TEMPORAL ASSEMBLY OF RECEIVED REGIONAL FEATURES FOR COLLABORATIVE PERCEPTION UNDER COUPLED DELAY AND LOSS

Anonymous authors
Paper under double-blind review

ABSTRACT

Collaborative perception extends an ego vehicle's view using observations from neighboring vehicles, but depends on timely communication. When vehicle-to-vehicle (V2V) communication degrades, delay and packet loss can jointly leave a neighbor's current update incomplete or unavailable at a perception deadline, potentially reducing detection accuracy. To study this phenomenon, we introduce CoDeV2V, a newly collected long-sequence multi-vehicle LiDAR dataset. CoDeV2V models each V2V link with a condition—good, bad, or outage—that can persist across updates. The link condition jointly affects transmission delay and loss; an outage makes the affected neighbor's update unavailable and may block several consecutive updates if it persists. Consequently, current neighbor information may be incomplete or unavailable at the perception deadline, while retained information becomes stale. Existing methods address aspects of communication degradation, but on CoDeV2V the evaluated methods still lose substantial detection accuracy compared with complete and timely neighbor information. To mitigate the remaining loss in detection accuracy, we propose AssemblyFusion. AssemblyFusion assembles regional features received from current and earlier observations, using how old the information is and where it is available to guide fusion with ego features. The ego vehicle also selects detections from each neighbor's latest available observation, updates older detections using motion estimates, and combines the selected detections with predictions from fused features. Together, regional feature assembly uses the portions of neighbor updates that arrive by the deadline, while motion compensation updates historical detections when a neighbor's current detections are unavailable, mitigating the impact of coupled delay and loss on perception. Under severe coupled communication degradation on the 5,400-frame CoDeV2V test split, AssemblyFusion achieves 53.20 $\pm$ 0.46 BEV AP@0.7 across three communication realizations, 6.70 points above CoAlign. Ablations show higher accuracy with motion-compensated historical detections than with discarded or uncorrected historical detections.

Keywords: Collaborative perception; V2V communication; 3D object detection; coupled delay and packet loss; information freshness; decodability-first admission.

1 INTRODUCTION

Single-vehicle perception relies on onboard sensors and is constrained by limited sensing range, occlusion, and sparse distant observations. Multi-agent collaborative perception shares raw observations, intermediate features, or object-level outputs among vehicles or roadside infrastructure and fuses the shared information with ego-vehicle perception (Xu et al., 2022b). A neighboring vehicle may see traffic hidden from the ego vehicle at an intersection, or provide denser observations of a target far from the ego vehicle but closer to the neighbor (Xu et al., 2022b; Wang et al., 2020). Collaboration therefore extends perception coverage and supplements observations within the ego vehicle's sensing region. These benefits support collaborative 3D detection and motion prediction (Wang et al., 2020), bird's-eye-view (BEV) semantic segmentation (Xu et al., 2023), tracking and trajectory forecasting (Yu et al., 2023), and cooperative systems integrating perception with online mapping, occupancy prediction, and planning (Yu et al., 2025).

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.

Collaborative perception uses information from neighboring vehicles to supplement ego-vehicle observations. Measurements show longer delays and higher packet loss when buildings obstruct communication between vehicles (Wang et al., 2021); highway measurements reveal correlated interruptions that average packet delivery rates alone cannot characterize (Renda et al., 2016). A closely related modeling framework is the joint hidden Markov model of Salvo Rossi et al. (2006) for periodically transmitted packets. The link condition follows a Markov chain: the previous condition affects the next, and each condition determines both loss probability and the delay distribution of delivered packets. The shared condition couples delay and loss, while condition persistence captures dependence across transmissions. After fitting to measured traffic, generated sequences capture loss, delay, and temporal-correlation statistics. Iannello et al. (2005) also evaluates this modeling approach on heterogeneous networks including wireless links. This shared-condition structure provides a basis for our communication model. Loss-pattern and delay metrics describe consecutive losses, their spacing, and differences in packet transmission times (Koodli & Ravikanth, 2002; Demichelis & Chimento, 2002); V2X reliability emphasizes delivery within a specified time (ETSI, 2020).

To study collaborative perception under degraded V2V communication, we introduce CoDeV2V, a newly collected long-sequence multi-vehicle LiDAR dataset. CoDeV2V models each V2V link with a simulated condition—good, bad, or outage—that can persist across updates and jointly affects transmission delay and loss. Outages can block consecutive updates, leaving current neighbor information unavailable while retained observations grow stale. Long sequences preserve observations before and after interruptions, enabling methods to be compared on the same driving observations, evaluation targets, and communication-degradation conditions.

Existing methods mitigate the effects of communication delay on collaborative perception. SyncNet estimates current features and collaboration attention from historical information (Lei et al., 2022); LRCP uses predicted motion to guide asynchronous BEV feature fusion (Wang & Nordström, 2025); and CoDynTrust combines motion compensation with uncertainty-based feature weighting (Xu et al., 2025b). These approaches address stale neighbor information through prediction, compensation, or adaptive fusion. CoDeV2V additionally models persistent link conditions coupling the delay and loss of separate transmissions carrying one neighbor update. Under persistent communication degradation, the receiver may receive only part of a current update while retaining earlier information. Before temporal compensation, the receiver must determine which features or detections can be decoded by the deadline and when the corresponding observations were made. On CoDeV2V, the evaluated methods, including these three approaches, still lose detection accuracy relative to complete, current-time neighbor input (Table 1).

To address incomplete or unavailable current neighbor updates and stale historical information, we propose AssemblyFusion (Figure 1). AssemblyFusion uses a common feature assembly process for delayed and partially lost updates: current and historical regional features are aligned and combined, with feature age and spatial coverage guiding fusion with ego features. Each regional transmission also carries a complete copy of the neighbor's detections and motion estimates from the same observation. For each eligible neighbor, AssemblyFusion selects detections from the latest observation with a complete copy received by the deadline, motion-compensates older detections, and combines the selected detections with predictions from fused features. Regional assembly organizes available features across space and time; motion compensation updates historical object positions.

Detection selection draws on communication-system freshness management. Age of Information (AoI) measures elapsed time since the generation of the freshest received update (Yates et al., 2021); collective-perception standards group messages by a common reference time (ETSI, 2023). These ideas motivate organizing neighbor detections by observation time and selecting the most recent eligible detections among complete copies received by the deadline. Historical-state propagation under unreliable links motivates the subsequent motion compensation (Dutta & Singh, 2023).

Under severe coupled communication degradation on the 5,400-frame CoDeV2V test split, AssemblyFusion achieves 53.20 $\pm$ 0.46 BEV AP@0.7 across three communication realizations, 6.70 points above CoAlign. Mechanism analyses separate historical-detection availability from the benefit of motion compensation.

Our contributions are: (1) CoDeV2V, a newly collected long-sequence multi-vehicle LiDAR dataset with a formulation of persistent communication conditions coupling transmission delay and loss;

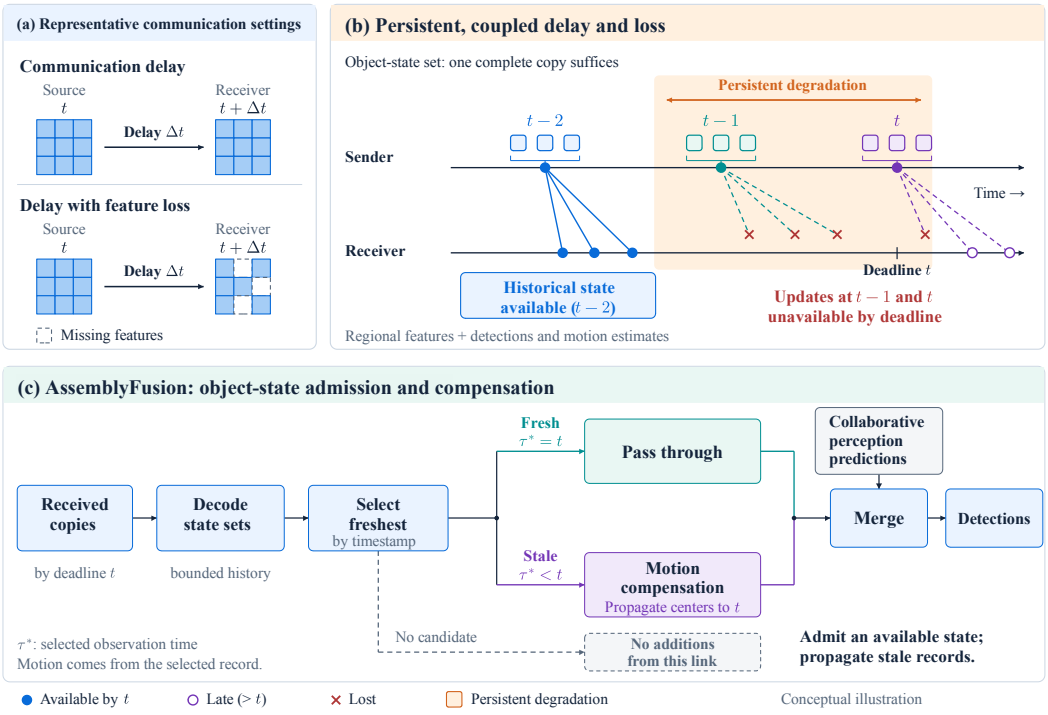


Figure 1: **Communication settings and AssemblyFusion.** (a) Message-level delay and delay with feature masking. (b) Persistent link conditions couple transmission delay and loss. Each square carries one region's features and a complete copy of the neighbor's detections and motion estimates from the same observation. Filled/open dots denote arrival by/after the deadline; crosses denote loss. (c) Per eligible neighbor, detections are selected by latest observation time among complete copies received by the deadline, motion-compensated if historical, and fused with feature-based predictions. Times, counts, and trajectories are schematic; t denotes send time in (a) and decision time in (b,c).

(2) AssemblyFusion, a regional feature assembly framework that addresses incomplete neighbor updates and stale historical information under coupled delay and loss by combining received features from current and earlier observations and incorporating detection selection and motion compensation; and (3) 5,400-frame comparisons with 14 external methods and paired receiver-strategy ablations separating historical-detection availability from the benefit of motion compensation.

2 Related Work

2.1 Collaborative Perception

Collaborative perception shares and fuses complementary observations among vehicles or roadside nodes. Early fusion combines raw observations, intermediate fusion aggregates learned features, and late fusion combines independently generated outputs. OPV2V evaluates these paradigms and provides an attention-based intermediate-fusion reference (Xu et al., 2022b). V2VNet exchanges compressed intermediate features through a spatially aware graph network for detection and prediction (Wang et al., 2020). DiscoNet distills a holistic-observation teacher into an intermediate-fusion student with spatially varying collaboration weights (Li et al., 2021). V2X-ViT alternates heterogeneous multi-agent and multiscale window attention (Xu et al., 2022a), while CoBEVT uses fused axial attention for cooperative BEV representations (Xu et al., 2023). Information selection and representation are complementary directions: Where2comm uses spatial confidence to communicate informative regions (Hu et al., 2022), while Point Cluster preserves object structure with compact point clusters and combines them through matching and merging (Ding et al., 2025).

2.2 COMMUNICATION FOUNDATIONS AND IMPAIRMENT-AWARE PERCEPTION

Communication metrics distinguish packet-delay variation and loss patterns from average delivery quality (Demichelis & Chimento, 2002; Koodli & Ravikanth, 2002). Measurements of reception blackouts motivate temporal dependence (Renda et al., 2016), while finite-state channel models provide a framework for representing persistent link conditions (Hassan et al., 2004). AoI and collective-perception reference timestamps motivate freshness management based on update timestamps (Yates et al., 2021; ETSI, 2023). These foundations guide the granularity, temporal structure, and timestamp organization of our formulation.

For asynchronous perception, SyncNet jointly estimates features and attention with time modulation (Lei et al., 2022); CoBEVFlow warps BEV features using estimated motion fields (Wei et al., 2023); TraF-Align samples and aggregates historical features along predicted trajectories (Song et al., 2025). Point Cluster instead estimates object motion from matched historical clusters (Ding et al., 2025). These approaches reduce temporal mismatch through prediction, alignment, or propagation.

For missing information, SCORPION combines BEV-feature reconstruction with spatial and temporal alignment under packet loss and pose/synchronization errors (Zhu et al., 2025). QPoint2Comm uses codebook indices and masked training for loss tolerance (Xu et al., 2026). RoCooper enhances corrupted features using ego and historical information, with a subsequent extension addressing robust and efficient cooperation (Tang et al., 2025; 2026). V2X-INCOP predicts missing collaborative information from previously fused features during communication interruptions (Ren et al., 2024). AFFormer addresses channel-induced feature corruption from path loss, fading, and noise (Zhou et al., 2026). AssemblyFusion jointly uses regional features received across observation times and a complete timestamped detection set from each eligible neighbor when available by the decision deadline, with motion compensation applied after set selection.

3 PROBLEM FORMULATION

3.1 COLLABORATIVE PERCEPTION UNDER COUPLED COMMUNICATION DEGRADATION

We index the ego vehicle and N_{c} collaborating vehicles by $i = 0, 1, \ldots, N_{\mathrm{c}}$, with $i = 0$ corresponding to the ego vehicle and $i = 1, \ldots, N_{\mathrm{c}}$ to the collaborating vehicles. Let $t \in \mathcal{T}$ be an integer observation index. The task is to estimate current targets Y_0^t from ego observation X_0^t and received neighboring information. Observations are separated by $\delta = 100$ ms; τ indexes a sender observation on the same axis. Communication governs access to neighboring information while current ego observations and evaluation targets remain fixed.

For method m, sender i represents the update from frame τ using $K_{i,\tau}^m$ transmission units $p_{i,\tau,k}^m$. Together, these units represent the sender's update for observation time τ; each retains the sender and observation-time identifiers. Their content and decoding rules determine the available features or object records. $\mathcal{N}_0^\tau$ and $\mathcal{N}_0^t$ denote neighbors eligible at the observation time and at the current time, respectively. Motivated by vehicular measurements (Wang et al., 2021; Renda et al., 2016), persistent link states jointly condition unit-level delay and loss, allowing consecutive update interruptions; sampling rules appear in Appendix B.1.

3.2 RECEIVER-AVAILABLE INFORMATION AND TASK OBJECTIVE

For communication realization ω, arrival time $a_{i,\tau,k}^m(\omega)$ is measured in observation periods; $+\infty$ denotes loss. The deadline d_t is fixed before communication outcomes. Both lost units and units arriving after d_t are unavailable for that decision. The received set is

$$\mathcal{R}_{0,m}^t(\omega) = \left\{ (p_{i,\tau,k}^m, a_{i,\tau,k}^m(\omega)) \;\middle|\; \tau \le t,\ i \in \mathcal{N}_0^\tau,\ 1 \le k \le K_{i,\tau}^m, \quad a_{i,\tau,k}^m(\omega) \le d_t \right\}. \tag{1}$$

Let $\mathcal{D}_{0,m}^t$ contain decoded perceptual content from neighbors eligible at both the observation time and the current decision. Permitted inputs also include a reception descriptor μ_m^t derived from received units, receiver context $\mathcal{C}_{\mathrm{rx}}^t$ such as the clock and poses, and history $\mathcal{L}_m^{t-}$ built from previously permitted inputs. The current prediction is

$$\widehat{Y}^t_{0,m} = f_m\big(X^t_0, \mathcal{D}^t_{0,m}, \mu^t_m, \mathcal{C}^t_{\mathrm{rx}}, \mathcal{L}^{t-}_m\big)\,. \tag{2}$$

The objective is accurate current-scene perception despite delayed or missing updates, using current ego observations and causally available receiver information.

4 AssemblyFusion

4.1 Overall Framework and Object-State Construction

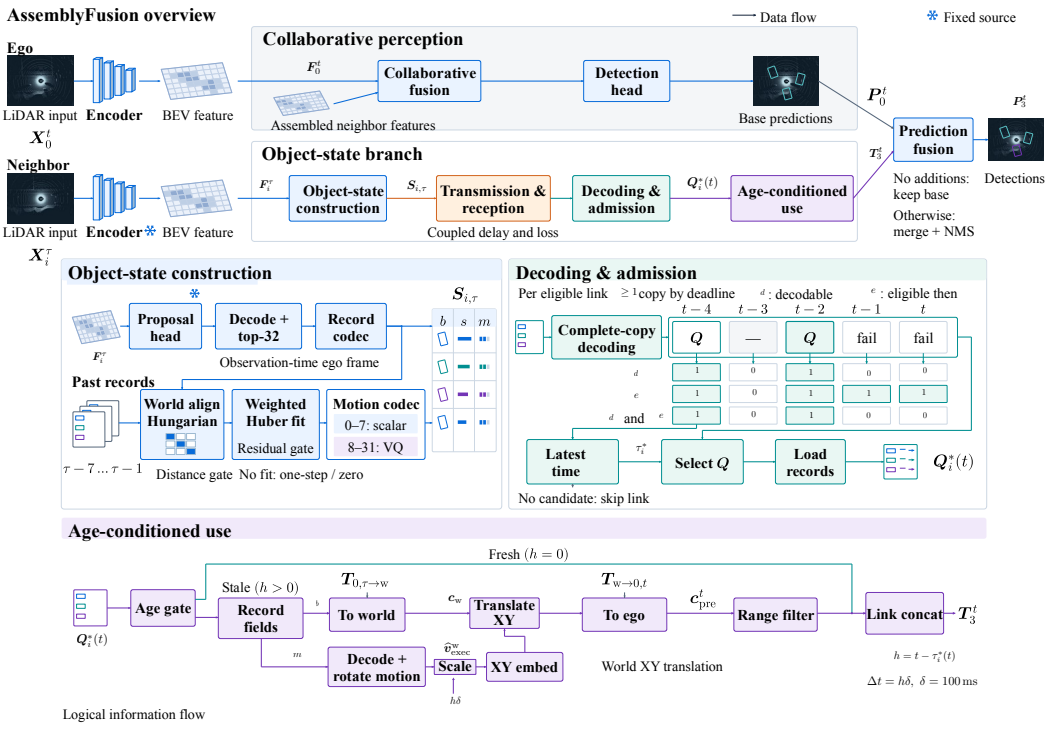


Figure 2: **AssemblyFusion overview.** The upper pipeline assembles received regional features for collaborative prediction. The lower panels show timestamped object-state construction, decoding and admission, and age-conditioned use. Each object-state set contains boxes, confidence scores, and encoded planar velocity estimates from one observation. F^t_0 and F^τ_i are ego and neighbor features; b, s, m denote geometry, confidence, and encoded motion; e indicates neighbor eligibility. VQ means vector quantization; snowflakes mark fixed sender components. LiDAR thumbnails are BEV projections of samples (Xu et al., 2022b).

Coupled delay and loss leave regional features with different availability and observation times. AssemblyFusion aligns and combines received current and historical regions in a common assembly process, using feature age and spatial coverage to guide fusion with ego features (Figure 2). Partial current updates and available history can therefore contribute together to current perception.

The sender encodes spatial point-cloud regions into early PointPillars representations (Lang et al., 2019). Each nonempty pillar carries pre-activation channel maxima, retained-point count, coordinate mean, and grid index, with observation timestamps. The receiver merges encodings from the same neighbor and observation with mean correction, aligns observation groups to the neighbor's current frame, and combines them by channelwise maximum. Aligned point-count weights and timestamps yield age and valid-region maps. Age adaptation and the BEV backbone process the assembled features before fusion with current ego features to produce P^t_0. Historical feature resampling approximates alignment across changing coordinate grids. At each decision deadline, delayed and lost transmissions both affect which regional features are available, allowing AssemblyFusion to handle their combined effects through the same assembly process.

In parallel, sender i groups detected boxes, confidence scores, and encoded planar velocity estimates at observation time τ into a timestamped object-state set $\mathcal{S}_{i,\tau}$. Postprocessed boxes use the destination ego's coordinate frame at τ and are ranked by confidence, retaining up to 32 proposals. Each regional transmission carries features from its own region together with the selected detections, confidence scores, and motion estimates from that observation. The regional features and accompanying detection information share the same arrival or loss outcome. Receiving a regional transmission makes the selected detections available, while regional feature coverage depends on which transmissions arrive.

Motion is estimated causally from proposal histories using world-coordinate association and confidence-weighted Huber fitting. A residual gate sets unreliable estimates to zero. When fitting is unavailable, AssemblyFusion uses an adjacent-frame estimate if available, and zero otherwise. The object-state set carries the encoded planar velocity estimates alongside object geometry and confidence.

4.2 Generation-Coherent, Decodability-First Admission

Inspired by timestamped status updates (Yates et al., 2021), AssemblyFusion makes decodability a prerequisite for freshness selection (Figure 3). Under the complete-copy representation, an object-state set is decodable if at least one qualifying copy is available by the deadline. Decoding returns a descriptor $Q^t_{i,\tau}$ identifying the corresponding object-state set for sender i and observation time τ, or failure $\perp_{\mathrm{dec}}$. Repeated copies correspond to the same object-state set.

For each currently eligible sender with at least one candidate, AssemblyFusion selects the latest observation time among fully decodable object-state sets within the current and four preceding observations:

$$\begin{aligned} d^t_{i,\tau} &= \mathbf{1}[Q^t_{i,\tau} \neq \perp_{\mathrm{dec}}], \qquad \mathcal{W}_t = \{t-4,\dots,t\} \cap \mathcal{T}, \\ \tau^\star_i(t) &= \max\{\tau \in \mathcal{W}_t :\ i \in \mathcal{N}^\tau_0,\ d^t_{i,\tau} = 1\}, \quad i \in \mathcal{N}^t_0, \\ Q^\star_i(t) &= \mathrm{Load}\big(Q^t_{i,\tau^\star_i(t)}\big). \end{aligned} \tag{3}$$

Here $d^t_{i,\tau}$ indicates decodability, and $Q^\star_i(t)$ contains the entries of the selected object-state set. Candidates are ordered by observation timestamp, and the selected state is loaded once. With no candidate, selection and loading are skipped. An empty selected object-state set also yields an empty addition on that link.

Generation coherence means that the geometry, confidence, and motion in each admitted object-state set remain associated with the same neighbor and observation time. Different links may contribute object-state sets of different ages. The perception path separately combines regions from different observation times. Caches indexed by sender and observation time realize the object-state admission flow in offline replay; Appendix C.2 gives the receiver algorithm.

4.3 Stale-State Compensation and Prediction Fusion

Current-observation states are used directly. For an admitted historical object state, let $\tau = \tau^\star_i(t)$ and $\Delta t = (t-\tau)\delta$. Following historical-state propagation in remote estimation (Dutta & Singh, 2023), AssemblyFusion transforms its box center in homogeneous coordinates c_τ to world coordinates, applies the decoded planar motion, and re-expresses it in the current ego frame:

$$\begin{aligned} c^t_{\mathrm{pre}} &= T_{w\to 0,t}\left[T_{0,\tau\to w}\, c_\tau + \Delta t\, \mathsf{E}_{xy}(\widehat{v}^w_{\mathrm{exec}})\right], \\ \mathsf{E}_{xy}([u_x,u_y]) &= [u_x,u_y,0,0]^\top. \end{aligned} \tag{4}$$

The transforms map between the destination ego and world frames at the specified times. The decoded world-frame velocity $\widehat{v}^w_{\mathrm{exec}}$ translates the center while preserving world-frame height and orientation, box dimensions, and confidence; coordinate transformation then re-expresses the box. Historical states undergo task-range filtering, while fresh states already use the current ego frame. The retained records $\mathcal{T}^t_3$ are concatenated with P^t_0 and fused by rotated non-maximum suppression (NMS); if $\mathcal{T}^t_3$ is empty, P^t_0 is returned unchanged. Appendix C.3 gives the exact fusion branches.

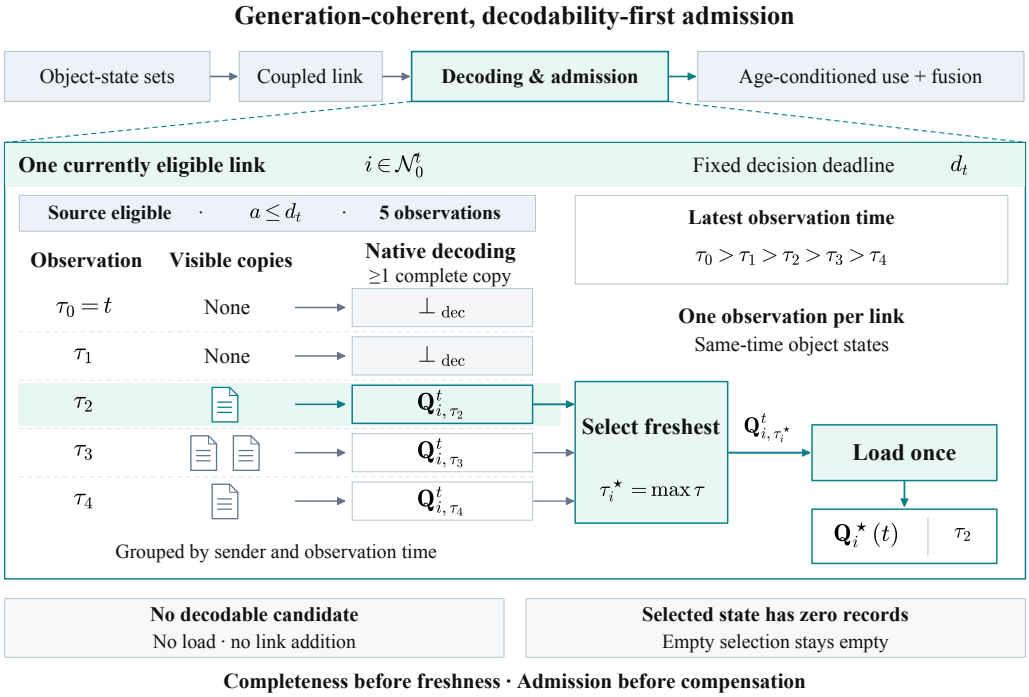


Figure 3: **Generation-coherent, decodability-first admission.** Complete copies are grouped by neighbor and observation time within a five-observation window. For a currently eligible neighbor, the latest fully decodable object-state set is selected; the schematic example selects τ_2. Current sets are used directly, while historical sets undergo motion compensation. The receiver algorithm is given in Appendix C.2.

5 Experiments

5.1 CoDeV2V: Continuous Sequences with Coupled Communication Degradation

CoDeV2V combines newly collected continuous multi-vehicle LiDAR observations with simulated coupled delay and loss. The communication model draws on joint loss–delay hidden Markov models evaluated on measured network traffic (Iannello et al., 2005; Salvo Rossi et al., 2006). Each V2V link has a communication condition that can persist across updates and affects both transmission delay and loss probability; outages can block consecutive neighbor updates. Long sequences retain observations before and after interruptions, supporting analysis of the availability and freshness of historical information. Methods are compared using the same driving observations, evaluation targets, and communication-degradation conditions.

AssemblyFusion training and validation use preprocessed OPV2V observations (Xu et al., 2022b); evaluation uses six CoDeV2V test sequences. Each test sequence contains 4–7 acquisition vehicles equipped with 64-channel LiDARs; point clouds, poses, and timestamps are saved at 10 Hz. Each sequence spans 900 consecutive instants, approximately 90 s, yielding 5,400 ego evaluation frames and 74,502 object–frame annotations in total, including frames without target objects. Appendix A gives acquisition settings, data partitions, and dataset statistics; Appendices B and C specify the evaluation protocol and receiver algorithms, respectively.

5.2 Experimental Setup

Evaluation scope. All AssemblyFusion results in this section and Appendix D use the sender-side regional-feature transport and receiver assembly described in Section 4.1.

We report BEV AP $\times 100$ at IoU thresholds 0.3, 0.5, and 0.7 using VOC2010 integration. With weights and frames fixed, means and sample standard deviations over three communication realizations quantify communication variability. AP-point differences use unrounded means. The severe coupled-degradation simulation uses a 100 ms grid, 12 logical opportunities per link and observation update, and the current decision slot as cutoff (Appendix B).

The 14 external systems share communication events while retaining their native representations, payload sizes, detectors, temporal designs, and training provenance. Matched Ego evaluates each detector with ego input; Native Ideal-current supplies its complete current-time collaborative input. No Fusion is an independent author-released OpenCOOD single-agent PointPillars (Xu et al., 2022b). All conditions share evaluation frames, region of interest (ROI), targets, and metrics. Ablations share perception weights and a training-data-fitted motion codebook. Implementation adaptations appear in Appendix C.1.

5.3 Comparison with Existing Methods

Table 1: Complete-split BEV AP@0.7 under severe simulated coupled delay and loss. Severe entries are mean $\pm$ sample SD across three communication realizations. Ego and Ideal-current are method-specific; No Fusion is independent.

Method	Matched Ego	Native Ideal-current	Severe
No Fusion	31.69	—	—
Late Fusion	—	—	44.12 ± 0.44
Attentive Fusion	35.35	64.08	35.23 ± 0.52
CoAlign	40.47	57.21	46.49 ± 0.32
CoBEVT	37.64	66.81	29.18 ± 0.64
CoDynTrust	36.94	60.90	30.69 ± 0.24
CoSDH	41.67	66.26	33.99 ± 0.59
DiscoNet	37.15	54.85	33.85 ± 0.44
F-Cooper	33.23	59.26	20.60 ± 0.34
How2comm	34.04	44.29	37.21 ± 0.20
LRCP	28.33	37.56	31.98 ± 0.08
SCOPE	41.79	58.76	40.09 ± 0.26
SyncNet	34.61	40.70	19.44 ± 0.05
V2VNet	35.05	62.02	25.44 ± 0.75
V2X-ViT	32.99	43.00	26.91 ± 0.41
Where2comm	33.21	62.62	31.25 ± 0.50
AssemblyFusion (ours)	33.43	61.09	$\mathbf{53.20 \pm 0.46}$

AssemblyFusion achieves 53.20 ± 0.46 BEV AP@0.7, exceeding CoAlign and SCOPE by 6.70 and 13.11 points, respectively (Table 1): 19.77 points above matched Ego and 7.90 below Ideal-current. AP@0.5 and AP@0.3 reach 77.69 ± 0.57 and 83.83 ± 0.53 (Table S2). Per-scene rankings vary.

In the evaluated implementations, F-Cooper (Chen et al., 2019) and Attentive Fusion (Xu et al., 2022b) use spatial fusion without explicit temporal compensation. Their AP@0.7 drops from 59.26 to 20.60 and from 64.08 to 35.23, respectively, from Ideal-current to Severe. These declines motivate jointly handling regional availability and observation age.

Figure 4 shows two cases; the remaining two appear in Appendix D.3.

5.4 Ablation Studies and Historical-State Availability

Admitting and using history. Fresh-only admission improves AP@0.7 from 40.26 to 43.64; full AssemblyFusion further improves it by 9.56 ± 0.29 points (Table 2). Static historical reuse retains ego-pose alignment but omits object displacement in Equation (4). Relative to fresh-only admission, it gains 4.97 points at IoU 0.3 yet loses 6.39 at IoU 0.7, consistent with stale geometry weakening precise localization. Motion compensation improves over static reuse by 15.94 ± 0.17 points. Full AssemblyFusion improves over no object-state addition by 12.93 ± 0.41 points.

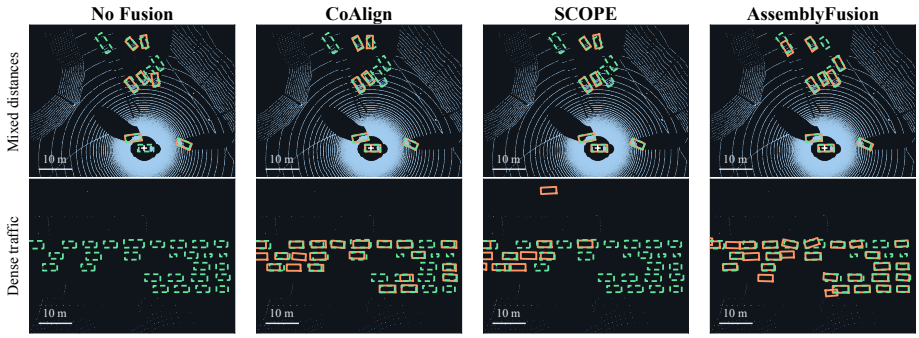


Figure 4: Matched BEV detections in mixed-distance and dense-traffic examples (seed 4612): cyan ego LiDAR, dashed green ground truth, orange predictions.

Table 2: Paired receiver strategies under severe coupled delay and loss (BEV AP $\times 100$; mean $\pm$ sample SD across three realizations). Perception weights are shared; the regional-feature path can retain historical information under fresh-only admission.

Configuration	AP@0.3	AP@0.5	AP@0.7
No object-state addition	70.72 ± 0.35	64.18 ± 0.33	40.26 ± 0.15
Fresh-only admission	73.94 ± 0.48	67.69 ± 0.51	43.64 ± 0.39
Static historical reuse	78.91 ± 0.60	64.73 ± 0.70	37.26 ± 0.29
AssemblyFusion (full)	$\mathbf{83.83 \pm 0.53}$	$\mathbf{77.69 \pm 0.57}$	$\mathbf{53.20 \pm 0.46}$

Availability and use of history. When no complete copy of the current object-state set has arrived by the deadline, history supplies an admissible set in 20,796 of 25,849 eligible decisions (80.45%). Outputs coincide without stale admissions; gains occur in the stale-present group. Appendix D details admission statistics and results conditioned on historical availability.

6 Discussion and Limitations

Persistent link states couple delay and loss across updates. Given the shared state and base delay, unit losses and delay perturbations are sampled independently in non-outage states; outages erase all units of an update. The 100 ms grid describes availability at perception cycles; finer-grained scheduling and measured wireless traces are further validation directions.

7 Conclusion

CoDeV2V combines newly collected long-sequence multi-vehicle LiDAR data with simulated delay and loss coupled by persistent link conditions. AssemblyFusion brings delayed and missing regional updates into a common feature assembly framework, combining received features from current and earlier observations to support perception when current neighbor information is incomplete or unavailable. The framework also selects each eligible neighbor's detections from the latest available observation and motion-compensates historical detections before fusion. Paired ablations distinguish historical-detection availability from the gains due to motion compensation.

AI USE STATEMENT

Generative AI tools assisted literature retrieval and synthesis; formulation of the communication model and receiver-information framework; discussion of mathematical relations and experimental design; implementation and debugging of method, data-processing, evaluation, and analysis code; interpretation of recorded results; manuscript drafting, editing, and translation; and schematic figure design and document preparation. The driving observations were collected in a simulator, and the quantitative figures use recorded evaluation outputs; generated visual examples in method diagrams are schematic. Verification included checking citations against primary sources, comparing mathematical and algorithmic descriptions with implementation, and checking numerical and figure provenance. The authors retain responsibility for the scientific claims, software, data, figures, and final manuscript, including AI-assisted material.

REPRODUCIBILITY STATEMENT

Appendix A describes acquisition and annotation rules; Appendix B specifies communication sampling, timing, and metrics; Appendix C records receiver inference and comparison details. Appendix D reports complementary metrics, historical-information use, and qualitative comparisons. Section 6 discusses limitations. Section 5.2 identifies the evaluated implementation and communication realizations.

ETHICS STATEMENT

The driving observations considered here are simulated. The study evaluates perception under modeled communication impairments. Practical use would require validation of sensing, communication, and system-level safety beyond the scenarios examined.

REFERENCES

Qi Chen, Xu Ma, Sihai Tang, Jingda Guo, Qing Yang, and Song Fu. F-Cooper: Feature Based Cooperative Perception for Autonomous Vehicle Edge Computing System using 3D Point Clouds. In *ACM/IEEE Symposium on Edge Computing (SEC)*, pp. 88–100, 2019. doi: 10.1145/3318216.3363300. URL https://doi.org/10.1145/3318216.3363300.

C. Demichelis and P. Chimento. IP Packet Delay Variation Metric for IP Performance Metrics (IPPM). Technical Report RFC 3393, RFC Editor, 2002. URL https://www.rfc-editor.org/info/rfc3393/.

Zihan Ding, Jiahui Fu, Si Liu, Hongyu Li, Siheng Chen, Hongsheng Li, Shifeng Zhang, and Xu Zhou. Point Cluster: A Compact Message Unit for Communication-Efficient Collaborative Perception. In *International Conference on Learning Representations (ICLR)*, pp. 2696–2712, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/07b446bc8bbadb47793711e4cda2f43e-Abstract-Conference.html.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An Open Urban Driving Simulator. In *Conference on Robot Learning (CoRL)*, volume 78, pp. 1–16. PMLR, 2017. URL https://proceedings.mlr.press/v78/dosovitskiy17a.html.

Manali Dutta and Rahul Singh. Optimal Scheduling Policies for Remote Estimation of Autoregressive Markov Processes over Time-Correlated Fading Channel. In *IEEE Conference on Decision and Control (CDC)*, pp. 6455–6462, 2023. doi: 10.1109/CDC49753.2023.10384144. URL https://doi.org/10.1109/CDC49753.2023.10384144.

ETSI. 5G; Service Requirements for Enhanced V2X Scenarios. Technical Specification ETSI TS 122 186 V16.2.0 / 3GPP TS 22.186 Release 16, European Telecommunications Standards Institute, 2020. URL https://www.etsi.org/deliver/etsi_ts/122100_122199/122186/16.02.00_60/ts_122186v160200p.pdf.

ETSI. Intelligent Transport System (ITS); Vehicular Communications; Basic Set of Applications; Collective Perception Service; Release 2. Technical Specification ETSI TS 103 324 V2.1.1, European Telecommunications Standards Institute, 2023. URL https://www.etsi.org/deliver/etsi_ts/103300_103399/103324/02.01.01_60/ts_103324v020101p.pdf.

M. Hassan, M. M. Krunz, and I. Matta. Markov-Based Channel Characterization for Tractable Performance Analysis in Wireless Packet Networks. *IEEE Transactions on Wireless Communications*, 3(3):821–831, 2004. doi: 10.1109/TWC.2004.827729. URL https://doi.org/10.1109/TWC.2004.827729.

Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-Efficient Collaborative Perception via Spatial Confidence Maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 4874–4886, 2022. doi: 10.52202/068431-0352. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/1f5c5cd01b864d53cc5fa0a3472e152e-Abstract-Conference.html.

Giulio Iannello, Francesco Palmieri, Antonio Pescapè, and Pierluigi Salvo Rossi. End-to-End Packet-Channel Bayesian Model Applied to Heterogeneous Wireless Networks. In *IEEE Global Telecommunications Conference (GLOBECOM)*, volume 1, pp. 484–489. IEEE, 2005. doi: 10.1109/GLOCOM.2005.1577673. URL https://doi.org/10.1109/GLOCOM.2005.1577673.

R. Koodli and R. Ravikanth. One-Way Loss Pattern Sample Metrics. Technical Report RFC 3357, RFC Editor, 2002. URL https://www.rfc-editor.org/info/rfc3357/.

Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. PointPillars: Fast Encoders for Object Detection from Point Clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12697–12705, 2019. doi: 10.1109/CVPR.2019.01298. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Lang_PointPillars_Fast_Encoders_for_Object_Detection_From_Point_Clouds_CVPR_2019_paper.html.

Zixing Lei, Shunli Ren, Yue Hu, Wenjun Zhang, and Siheng Chen. Latency-Aware Collaborative Perception. In *European Conference on Computer Vision (ECCV)*, volume 13692 of *Lecture Notes in Computer Science*, pp. 316–332. Springer, 2022. doi: 10.1007/978-3-031-19824-3_19. URL https://doi.org/10.1007/978-3-031-19824-3_19.

Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, and Wenjun Zhang. Learning Distilled Collaboration Graph for Multi-Agent Perception. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 29541–29552, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/f702defbc67edb455949f46babab0c18-Abstract.html.

Yifan Lu, Quanhao Li, Baoan Liu, Mehrdad Dianati, Chen Feng, Siheng Chen, and Yanfeng Wang. Robust Collaborative 3D Object Detection in Presence of Pose Errors. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4812–4818, 2023. doi: 10.1109/ICRA48891.2023.10160546. URL https://doi.org/10.1109/ICRA48891.2023.10160546.

Shunli Ren, Zixing Lei, Zi Wang, Mehrdad Dianati, Yafei Wang, Siheng Chen, and Wenjun Zhang. Interruption-Aware Cooperative Perception for V2X Communication-Aided Autonomous Driving. *IEEE Transactions on Intelligent Vehicles*, 9(4):4698–4714, 2024. doi: 10.1109/TIV.2024.3371974. URL https://doi.org/10.1109/TIV.2024.3371974.

M. Elena Renda, Giovanni Resta, Paolo Santi, Francesca Martelli, and Alessandro Franchini. IEEE 802.11p VANets: Experimental Evaluation of Packet Inter-Reception Time. *Computer Communications*, 75:26–38, 2016. doi: 10.1016/j.comcom.2015.06.003. URL https://www.sciencedirect.com/science/article/abs/pii/S0140366415002182.

Pierluigi Salvo Rossi, Gianmarco Romano, Francesco Palmieri, and Giulio Iannello. Joint End-to-End Loss–Delay Hidden Markov Model for Periodic UDP Traffic Over the Internet. *IEEE*

Transactions on Signal Processing, 54(2):530–541, 2006. doi: 10.1109/TSP.2005.861066. URL https://doi.org/10.1109/TSP.2005.861066.

Zhiying Song, Lei Yang, Fuxi Wen, and Jun Li. TraF-Align: Trajectory-Aware Feature Alignment for Asynchronous Multi-Agent Perception. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12048–12057, 2025. doi: 10.1109/CVPR52734.2025.01125. URL https://openaccess.thecvf.com/content/CVPR2025/html/Song_TraF-Align_Trajectory-aware_Feature_Alignment_for_Asynchronous_Multi-agent_Perception_CVPR_2025_paper.html.

Tao Tang, Chaokun Zhang, Gong Chen, and Jinlong E. RoCooper: Robust Cooperative Perception under Vehicle-to-Vehicle Communication Impairments. In *IEEE Conference on Computer Communications (INFOCOM)*, pp. 1–10, 2025. doi: 10.1109/INFOCOM55648.2025.11044682. URL https://doi.org/10.1109/INFOCOM55648.2025.11044682.

Tao Tang, Chaokun Zhang, Gong Chen, Jinlong E, and Pengcheng Lyu. Robust and Efficient Cooperative Perception under Vehicle-to-Vehicle Communication Impairments. *IEEE Transactions on Mobile Computing*, 25(8):12058–12073, 2026. doi: 10.1109/TMC.2026.3668902. URL https://doi.org/10.1109/TMC.2026.3668902.

Junjie Wang and Tomas Nordström. Latency Robust Cooperative Perception using Asynchronous Feature Fusion. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4862–4871, 2025. doi: 10.1109/WACV61041.2025.00476. URL https://openaccess.thecvf.com/content/WACV2025/html/Wang_Latency_Robust_Cooperative_Perception_using_Asynchronous_Feature_Fusion_WACV_2025_paper.html.

Runmin Wang, Xinrui Zhang, Zhigang Xu, Xiangmo Zhao, and Xiaochi Li. Research on Performance and Function Testing of V2X in a Closed Test Field. *Journal of Advanced Transportation*, 2021:9970978, 2021. doi: 10.1155/2021/9970978. URL https://doi.org/10.1155/2021/9970978.

Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. V2VNet: Vehicle-to-Vehicle Communication for Joint Perception and Prediction. In *European Conference on Computer Vision (ECCV)*, volume 12347 of *Lecture Notes in Computer Science*, pp. 605–621. Springer, 2020. doi: 10.1007/978-3-030-58536-5_36. URL https://doi.org/10.1007/978-3-030-58536-5_36.

Sizhe Wei, Yuxi Wei, Yue Hu, Yifan Lu, Yiqi Zhong, Siheng Chen, and Ya Zhang. Asynchrony-Robust Collaborative Perception via Bird's Eye View Flow. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 28462–28477, 2023. doi: 10.52202/075280-1236. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/5a829e299ebc1c1615ddb09e98fb6ce8-Abstract-Conference.html.

Junhao Xu, Yanan Zhang, Zhi Cai, and Di Huang. CoSDH: Communication-Efficient Collaborative Perception via Supply-Demand Awareness and Intermediate-Late Hybridization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6834–6843, 2025a. doi: 10.1109/CVPR52734.2025.00641. URL https://openaccess.thecvf.com/content/CVPR2025/html/Xu_CoSDH_Communication-Efficient_Collaborative_Perception_via_Supply-Demand_Awareness_and_Intermediate-Late_Hybridization_CVPR_2025_paper.html.

Runsheng Xu, Yi Guo, Xu Han, Xin Xia, Hao Xiang, and Jiaqi Ma. OpenCDA: An Open Cooperative Driving Automation Framework Integrated with Co-Simulation. In *IEEE International Intelligent Transportation Systems Conference (ITSC)*, pp. 1155–1162, 2021. doi: 10.1109/ITSC48978.2021.9564825. URL https://doi.org/10.1109/ITSC48978.2021.9564825.

Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer. In *European Conference on Computer Vision (ECCV)*, volume 13699 of *Lecture Notes in Computer Science*, pp. 107–124. Springer, 2022a. doi: 10.1007/978-3-031-19842-7_7. URL https://doi.org/10.1007/978-3-031-19842-7_7.

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 *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2583–2589, 2022b. doi: 10.1109/ICRA46639.2022.9812038. URL https://doi.org/10.1109/ICRA46639.2022.9812038.

Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. CoBEVT: Cooperative Bird's Eye View Semantic Segmentation with Sparse Transformers. In *Conference on Robot Learning (CoRL)*, volume 205, pp. 989–1000. PMLR, 2023. URL https://proceedings.mlr.press/v205/xu23a.html.

Sheng Xu, Enshu Wang, Hongfei Xue, Jian Teng, Bingyi Liu, Yi Zhu, Pu Wang, Libing Wu, and Chunming Qiao. Send Less, Perceive More: Masked Quantized Point Cloud Communication for Loss-Tolerant Collaborative Perception. arXiv preprint arXiv:2602.21667, 2026. URL https://arxiv.org/abs/2602.21667.

Yunjiang Xu, Lingzhi Li, Jin Wang, Benyuan Yang, Zhiwen Wu, Xinhong Chen, and Jianping Wang. CoDynTrust: Robust Asynchronous Collaborative Perception via Dynamic Feature Trust Modulus. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 336–342, 2025b. doi: 10.1109/ICRA55743.2025.11127779. URL https://doi.org/10.1109/ICRA55743.2025.11127779.

Dingkang Yang, Kun Yang, Yuzheng Wang, Jing Liu, Zhi Xu, Rongbin Yin, Peng Zhai, and Lihua Zhang. How2comm: Communication-Efficient and Collaboration-Pragmatic Multi-Agent Perception. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 25151–25164, 2023a. doi: 10.52202/075280-1093. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4f31327e046913c7238d5b671f5d820e-Abstract.html.

Kun Yang, Dingkang Yang, Jingyu Zhang, Mingcheng Li, Yang Liu, Jing Liu, Hanqi Wang, Peng Sun, and Liang Song. Spatio-Temporal Domain Awareness for Multi-Agent Collaborative Perception. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23383–23392, 2023b. doi: 10.1109/ICCV51070.2023.02137. URL https://openaccess.thecvf.com/content/ICCV2023/html/Yang_Spatio-Temporal_Domain_Awareness_for_Multi-Agent_Collaborative_Perception_ICCV_2023_paper.html.

Roy D. Yates, Yin Sun, D. Richard Brown, III, Sanjit K. Kaul, Eytan Modiano, and Sennur Ulukus. Age of Information: An Introduction and Survey. *IEEE Journal on Selected Areas in Communications*, 39(5):1183–1210, 2021. doi: 10.1109/JSAC.2021.3065072. URL https://doi.org/10.1109/JSAC.2021.3065072.

Haibao Yu, Wenxian Yang, Hongzhi Ruan, Zhenwei Yang, Yingjuan Tang, Xu Gao, Xin Hao, Yifeng Shi, Yifeng Pan, Ning Sun, Juan Song, Jirui Yuan, Ping Luo, and Zaiqing Nie. V2X-Seq: A Large-Scale Sequential Dataset for Vehicle-Infrastructure Cooperative Perception and Forecasting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5486–5495, 2023. doi: 10.1109/CVPR52729.2023.00531. URL https://openaccess.thecvf.com/content/CVPR2023/html/Yu_V2X-Seq_A_Large-Scale_Sequential_Dataset_for_Vehicle-Infrastructure_Cooperative_Perception_and_CVPR_2023_paper.html.

Haibao Yu, Wenxian Yang, Jiaru Zhong, Zhenwei Yang, Siqi Fan, Ping Luo, and Zaiqing Nie. End-to-End Autonomous Driving through V2X Cooperation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(9):9598–9606, 2025. doi: 10.1609/aaai.v39i9.33040. URL https://ojs.aaai.org/index.php/AAAI/article/view/33040.

Xi Zhou, Tao Huang, Qing-Long Han, Rana Abbas, and Mostafa Rahimi Azghadi. AFFormer: Adaptive Feature Fusion Transformer for V2X Cooperative Perception under Channel Impairments. arXiv preprint arXiv:2605.01888, 2026. URL https://arxiv.org/abs/2605.01888v1.

Ruiyang Zhu, Minkyoung Cho, Shuqing Zeng, Fan Bai, and Z. Morley Mao. SCORPION: Robust Spatial-Temporal Collaborative Perception Model on Lossy Wireless Network. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 14194–14201, 2025. doi: 10.1109/IROS60139.2025.11247050. URL https://doi.org/10.1109/IROS60139.2025.11247050.

A Dataset and Annotations

A.1 Acquisition and annotation rules

CoDeV2V contains 24 newly acquired sequences and 21,600 synchronized instants, organized into 12/6/6 acquisition train/validation/test sequences. Evaluation uses the six test sequences, with one fixed ego and 900 consecutive instants per sequence: 5,400 ego frames and 27,900 synchronized vehicle-level point-cloud/label pairs.

Collection uses CARLA 0.9.11 (Dosovitskiy et al., 2017) and OpenCDA (Xu et al., 2021). Each vehicle has a 64-channel LiDAR with a 120 m range, 1.3 million points/s, and a vertical field of view of $-25°$ to $+2°$. Simulation and scanning run at 20 Hz, with observations saved at 10 Hz. The six evaluation sequences span approximately 90 s each and include merging, crossing traffic, curves, stop-and-go motion, and occlusion. LiDAR general dropoff, intensity limit, zero-intensity dropoff, and noise standard deviation are 0.1, 0.7, 0.15, and 0.02, respectively.

Ground truth uses simulator geometry and semantic-LiDAR instance hits for target eligibility. Labels from acquisition vehicles within 160 m, capped at seven vehicles, are deduplicated by object identity and transformed into the current ego LiDAR frame. Valid boxes have finite geometry, positive dimensions, and all eight corners inside the inclusive ROI $[-140.8, 140.8] \times [-38.4, 38.4] \times [-3, 1]$ m. Communication conditions share the same current ego observations and targets. All 5,400 frames are evaluated, including 145 empty-target frames whose false positives are counted.

A.2 Perception difficulty under ideal communication

To understand perception difficulty under Ideal-current, Table S1 compares CoDeV2V with the original OPV2V Default Towns test split using the same target projection, deduplication, and ROI rules. Ideal-current supplies each method with its complete current-time collaborative input, removing the simulated delay and loss. Detection still depends on the available observations and the model's ability to handle the scene's objects. The checked LiDAR configurations match; the comparison therefore focuses on observed target distributions. Instances are counted per target and frame, and evaluation retains all eligible targets.

Table S1: Dataset characteristics under a common target definition. OPV2V refers to the original Default Towns test split. Instances are counted per target and frame; vehicle counts describe acquisition participants.

Attribute	OPV2V Default Towns test	CoDeV2V eval.
Ego instants	2,170	5,400
Duration per scene	≈7–22 s	≈90 s
Acquisition vehicles	2–5	4–7
Target–frame instances	32,498	74,502
Empty-target frames	0	145
Targets per ego frame	14.98	13.80
Median target distance	32.15 m	49.61 m
Distance >60 m	17.23%	43.15%
Distance >80 m	7.47%	29.04%
Horizontal length >5.5 m	0.00%	18.43%

CoDeV2V has a larger proportion of targets far from the ego vehicle and a different size distribution: median target distance increases from 32.15 to 49.61 m, the proportion beyond 80 m from 7.47% to 29.04%, and the proportion with horizontal length above 5.5 m from 0.00% to 18.43%. These shifts provide context for lower Ideal-current AP: complete communication preserves available views, while detection remains subject to visibility, observation quality, and generalization to the target distribution. Observation quality also depends on each collaborating vehicle's viewpoint. A separate LiDAR-support diagnostic samples 150 detector-training frames containing 2,000 targets and 180 evaluation frames containing 2,470 targets. The fraction with ego-point support decreases from 94.10% to 75.91%, while the neighbor-only fraction increases from 5.75% to 24.05%. This sampled

training–evaluation comparison describes observation support; Table S1 instead compares the two complete test splits. Together, these observations motivate reporting Matched Ego, Ideal-current, and degraded-communication results on the same targets. These statistics characterize the observed distribution differences.

B Communication and Evaluation Protocol

B.1 Coupled delay and loss

A transmission unit $p^m_{i,\tau,k} = (\xi^m_{i,\tau,k}, e^m_{i,\tau,k})$ contains its method-native payload and co-arriving metadata identifying the sender, observation time, unit, cardinality, and representation. Source construction uses information available by τ. Communication events depend on link conditions independently of the perception method, payload content, predictions, and targets. A persistent Markov link state jointly conditions loss and delay, with a shared base delay within each link update.

Each eligible observation advances the link state. Transition entries below give the next-state probabilities in good/bad/outage order, conditional on the preceding state and current link distance.

Link distance	From good	From bad	From outage
Up to 30 m	(0.88, 0.10, 0.02)	(0.45, 0.45, 0.10)	(0.15, 0.45, 0.40)
Above 30, up to 55 m	(0.72, 0.24, 0.04)	(0.28, 0.60, 0.12)	(0.05, 0.35, 0.60)
Above 55, up to 160 m	(0.50, 0.38, 0.12)	(0.16, 0.62, 0.22)	(0.02, 0.28, 0.70)

State	Base-delay sampling list (steps)	Unit loss probability
Good	(0, 0, 1, 2, 3, 4)	0.02
Bad	(0, 2, 3, 4, 5, 6, 7, 8)	0.12
Outage	All opportunities erased	1

The preceding state starts as good; a source outside 160 m provides no opportunities and sets the preceding state to outage. Each eligible link update has 12 logical opportunities sharing the state and, outside outage, one base delay sampled uniformly from the listed entries, including the repeated zero. Conditional on state and base delay, each unit samples loss and a delay perturbation independently. With probability 0.20, the perturbation is -1 or +1 step with equal probability; delays are clipped to [0, 8]. Outage removes all units of that update. A surviving unit arrives at its observation slot plus the sampled delay. These parameters specify the simulation used for evaluation.

B.2 Timing and metrics

The simulation uses a 100 ms grid aligned with saved observations, and the decision cutoff is the current slot ($d_t = t$). Zero-step delay means same-slot availability. Units associated with an observation share its transmission timestamp; the model resolves perception-cycle availability and leaves within-cycle scheduling abstract. Lost units have arrival time $+\infty$; finite arrivals after the cutoff become available only for later decisions.

Evaluation uses BEV/XY polygon IoU and VOC2010 AP. For each of three communication realizations, predictions are pooled over the declared evaluation set before AP computation; we report mean and sample standard deviation. Differences between methods are paired within each realization and computed before rounding. Weights remain fixed across these realizations. Seeds, deterministic tie ordering, and evaluator identifiers are supplied in the reproducibility material.

C Algorithms and Implementation

C.1 Compared methods and controls

Table 1 evaluates Attentive Fusion (Xu et al., 2022b), CoAlign (Lu et al., 2023), CoBEVT (Xu et al., 2023), CoDynTrust (Xu et al., 2025b), CoSDH (Xu et al., 2025a), DiscoNet (Li et al., 2021),

F-Cooper (Chen et al., 2019), How2comm (Yang et al., 2023a), LRCP (Wang & Nordström, 2025), SCOPE (Yang et al., 2023b), SyncNet (Lei et al., 2022), V2VNet (Wang et al., 2020), V2X-ViT (Xu et al., 2022a), and Where2comm (Hu et al., 2022). Attentive Fusion is the OPV2V/OpenCOOD intermediate-fusion reference.

No Fusion uses the author-released OpenCOOD single-agent PointPillars detector; Matched Ego and Native Ideal-current retain each method's own detector. Evaluation shares frames, targets, ROI, metrics, and communication events while preserving native representations, neighborhoods, and histories, including How2comm's sequential inputs. Shared logical opportunities standardize availability conditions. Late Fusion combines ego detections with neighbor detections using the same detection selection and motion compensation as AssemblyFusion. Incorporating regional feature assembly and fusion further improves AP@0.7 from 44.12 to 53.20 under the same communication conditions. This improvement indicates that assembled regional features provide additional value beyond detection-level collaboration.

Payload composition and overhead. We estimate numerical payload sizes from 192 source updates, with 32 updates sampled from each sequence. Comparisons use the 188 updates shared by the four evaluated layouts, with scene weights derived from eligible source-update counts. Under these layouts, the estimated payloads per neighbor update are 1,854.48 KiB for AssemblyFusion, 15,401.39 KiB for CoAlign, 9,977.03 KiB for SCOPE, and 943.21 KiB for Where2comm. AssemblyFusion sends features from different spatial regions across 12 transmission opportunities. Each opportunity additionally carries detection and motion information in a logical quantized layout reserving 544 B, totaling 6.375 KiB per update. The estimates exclude common metadata and network-layer overhead; AssemblyFusion's accounting includes all scheduled opportunities, including transmissions subsequently lost or delayed.

Native feature representation sizes. We additionally inspect feature representations in the evaluated implementations. The dense FP32 representations occupy 15,400 KiB for Attentive Fusion and SyncNet, 35,200 KiB for F-Cooper, 33,792 KiB for CoBEVT, 8,448 KiB for V2X-ViT, and 52,800 KiB for DiscoNet per source observation. V2VNet uses 8,800 KiB per directed message-passing round, with two rounds configured. LRCP's original multiscale features occupy 15,400 KiB, while the projected current-observation features occupy 23,100 KiB. CoDynTrust's pre-compensation pillar features and backbone features occupy 35,200 and 30,800 KiB, respectively. These alternative processing-stage representations are reported separately rather than added together. CoSDH's dense compressed FP16 bottleneck occupies 462 KiB, excluding auxiliary messages. These component sizes provide implementation-specific context; complete transmitted totals additionally depend on message construction, auxiliary exchanges, and history caching. How2comm's complete total remains unresolved. Overall, AssemblyFusion achieves its reported detection performance with a moderate estimated payload within the range of the audited comparison layouts. The accompanying detection and motion information accounts for approximately 0.34% of AssemblyFusion's counted payload, adding only a small byte overhead to regional feature transmission.

Additional related methods. CoBEVFlow (Wei et al., 2023), TraF-Align (Song et al., 2025), and SCORPION (Zhu et al., 2025) are discussed in Section 2 but excluded from the numerical comparison. At the audited CoBEVFlow revision `ab8f9ff`, the inspected learned-flow path overwrites predicted flow with ground truth before feature warping. The audited TraF-Align implementation groups features into two agents. These paths require changes to prediction or feature-grouping logic to support our multi-neighbor evaluation without ground-truth receiver input, exceeding our input-only adaptation policy. SCORPION's official project page provided the paper and video but no model-code release at the time of our audit. These implementation constraints determined their exclusion from this comparison.

C.2 Receiver inference

Algorithm 1 summarizes feature assembly, detection selection, motion compensation, and prediction fusion. A separate fixed sender network generates detections and motion estimates from observation-time history. A completely received empty detection set remains eligible for latest-observation selection and contributes no boxes. External methods retain their native decoding and information-use rules.

Algorithm 1 AssemblyFusion receiver inference at observation index t

Require: Current ego observation X_0^t; units received by d_t; receiver history; available poses, timestamps, and eligible-neighbor sets $\mathcal{N}_0^\tau$.
Ensure: Current predictions P_3^t and updated regional cache.

```
Update the regional cache using arrivals visible by d_t.
Assemble neighbor features, age maps, and support maps (Section 4.1).
Encode current ego input; apply age adaptation, backbone, collaborative fusion,
    and detection head to obtain P_0^t.
T_3^t ← ∅; W_t ← {t − 4, ..., t} ∩ T.
for each i ∈ N_0^t do
    K_i ← ∅.
    for each τ ∈ W_t with i ∈ N_0^τ do
        if at least one complete copy of S_{i,τ} is received by d_t then
            Decode Q_{i,τ}^t; add τ to K_i.
        end if
    end for
    if K_i ≠ ∅ then
        τ* ← max K_i; load the object states in Q_{i,τ*}^t once.
        if τ* = t then
            Append the fresh boxes and scores to T_3^t.
        else
            Transform boxes to world coordinates; decode their carried motion.
            Propagate centers by (t − τ*)δ and transform to the current ego frame.
            Apply task-range filtering; append retained boxes and scores to T_3^t.
        end if
    end if
end for
return Prediction fusion of P_0^t and T_3^t in Equation (C.1).
```

An empty decoded or filtered set contributes no boxes; selection stays unchanged.

C.3 Prediction fusion

Let $\mathcal{T}_3^t$ contain retained fresh and motion-compensated historical objects and P_0^t the collaborative prediction. The output uses ordered concatenation $\|$ and rotated NMS. If the object-state branch contributes no retained objects, the collaborative prediction is returned directly:

$$P_3^t = \begin{cases} P_0^t, & \mathcal{T}_3^t = \varnothing, \\ \mathrm{NMS}(P_0^t \| \mathcal{T}_3^t), & \mathcal{T}_3^t \neq \varnothing,\ P_0^t \text{ present}, \\ \mathrm{NMS}(\mathcal{T}_3^t), & \mathcal{T}_3^t \neq \varnothing,\ P_0^t \text{ absent}. \end{cases} \tag{C.1}$$

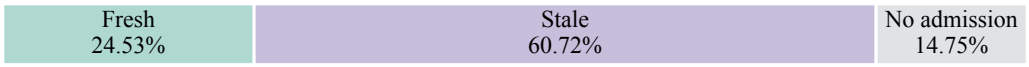
D SUPPLEMENTARY EXPERIMENTS

D.1 ADDITIONAL IOU THRESHOLDS

Table S2: BEV AP@0.3 and AP@0.5 on the complete CoDeV2V evaluation split under severe coupled communication degradation. Means and sample standard deviations summarize three communication realizations; No Fusion is deterministic.

Method	AP@0.3	AP@0.5
Attentive Fusion	67.74 ± 0.63	57.61 ± 0.62
CoAlign	73.20 ± 0.38	67.76 ± 0.39
CoBEVT	59.65 ± 0.79	48.10 ± 0.98
CoDynTrust	57.60 ± 0.48	51.02 ± 0.46
CoSDH	71.00 ± 0.89	58.15 ± 0.98
DiscoNet	60.36 ± 0.83	53.51 ± 0.70
F-Cooper	59.58 ± 0.77	46.18 ± 0.73
How2comm	58.93 ± 0.24	54.84 ± 0.24
LRCP	73.89 ± 0.21	67.82 ± 0.12
SCOPE	70.05 ± 0.46	62.80 ± 0.41
SyncNet	30.31 ± 0.14	27.86 ± 0.19
V2VNet	68.43 ± 0.63	53.58 ± 0.80
V2X-ViT	53.25 ± 0.76	45.79 ± 0.76
Where2comm	65.51 ± 0.80	53.95 ± 0.92
AssemblyFusion (ours)	83.83 ± 0.53	77.69 ± 0.57
No Fusion	45.48	43.33

D.2 AVAILABILITY AND USE OF HISTORICAL STATES



Current update undecodable: history admits a state in 80.45% of decisions

Figure S1: Object-state availability over 34,251 eligible link decisions. The historical-availability rate of 80.45% is conditional on the 25,849 decisions with an undecodable current-observation set. Admission counts include decoded empty sets.

Fresh, historical, and absent admissions account for 8,402 (24.53%), 20,796 (60.72%), and 5,053 (14.75%) eligible link decisions, respectively (Figure S1). Current-update unavailability forms 5,567 consecutive runs, with a longest run of 60 observations. Available historical information therefore bridges a subset of update interruptions.

AssemblyFusion assembles the regional features and neighbor detections available under coupled delay and loss into predictions for the current scene. Regional feature assembly aligns and combines received features across observation times, while prediction fusion incorporates historical neighbor detections after motion compensation. The assembly framework thus allows partially received current updates and available history to support current perception together.

Table S3 examines how the assembly framework benefits from historical detections. With regional-feature processing held fixed, incorporating motion-compensated historical detections improves BEV AP@0.7 by 11.62 ± 0.31 points over fresh-only admission on the 11,069 frame–realization instances containing historical admissions. Predictions coincide on the remaining 5,131 instances. Groups are defined using communication and admission metadata before range filtering and NMS;

AP is computed within each group and communication realization, followed by paired differences. Age groups contain different frames and targets, so their results describe where the gains occur.

Table S3: AssemblyFusion gains in BEV AP@0.7 points over fresh-only admission under severe coupled communication degradation. Subgroups partition the historical-admission row; age is the interval from the observation time to the current decision. Counts are frame–realization instances.

Frame group	Instances	vs. fresh-only
Historical admission present	11,069	$+11.62 \pm 0.31$
Maximum age 100 ms	2,524	$+8.79 \pm 0.83$
Maximum age 200 ms	3,183	$+11.61 \pm 0.65$
Maximum age 300 ms	3,013	$+12.85 \pm 0.27$
Maximum age 400 ms	2,349	$+12.58 \pm 0.39$

D.3 REGIONAL-FEATURE QUALITATIVE COMPARISONS

The following figures use the same fixed regional-feature model as the complete-split numerical evaluation. Figure 4 shows mixed-distance and dense-traffic examples; Figure S2 adds sparse-return and nearby-vehicle examples from the same four-sequence comparison. External boxes reproduce the archived No Fusion, CoAlign, and SCOPE predictions on the same observations.

All methods share current ego LiDAR, ground truth, coordinates, and spatial windows, with every postprocessed prediction retained. Each local crop uses identical bounds, point-cloud colors, and box styles across methods. Successful examples are selected for differences between AssemblyFusion and fresh-only admission; a difficult example complements them.

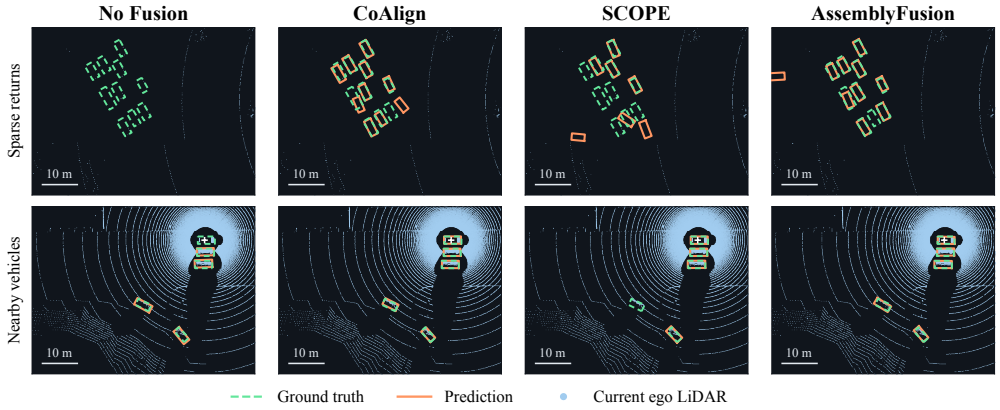


Figure S2: Additional qualitative comparisons with sparse ego LiDAR returns and nearby vehicles (communication seed 4612). Light-blue points show current ego LiDAR, green dashed boxes denote ground truth, and orange solid boxes denote predictions; white scale bars indicate 10 m. All methods use identical local views, and collaborative methods share one communication realization.

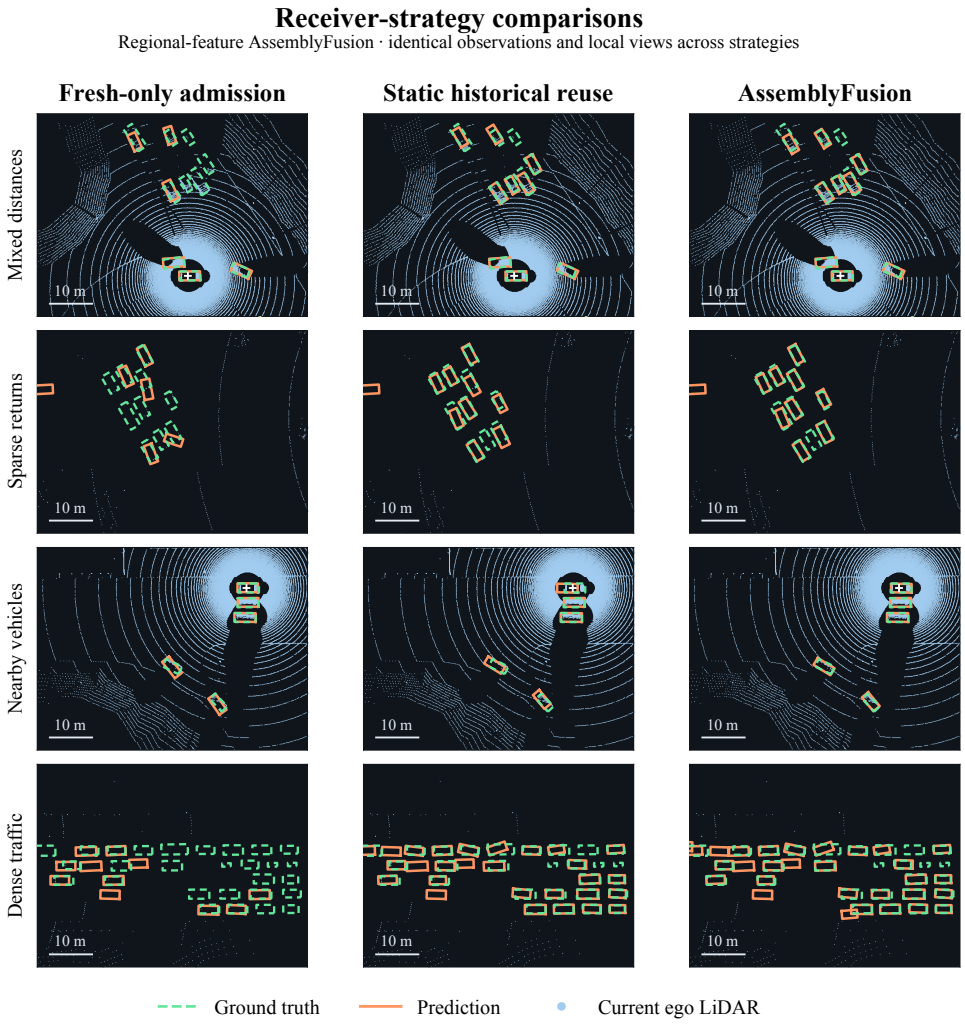


Figure S3: Receiver-strategy comparisons on the same observations. Fresh-only admission, static historical reuse, and AssemblyFusion share the same underlying collaborative predictions P_0^t. The examples distinguish admitting historical information from correcting its geometry for the current decision.

D.4 A DIFFICULT EXAMPLE

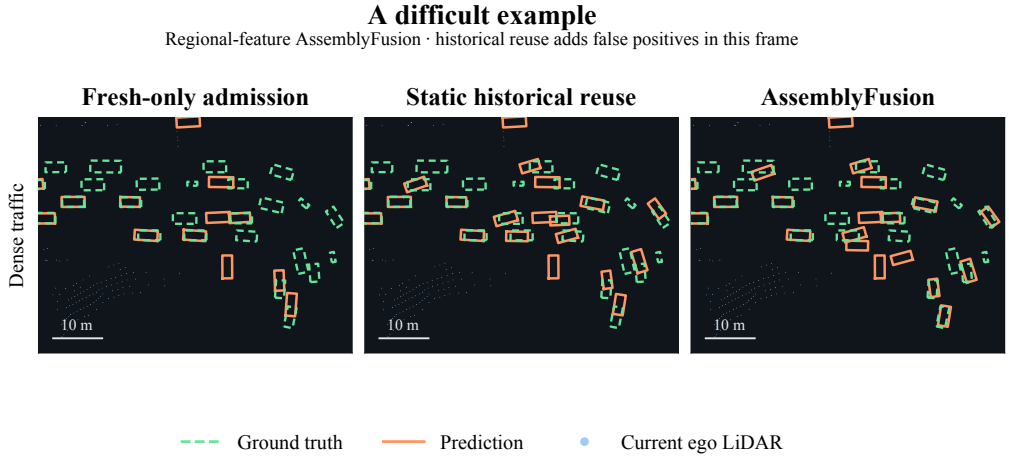


Figure S4: A difficult dense-traffic observation. The additional historical detections illustrate a case where admission and motion compensation fail to improve current perception. All postprocessed predictions are retained.