow for all three backbone configurations. Improvements on InterFusion and AttFuse support applicability across the evaluated fusion designs, while those on AttFuse w/MDD persist alongside modality denoising. Across these configurations, the adverse-weather gains are accompanied by higher normal-weather performance.



(a) InterFusion (b) AttFuse (c) AttFuse w/MDD

Figure 2: AP@0.7 under normal, fog, and snow for (a) InterFusion, (b) AttFuse, and (c) AttFuse w/MDD. Blue bars denote each baseline, and red bars denote the same baseline with Complement, Consistency, and RAAlign.

Table 2: Detection results of representative methods under different weather conditions.

Method	Modality	3D mAP@Normal			3D mAP@Fog			3D mAP@Snow		
		IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7
AttFuse	L	89.37	87.41	77.42	66.61	64.95	55.42	43.05	41.35	30.72
AttFuse	4DR	80.19	72.18	46.92	81.65	74.88	48.41	79.47	72.49	47.48
AttFuse w/MDD	L+4DR	88.08	87.77	80.34	83.34	80.62	63.12	83.82	81.98	67.25
AttFuse + RAAlign	L+4DR	90.35	89.61	80.49	85.05	82.64	64.37	87.26	85.43	69.86
AttFuse+Comp.+Cons.+RAAlign	L+4DR	90.66	90.21	81.97	87.56	83.35	65.98	87.54	85.70	71.02
AttFuse w/MDD+Comp.+Cons.+RAAlign	L+4DR	90.02	88.72	82.41	86.11	82.67	65.27	86.73	84.83	73.25

Table 2 provides the per-weather results at all three IoU thresholds. LiDAR-only AP@0.7 falls from 77.42 under normal conditions to 30.72 in snow, whereas the 4D Radar-only result changes from 46.92 to 47.48. For AttFuse w/MDD, the full framework raises AP@0.7 from 80.34 to 82.41 under normal conditions, from 63.12 to 65.27 in fog, and from 67.25 to 73.25 in snow. The largest

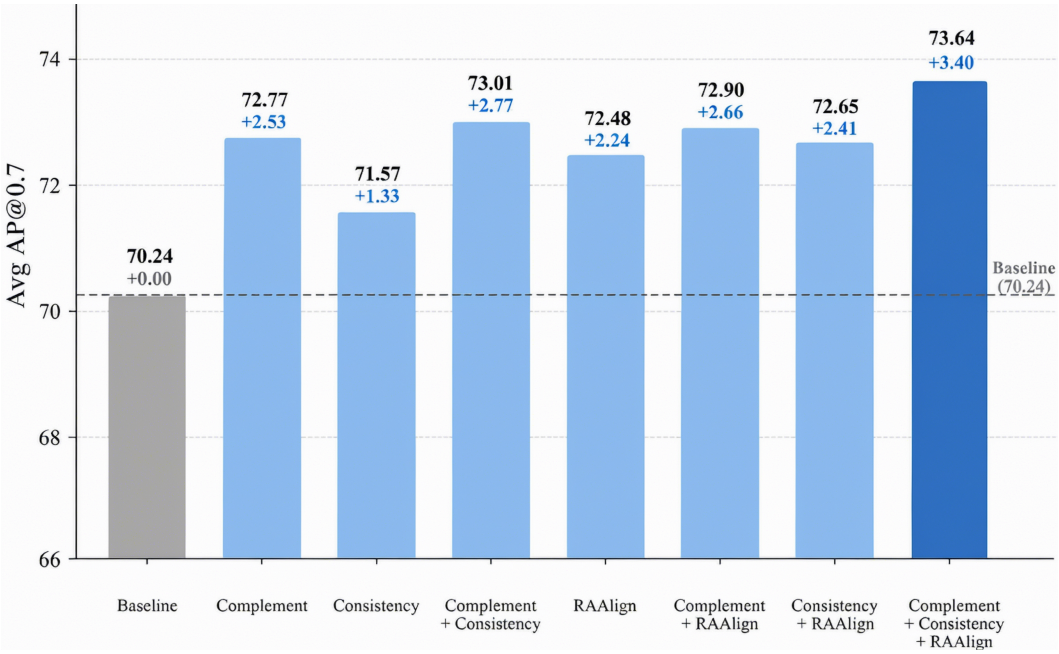


Figure 3: Module contributions on AttFuse w/MDD, measured by Avg AP@0.7 across normal, fog, and snow conditions. Labels above each bar give AP and its gain in percentage points over the baseline of 70.24, marked by the dashed line. The dark blue bar denotes the full configuration.

gain is 6.00 percentage points in snow, where the single-modality results show pronounced LiDAR degradation.

These experiments retain the original GPS/IMU noise setting and include no dedicated pose-error or communication-delay perturbation tests. Higher AP@0.7 supports improved detection at a stricter IoU threshold, but does not isolate robustness to alignment errors.

4.3 Ablation Studies

To examine the contributions underlying these gains, we evaluate all module combinations on AttFuse w/MDD. Table 3 reports results at three IoU thresholds, and Fig. 3 summarizes Avg AP@0.7.

Table 3: Module ablation on the MDD backbone.

Complement	Consistency	RAAlign	Avg AP@0.3	Avg AP@0.5	Avg AP@0.7	ΔAP@0.7
-	-	-	85.08	83.46	70.24	+0.00
✓	-	-	87.57	85.27	72.77	+2.53
-	✓	-	86.57	84.17	71.57	+1.33
✓	✓	-	86.96	84.39	73.01	+2.77
-	-	✓	87.71	85.51	72.48	+2.24
✓	-	✓	87.84	85.50	72.90	+2.66
-	✓	✓	87.84	85.80	72.65	+2.41
✓	✓	✓	87.62	85.41	73.64	+3.40

Each module improves Avg AP@0.7 over the 70.24 baseline. Complement provides the largest individual gain of 2.53 percentage points, followed by the alignment-aware fusion module at 2.24 and Consistency at 1.33. Complement with Consistency is the strongest two-module combination at 73.01. Adding alignment-aware fusion increases Avg AP@0.7 by another 0.63 points to 73.64, yielding an overall gain of 3.40 points. The full configuration ranks highest at AP@0.7, while other variants attain higher Avg AP at IoU thresholds of 0.3 and 0.5.

The module comparisons also reveal their inference cost, with Table 4 reporting accuracy, computation, and latency across the evaluated backbone configurations. For InterFusion, adding Complement and Consistency to the RAAlign fusion module improves Avg AP@0.7 from 57.23 to 59.46. Model inference latency increases from 26.72 to 27.87 ms, parameters from 25.171 M to 25.186 M, and FLOPs from 47.518 G to 48.471 G. Consistency is used only during training, while Complement adds a gating branch at inference.

Table 4: Efficiency and end-to-end latency comparison of representative methods.

Method	Avg AP@0.7	Params(M)	FLOPs(G)	Model Latency(ms)	E2E Latency(ms)
AttFuse + RAAlign	71.57	25.163	230.081	41.00	73.40
AttFuse + Comp.+Cons.+RAAlign	72.99	25.180	234.803	47.28	79.80
AttFuse + MDD	70.24	25.144	250.644	80.36	92.60
AttFuse + MDD + Comp.+Cons.+RAAlign	73.64	25.218	255.742	87.75	98.90
InterFusion + RAAlign	57.23	25.171	47.518	26.72	58.30
InterFusion + Comp.+Cons.+RAAlign	59.46	25.186	48.471	27.87	60.10

The full AttFuse and AttFuse w/MDD configurations achieve Avg AP@0.7 of 72.99 and 73.64, with end-to-end latencies of 79.80 and 98.90 ms, respectively. The MDD configuration thus gains 0.65 AP points at an additional 19.10 ms of end-to-end latency. These measurements characterize the accuracy–latency trade-off on our experimental platform within the V2X-R/OpenCooD pipeline.

5 Conclusion and Discussion

RAAlign combines gated complementary messages, object-level consistency, and reliability-aware fusion for cooperative LiDAR–4D Radar 3D object detection. On the evaluated V2X-R configurations, it improved detection under normal, fog, and snow with AttFuse and its MDD-enhanced variant. Ablations supported the combined use of these modules, while the measured latency increase was modest on our experimental platform.

Limitations and future work. All evaluations are based on simulated V2X-R data under the original cooperative protocol. RAAlign estimates a residual 2D translation for each neighbor in BEV and uses a time-delay prior to weight its features. It does not directly estimate rotation or spatially varying displacements caused by object motion during a delay. Robustness to packet loss, sensor drift, and asynchronous sampling has not been tested. Future work will target lower end-to-end latency for real-time autonomous driving.

References

Y. Chae, H. Kim, and K.-J. Yoon. Towards robust 3D object detection with LiDAR and 4D Radar fusion in various weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15162–15172, 2024.

Y. Chae, H. Park, H. Kim, et al. Doppler-aware LiDAR-RADAR fusion for weather-robust 3D detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27197–27208, 2025.

C. Ge, J. Chen, E. Xie, et al. MetaBEV: Solving sensor failures for 3D detection and map segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8721–8731, 2023.

Y. Hu, J. Peng, S. Liu, et al. Communication-efficient collaborative perception via information filling with codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15481–15490, 2024.

X. Huang, J. Wang, Q. Xia, et al. V2X-R: Cooperative LiDAR-4D Radar fusion with denoising diffusion for 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27390–27400, 2025.

Z. Liu, H. Tang, A. Amini, et al. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *Proceedings of the 2023 IEEE International Conference on Robotics and Automation*, pp. 2774–2781, 2023.

J. Park, K. Kim, and H. Shim. Rethinking data augmentation for robust LiDAR semantic segmentation in adverse weather. In *Computer Vision – ECCV 2024*, pp. 320–336, 2025.

Z. Song, L. Yang, F. Wen, et al. TraF-Align: Trajectory-aware feature alignment for asynchronous multi-agent perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12048–12057, 2025.

H. Xiang, R. Xu, and J. Ma. HM-ViT: Hetero-modal vehicle-to-vehicle cooperative perception with vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 284–295, 2023.

R. Xu, H. Xiang, X. Xia, et al. OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *Proceedings of the 2022 IEEE International Conference on Robotics and Automation*, pp. 2583–2589, 2022.

R. Xu, X. Xia, J. Li, et al. V2V4Real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13712–13722, 2023.

Y. Yuan, Y. Xia, D. Cremers, et al. SparseAlign: A fully sparse framework for cooperative object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22296–22305, 2025.

J. Yue, Z. Lin, X. Lin, et al. RobuRCDet: Enhancing robustness of radar-camera fusion in bird's eye view for 3D object detection. In *International Conference on Learning Representations*, pp. 12726–12741, 2025.

J. Zhang, K. Yang, Y. Wang, et al. ERMVP: Communication-efficient and collaboration-robust multi-vehicle perception in challenging environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12575–12584, 2024.

Y. Zhu, Z. Wang, and Y. Wang. MamV2XCalib: V2X-based target-less infrastructure camera calibration with state space model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26696–26705, 2025.

W. Zimmer, G. A. Wardana, S. Sritharan, et al. TUMTraf V2X cooperative perception dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22668–22677, 2024.

A APPENDIX

A.1 RELIABILITY-AWARE FUSION PROCEDURE

Algorithm 1 summarizes the residual alignment and reliability-aware fusion operations defined in (5)–(10).

Algorithm 1 Alignment-error-sensitive Reliability-aware Adaptive Fusion process

Input: Ego BEV feature Y_e; Neighbor BEV features Y_i $(i = 1, \ldots, N-1)$; Time-delay prior δ_i $(i = 1, \ldots, N-1)$; Embedding function φ; Search radius R; Temperature τ; Confidence power γ; Time-delay coefficient λ; Local-map coefficients α, β.
Output: fused ego BEV feature H_e.

$E_e \leftarrow \mathrm{Normalize}(\varphi(Y_e))$
$X \leftarrow \{Y_e\}$
$S_R \leftarrow \{(d_x, d_y) \mid d_x, d_y \in [-R, R] \cap \mathbb{Z}\}$
for $i \leftarrow 1, 2, \ldots, N-1$ **do**
  $E_i \leftarrow \mathrm{Normalize}(\varphi(Y_i))$
  for each $s \in S_R$ **do**
    $c_i(s) \leftarrow |\Omega_s|^{-1} \cdot \sum_{p \in \Omega_s} \langle E_e(p), E_i(p+s) \rangle$
  end for
  $\pi_i(s) \leftarrow \exp(c_i(s)/\tau) / \sum_{u \in S_R} \exp(c_i(u)/\tau)$
  $\Delta_i \leftarrow \sum_s \pi_i(s) \cdot s$
  $r_i^{sp} \leftarrow (\max_{s \in S_R} \pi_i(s))^{\gamma}$
  $r_i^{time} \leftarrow \exp(-\lambda |\delta_i|)$
  $r_i \leftarrow r_i^{sp} \cdot r_i^{time}$
  $\widehat{Y}_i \leftarrow \mathrm{Warp}(Y_i, \Delta_i)$
  if local reliability map is used **then**
    $\widehat{E}_i \leftarrow \mathrm{Normalize}(\varphi(\widehat{Y}_i))$
    $sim_i(p) \leftarrow \cos(E_e(p), \widehat{E}_i(p))$
    $A_i(p) \leftarrow \sigma(\alpha \cdot sim_i(p) - \beta)$
    $Y_i^{rel}(p) \leftarrow r_i \cdot A_i(p) \cdot \widehat{Y}_i(p)$
  else
    $Y_i^{rel}(p) \leftarrow r_i \cdot \widehat{Y}_i(p)$
  end if
  $X \leftarrow X \cup \{Y_i^{rel}\}$
end for
$X \leftarrow \mathrm{Stack}(X)$
$Q, K, V \leftarrow \mathrm{Linear}(X)$
$H \leftarrow \mathrm{Attn}(X, X, X) = \mathrm{Softmax}(QK^T/\sqrt{d})V$
$H_e \leftarrow H[ego]$
return H_e

The algorithm accepts ego and neighbor BEV features and returns a fused ego BEV feature, preserving the interface of the V2X-R pipeline. Residual displacement is estimated from correlations in a low-dimensional embedding space, and the resulting distribution determines spatial reliability. When enabled, the local reliability map further suppresses misaligned or weakly correlated BEV locations. Neighbor features are thus weighted by estimated reliability before attention aggregation when weather degradation and alignment uncertainty coexist.

A.2 TRAINING OBJECTIVE AND EVALUATION

The training objective combines the classification and 3D box regression losses with the optional MDD denoising loss and the object-level consistency loss:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{reg} + \lambda_{mdd}\mathcal{L}_{mdd} + \lambda_{con}\mathcal{L}_{con}. \tag{11}$$

In (11), $\mathcal{L}_{cls}$ and $\mathcal{L}_{reg}$ denote the classification and 3D box regression losses. The multimodal denoising term $\mathcal{L}_{mdd}$ is included when the MDD branch is enabled, while $\mathcal{L}_{con}$ imposes object-level cross-agent cross-modal consistency. Complement gates communication content, RAAlign modulates neighbor features along the fusion path, and Consistency constrains object representations during training.

Main results evaluate the full framework, while weather-specific results examine LiDAR–4D Radar complementarity and reliability-aware fusion under degradation. Module ablations and stepwise integration isolate the module contributions, and efficiency analysis measures deployment cost.

A.3 SENSOR CONFIGURATION AND DATA FORMATS

V2X-R contains 12,079 scenarios, 37,727 LiDAR/4D radar point-cloud frames, 150,908 images, and 170,859 annotated 3D vehicle boxes (Huang et al., 2025). Table 5 lists sensor specifications and GPS/IMU noise settings; Table 6 lists data formats and per-sample attributes. Camera entries describe the dataset configuration; images are excluded from our experiments.

Table 5: V2X-R sensor configuration (Huang et al., 2025).

Sensors	**Details**
4x Camera	4 RGB units, positions: (2.5,0,1.0,0), (0.0,0.3,1.8,100), (0.0,-0.3,1.8,-100), (-2.0,0.0,1.5,180)
1x LiDAR	64 channels, 120 m range, -25° to 2° vertical FOV, 0.02 noise standard deviation, 20 Hz rotation frequency
1x 4D Radar	150 m range, 120° horizontal FOV, 30° vertical FOV
GPS & IMU	Vehicle GNSS: altitude noise 0.001 m; vehicle IMU: heading noise 0.1°, speed noise 0.2 m/s; RSU GNSS: altitude noise 0.05 m

Table 6: V2X-R data formats and attributes (Huang et al., 2025).

Sensor	**Data Structure**	**File Type**	**Attributes**
LiDAR	Point cloud	.pcd	$N \times 4$; [x, y, z, intensity]
4D Radar	Point cloud	.pcd	$N \times 4$; [x, y, z, velocity]
Camera	Image	.png	$800 \times 600 \times 3$; [R, G, B]

A.4 STEPWISE MODULE INTEGRATION

Table 7 examines module integration on AttFuse, including the baseline, RAAlign alone, and variants that add Complement, Consistency, or both to RAAlign.

On AttFuse, RAAlign raises Avg AP@0.7 from 69.05 to 71.57. Adding either Complement or Consistency increases this value to 72.13, and combining all three modules reaches 72.99. Relative to the baseline, the full configuration adds 0.074 M parameters and increases model inference latency from 38.67 to 47.28 ms. The accuracy gain therefore comes with little parameter growth but a measurable latency cost.

Table 7: Stepwise module integration on the AttFuse backbone.

Stage	**Model Variant**	**Avg AP@0.3**	**Avg AP@0.5**	**Avg AP@0.7**	**Params(M)**	**FLOPs(G)**	**Latency(ms)**	**△AP@0.7**
S0	Baseline	85.20	83.10	69.05	25.106	229.705	38.67	+0.00
S1	+ RAAlign	87.55	85.89	71.57	25.163	230.081	41.00	+2.52
S2	+ Complement + RAAlign	87.93	85.91	72.13	25.180	234.803	43.12	+3.08
S3	+ Consistency + RAAlign	87.96	86.24	72.13	25.163	230.081	42.11	+3.08
S4	+ Complement + Consistency + RAAlign	88.59	86.42	72.99	25.180	234.803	47.28	+3.94