FFBL-COOP: ASSOCIATION-DECOUPLED COOPERATIVE 3D MULTI-OBJECT TRACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Cooperative 3D tracking must integrate complementary observations across agents and time while maintaining consistent identities. When evidence integration and identity inheritance share a matching decision, errors arising from cross-view appearance differences and spatial misalignment can compromise both feature fusion and track continuity. We propose FFBL-Coop, a *fuse first, bind later* framework that separates instance admission from identity management. Confidence-ranked Slot Admission (CSA) allocates cooperative queries to available ego slots using confidence and spatial proximity. Unified Representation Aggregation (URA) uses cooperative semantic features and aligned anchors to guide ego-feature retrieval, refining the augmented query bank within a shared transformer decoder. After refinement, Cooperative-Priority Identity Anchoring (CPIA) combines learned association with persistent mappings to establish accepted identity assignments across frames. A shared codebook reduces transmitted payload while retaining AP and AMOTA close to the uncompressed variant. FFBL-Coop achieves AMOTA/AP of $0.611/0.548$ on V2X-Seq and $0.688/0.653$ on Griffin-25M. Code will be released.

1 INTRODUCTION

Cooperative 3D tracking extends ego perception through observations from nearby vehicles, roadside infrastructure, or aerial platforms. These complementary views reveal objects obscured or outside the ego field of view. Exploiting this coverage over time requires integrating incoming evidence while preserving identities across changes in viewpoint and visibility.

Intermediate fusion supports evidence integration by exchanging learned representations, such as bird's-eye-view (BEV) features (Xu et al., 2022; 2023). To reduce communication cost, recent systems exchange sparse object queries and combine cross-agent association with feature aggregation (Yu et al., 2025; Zhong et al., 2025; Wang et al., 2026b). Within this sparse formulation, SparseCoop and Long-SCOPE accommodate unmatched queries, and SparseCoop further refines representations using ego image features (Wang et al., 2026b;c). Incorporating these representations into a recurrent tracker also requires deciding which trajectories the updated hypotheses continue.

Cooperative integration and temporal tracking impose different requirements on association. Specifically, integration must admit useful cooperative evidence despite cross-view appearance differences and alignment errors, whereas tracking must preserve identity across frames. Yet when one cross-agent match controls both query placement and identity inheritance, a correspondence error can compromise both representation updates and trajectory continuity (Figure 1(a)). Crucially, uncertainty about which existing track an observation corresponds to does not necessarily eliminate its value for detection. This distinction motivates our principle: *integrate and refine cooperative evidence first, then bind the resulting hypotheses to persistent trajectories*.

Following this principle, FFBL-Coop separates evidence admission from ego-track identity assignment through a *fuse first, bind later* pipeline. CSA selects cooperative queries by confidence and spatial proximity and places them in available ego slots without assigning ego IDs. URA then uses their semantic features and aligned anchors to guide ego-feature retrieval and refine the augmented bank. Only after final decoding does CPIA resolve identities using refined ego states and persistent cross-agent mappings, independently of admission weights. Learned association links queries to existing tracks, accepted assignments and new-track identities are recorded in persistent ID mappings.

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.

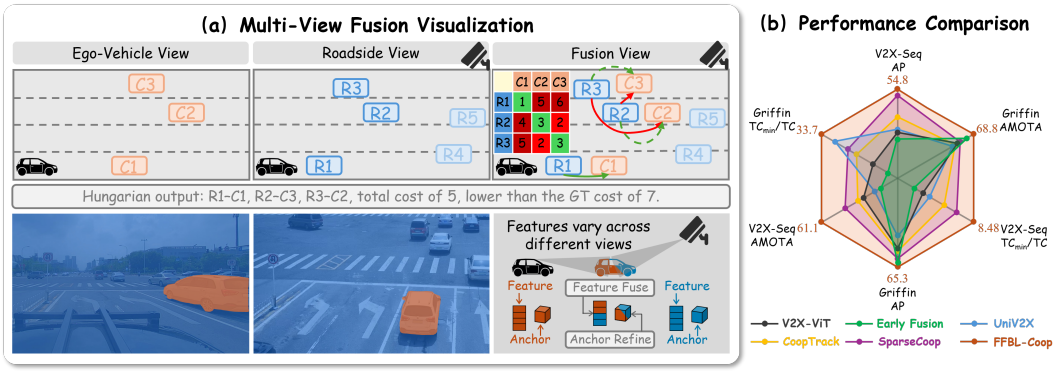


Figure 1: **(a) Cross-view matching challenges.** *Top:* Geometric discrepancies can make the minimum-cost assignment inconsistent with ground-truth object correspondences. *Bottom:* The same object has different feature representations and anchor estimates across agent views. **(b) System comparison.** Detection, tracking, and communication performance on V2X-Seq and Griffin-25M. Annotations show AP/AMOTA in percent and transmission rates in kB/s.

Figure 2 summarizes the complete pipeline, from cross-agent alignment and cooperative admission to shared decoding and identity binding. It shows how cooperative queries enter the recurrent ego bank and are refined alongside native ego queries. Valid out-of-region queries follow a separate box-output path to retain coverage beyond local integration, while a shared codebook compresses the cooperative features used for refinement. Experiments on V2X-Seq and Griffin-25M evaluate detection, tracking, and communication performance (Figure 1(b)). Controlled ablations examine whether separating identity management from feature admission improves tracking and how persistent ID mappings contribute to continuity.

Our main contributions are:

- We formulate cooperative query tracking as a *fuse first, bind later* process: cooperative evidence is admitted and refined before identity assignment. This separates the decision to integrate an observation from the decision of which trajectory it continues.
- We instantiate this principle through three stages. CSA ranks queries by confidence and proximity for admission, URA uses cooperative semantics to guide ego-feature retrieval, and CPIA combines association with persistent mappings to maintain identities across slot changes.
- Experiments on V2X-Seq and Griffin-25M demonstrate the effectiveness of FFBL-COOP for cooperative 3D detection and tracking. Codebook compression reduces payload while retaining AP and AMOTA close to uncompressed transmission.

2 RELATED WORK

2.1 3D MULTI-OBJECT TRACKING

Tracking-by-detection methods maintain trajectories by associating per-frame detections using geometric and motion cues (Weng et al., 2020; Yin et al., 2021). A complementary line of work learns object representations through queries. Specifically, query-based 3D detectors such as DETR3D (Wang et al., 2022) and PETR (Liu et al., 2022) aggregate multi-view image features into sparse object queries. Extending this representation across time, StreamPETR (Wang et al., 2023) and Sparse4D (Lin et al., 2023) incorporate historical information into object-centric representations, providing a foundation for recurrent updates.

Beyond temporal feature aggregation, query propagation also supports identity maintenance. MOTR (Zeng et al., 2022) propagates track queries across video frames, while DQTrack (Li et al., 2023) separates object and track queries to address representation conflicts between detection and tracking in 3D. ADA-Track (Ding et al., 2024) alternates detection and association within decoder layers. In a cooperative setting, recurrent updates additionally incorporate queries maintained by another agent. FFBL-Coop separates cooperative-query admission from ego-track identity inheritance using persistent cross-agent ID mappings to support continuity across slot changes.

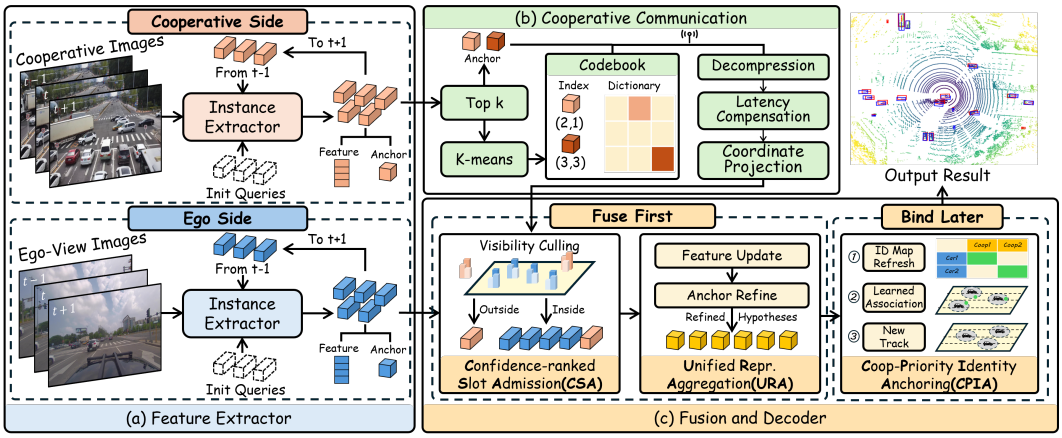


Figure 2: **FFBL-Coop framework.** Cooperative queries aligned to the ego frame through motion compensation and coordinate transformation. Following *fuse first, bind later*, CSA admits cooperative evidence into available slots , and URA uses cooperative semantics to guide ego-feature retrieval and refinement. CPIA assigns identities through learned association and ID mappings.

2.2 Cooperative Perception

Cooperative perception extends an agent's coverage by incorporating observations from other viewpoints. Within this paradigm, intermediate fusion exchanges learned features to integrate complementary evidence. F-Cooper (Chen et al., 2019) studies feature-level fusion, while V2X-ViT (Xu et al., 2022) and CoBEVT (Xu et al., 2023) use attention to aggregate multi-agent representations. To make communication more selective, When2com (Liu et al., 2020) and Where2comm (Hu et al., 2022) use learned communication decisions and spatial confidence maps, respectively. Complementing these efficiency-oriented designs, CoAlign (Lu et al., 2023) addresses pose uncertainty through agent and object alignment. At the object level, UniV2X (Yu et al., 2025) exchanges sparse queries for cooperative perception and planning.

Building on sparse representations, several methods combine cross-agent association, ture refinement, and temporal propagation (Zhong et al., 2025; Wang et al., 2026b;c). CoopTrack uses graph-based affinities and Hungarian matching to aggregate paired queries. SparseCoop combines geometric and appearance cues, refining queries with ego-image context. Long-SCOPE uses contextual descriptors and optimal-transport matching for long-range association. Despite advances in matching and refinement, coupling identity inheritance to query allocation can let matching errors affect both evidence integration and track continuity. FFBL-Coop addresses this coupling through a *fuse first, bind later* pipeline that separates feature admission from identity inheritance and records accepted assignments as persistent cross-agent ID mappings.

3 Method

3.1 Overview

FFBL-Coop follows a *fuse first, bind later* pipeline: cooperative evidence is integrated into the ego query bank before the hypotheses receive track identities (Figure 2). Each query is represented by (f, a, s, id), where $f \in \mathbb{R}^C$ is an instance feature, a is a 3D anchor encoding the object state, s is a confidence score, and id is the track identifier maintained by the originating agent.

At each communication step, the cooperative agent provides a query set

$$Q_t^{\mathrm{coop}} = \left\{ (f_j^{\mathrm{coop}}, a_j^{\mathrm{coop}}, s_j^{\mathrm{coop}}, \mathrm{id}_j^{\mathrm{coop}}) \right\}_{j=1}^{N_{\mathrm{coop}}}. \tag{1}$$

The feature component is transmitted through a shared codebook. The ego agent reconstructs the features and aligns the cooperative states to its current coordinate frame and timestamp (Sec. 3.3).

Geometrically valid aligned queries whose centers project within an ego camera image form the cooperative candidate pool (Sec. G). CSA allocates selected candidates to available slots, URA

refines their features with ego observations, and CPIA resolves identities. Valid queries outside all ego camera images bypass decoding and contribute boxes through a separate output path. The final output combines bank predictions and these cooperative detections:

$$Y_t = \left\{(\hat{a}_i, \hat{s}_i, \widehat{\mathrm{id}}_i)\right\}_{i=1}^{N_t^{\mathrm{out}}}, \tag{2}$$

where N_t^{out} is the number of output hypotheses, distinct from the fixed bank capacity N. The integration and identity rules are described in Sec. 3.4.

3.2 Agent-Specific Perception Stacks

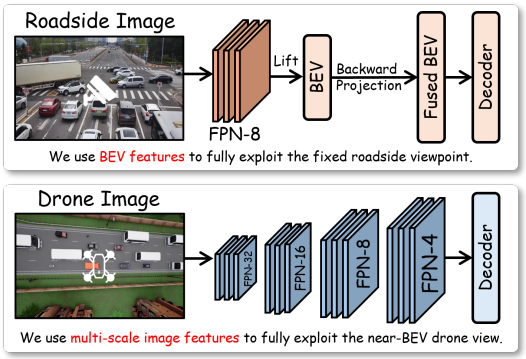


Figure 3: **Agent-specific perception stacks.** (a) Roadside queries aggregate evidence from constructed BEV features. (b) Aerial queries retrieve evidence from multi-scale FPN features.

The perception frontends use specific spatial representations and expose the same query interface (f, a, s, id) for cooperation. Roadside units construct BEV features, while ego vehicles and aerial agents sample image features through projected 3D anchors. Coordinate alignment and feature adaptation are applied before cross-agent integration.

Feature Encoding. Each agent processes its images with a backbone and feature pyramid network (FPN), producing multi-scale features F_t^I. For ego vehicles and aerial agents, queries project their 3D anchors into the camera views and aggregate features around the projected locations. This readout uses camera geometry without requiring the image plane to approximate a BEV layout.

For roadside units, LSS (Philion & Fidler, 2020) and deformable refinement construct BEV features F_t^{BP} from multi-scale image features. An image-conditioned hypernetwork predicts spatially varying modulation parameters (γ, β) and a gate g, yielding the final representation:

$$F_t^{\mathrm{BEV}} = (1 - g) \odot F_t^{\mathrm{BP}} + g \odot \left(\gamma \odot F_t^{\mathrm{BP}} + \beta\right), \tag{3}$$

where $\odot$ denotes element-wise multiplication, with broadcasting over feature dimensions where required. The gate controls the contribution of the image-conditioned modulation.

Sparse Queries and Recurrent Decoding. Each agent maintains a recurrent bank B_{t-1} of N query slots. A decoder layer updates the current queries, followed by cross-attention to the previous bank, producing B_t^{temp}. Subsequent decoder layers update query features and anchors:

$$\left(f_i^{\ell+1}, a_i^{\ell+1}\right) = \mathcal{D}_\ell\left(f_i^\ell, a_i^\ell, F_t, B_t^{\mathrm{temp}}\right), \tag{4}$$

where ℓ indexes decoder layers and F_t denotes the spatial features supplied by the frontend. This recurrent object-centric processing follows the temporal-query design of existing sparse perception models (Lin et al., 2023; Wang et al., 2023). In the ego pipeline, admission and identity updates occur after the temporal update. Subsequent layers jointly refine cooperative and native ego queries.

3.3 Communication and Cross-Agent Alignment

Codebook Compression. CodeFilling (Hu et al., 2024) explores codebook-based communication for collaborative perception. Here, we apply vector quantization (Linde et al., 1980) to sparse cooperative instance features. We construct a shared codebook $\mathcal{C} \in \mathbb{R}^{K_{\mathrm{code}} \times C}$ by applying k-means to training-set features, where K_{code} denotes the codebook size. Each cooperative feature is encoded by its nearest codeword:

$$z_j = \arg \min_{k \in \{1, \ldots, K_{\mathrm{code}}\}} \left\| f_j^{\mathrm{coop}} - \mathcal{C}_k \right\|_2^2. \tag{5}$$

The transmitted query contains $(z_j, a_j^{\text{coop}}, s_j^{\text{coop}}, \text{id}_j^{\text{coop}})$. The ego agent reconstructs the feature by table lookup, $\hat{f}_j^{\text{coop}} = \mathcal{C}_{z_j}$, while retaining the object state, confidence, and source track ID.

For queries integrated into the ego bank, reconstructed features provide object-level semantic cues for initialization and aggregation. Subsequent URA refinement incorporates local image evidence from ego spatial features. This use of cooperative features motivates codebook compression.

Latency and Coordinate Alignment. Before integration, the ego agent compensates for timestamp differences and transforms cooperative object states into its coordinate frame. Let $\Delta t = t^{\text{ego}} - t^{\text{coop}}$. Under a constant-velocity approximation, the cooperative center is updated as

$$p_j^{+} = p_j^{\text{coop}} + v_j^{\text{coop}} \Delta t, \tag{6}$$

where p_j^{coop} and v_j^{coop} are expressed in the cooperative frame. Applying $T_{\text{coop}\to\text{ego}}$ to the compensated state yields the aligned anchor $\tilde{a}_j^{\text{coop}}$, with its center and orientation transformed and velocity rotated into the ego frame.

A learned linear adapter conditions reconstructed features on the relative agent rotation:

$$\tilde{f}_j^{\text{coop}} = A_{\text{feat}} \left(\left[\hat{f}_j^{\text{coop}}; \text{vec}\big(R_{\text{coop}\to\text{ego}}\big) \right] \right), \tag{7}$$

where A_{feat} maps the concatenated input to $\mathbb{R}^C$ and $\text{vec}(\cdot)$ flattens the rotation matrix.

3.4 Association-Decoupled Cooperative Integration

FFBL-Coop separates evidence admission from identity commitment. After the first decoder block and temporal update, CSA places cooperative evidence in available slots without assigning track IDs. URA refines the augmented bank through the remaining layers. Only after final decoding does CPIA bind admitted hypotheses to identities using refined ego states and persistent correspondences, independently of CSA allocation weights.

Confidence-ranked Slot Admission (CSA). CSA selects cooperative evidence and determines its placement without assigning ego-track IDs. In-region queries $\widetilde{Q}_t^{\text{in}}$ are ranked by cooperative confidence, with budget K limiting the selected set $\mathcal{J}_t$. Let $\mathcal{H}_t$ and $\mathcal{F}_t^0$ index live tracks and free slots in B_t^{temp}. To restrict admission to free slots away from live-track centers, eligible slots satisfy

$$\mathcal{F}_t = \left\{ i \in \mathcal{F}_t^0 : \min_{h \in \mathcal{H}_t} \|p_i - p_h\|_{\text{BEV}} \geq \tau_{\text{CSA}} \right\}, \tag{8}$$

where p_i and p_h are centers at the bank-update stage. All free slots are eligible when $\mathcal{H}_t$ is empty. Admission is skipped if the candidate or eligible-slot set is empty.

During training, each selected query distributes its contribution across eligible slots according to distance:

$$\pi_{ji} = \frac{\exp(-d_{ji}/T_{\text{CSA}})}{\sum_{i' \in \mathcal{F}_t} \exp(-d_{ji'}/T_{\text{CSA}})}, \qquad d_{ji} = \|\tilde{p}_j^{\text{coop}} - p_i\|_{\text{BEV}}. \tag{9}$$

Both datasets normalize weights over slots without an occupancy dustbin. For slots with incoming mass $\rho_i > \epsilon_\rho$, contributions are aggregated as

$$\rho_i = \sum_{j \in \mathcal{J}_t} \pi_{ji}, \qquad \bar{f}_i^{\text{coop}} = \frac{\sum_{j \in \mathcal{J}_t} \pi_{ji} \tilde{f}_j^{\text{coop}}}{\rho_i}, \qquad \bar{p}_i^{\text{coop}} = \frac{\sum_{j \in \mathcal{J}_t} \pi_{ji} \tilde{p}_j^{\text{coop}}}{\rho_i}. \tag{10}$$

Soft initialization adds the aggregated feature and averages the slot and cooperative anchor centers:

$$f_i \leftarrow f_i + \bar{f}_i^{\text{coop}}, \qquad a_i^{xyz} \leftarrow \tfrac{1}{2} a_i^{xyz} + \tfrac{1}{2} \bar{p}_i^{\text{coop}}. \tag{11}$$

Other anchor components remain unchanged, and slots with $\rho_i \leq \epsilon_\rho$ retain their states.

At inference, candidates are processed in descending confidence order. Each overwrites the feature and complete anchor of its nearest unused eligible slot, until either set is exhausted. CSA retains each admitted query's source ID and assigned slot $i(j)$ for post-decoding identity binding, allocation itself does not determine the ego-track ID.

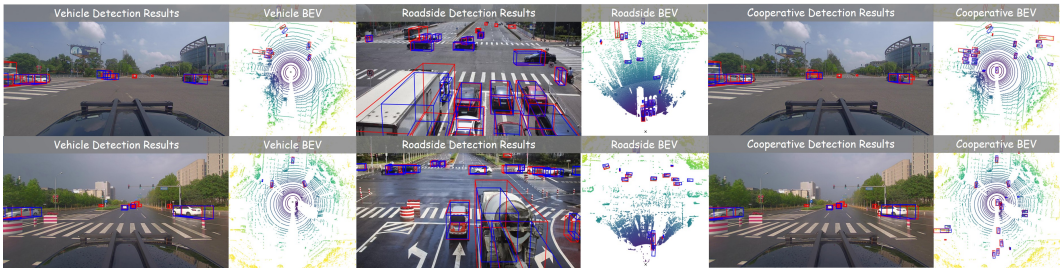


Figure 4: **Cooperative detection on V2X-Seq.** Each row shows vehicle-only, infrastructure-only, and cooperative results in paired image and BEV views. Blue: predictions, red: ground truth.

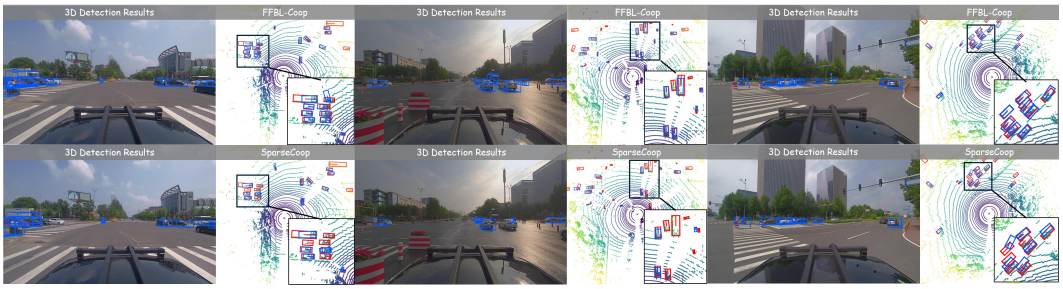


Figure 5: **Comparison with SparseCoop on V2X-Seq.** Top: FFBL-Coop, bottom: SparseCoop. Image and BEV pairs include enlarged regions highlighting detection differences.

Unified Representation Aggregation (URA). Cooperative and ego features may encode different visible parts of the same object. URA uses cooperative semantics to guide ego-feature retrieval: CSA-initialized queries enter the shared decoder, with anchors locating ego-image sampling regions and query states conditioning deformable aggregation:

$$\left(f_i^{\ell+1}, a_i^{\ell+1}\right) = \mathcal{D}_\ell\left(f_i^\ell, a_i^\ell, F_t, B_t^{\text{temp}}\right). \tag{12}$$

Here, F_t contains ego spatial features. This query-conditioned spatial readout gathers ego evidence beyond the information already pooled into instance vectors by direct linear or mean fusion. Admitted and native ego queries are refined in place through the remaining decoder layers, producing final states $(\hat{f}_i, \hat{a}_i)$ for CPIA identity association.

Let Ω_t denote slots initialized with cooperative instance representations. A complementary local readout supplies nearby cooperative context to queries outside Ω_t:

$$c_i = \text{Attn}\Big(f_i, \{\tilde{f}_j^{\text{coop}}\}_{j\in\mathcal{N}_{k_{\text{nbr}}}(i)}\Big), \qquad f_i \leftarrow f_i + g_i \odot c_i, \quad i \notin \Omega_t. \tag{13}$$

Here, $\mathcal{U}_t$ denotes cooperative query indices eligible for local readout, and $\mathcal{N}_{k_{\text{nbr}}}(i) \subseteq \mathcal{U}_t$ contains their nearest k_{nbr} neighbors within radius r. The gate is $g_i = \text{sigmoid}(\text{MLP}([f_i; c_i; d_i^{\min}]))$, where $d_i^{\min}$ is the nearest valid cooperative query distance. Empty neighborhoods skip the update. In-place decoding and local readout use the same forward operations during training and inference.

Cooperative-Priority Identity Anchoring (CPIA). CPIA resolves identities after final decoding, independently of CSA allocation weights. Association uses a fixed snapshot of final ego features and centers, binding changes identity ownership while preserving decoded features and boxes.

Learned association. An MLP predicts logits S_{jh} from cooperative-query features, final ego-track features $\hat{f}_h$, and their spatial relations. For a scoring set of m cooperative queries and n_{live} candidate ego tracks, we append an unmatched logit:

$$A = \text{softmax}_{\text{row}}\left([S/T_{\text{id}} \mid b_{\text{dust}}\mathbf{1}_m]\right), \tag{14}$$

where $T_{\text{id}} = 0.2$ and $b_{\text{dust}} = 0$. For $m > 1$, we initialize $v^{(0)} = \mathbf{1}_{n_{\text{live}}+1}$ and apply 20 scaling iterations:

$$u^{(q)} = \mathbf{1}_m \oslash (Av^{(q-1)}), \qquad v^{(q)} = \mathbf{1}_{n_{\text{live}}+1} \oslash (A^\top u^{(q)}), \tag{15}$$

where $\oslash$ denotes element-wise division. For $m = 1$, we retain A to avoid degenerate column scaling. For $m > 1$, unmatched queries compete for the unmatched column's unit total mass:

$$P = \begin{cases} A, & m = 1, \\ \mathrm{diag}(u^{(20)}) A \, \mathrm{diag}(v^{(20)}), & m > 1, \end{cases} \qquad \mathcal{L}_{\mathrm{id}} = -\sum_{j=1}^{m} \log P_{j,y_j}. \tag{16}$$

The target y_j indexes the corresponding live track, or the unmatched column $n_{\mathrm{live}} + 1$.

Persistent identity binding. At inference, CPIA labels admitted hypotheses in their assigned slots $i(j)$. Map M stores cooperative-to-ego correspondences from new-track initialization or accepted association. Valid mappings take priority if the mapped track is live and its ID is unclaimed by other admitted hypotheses. For unresolved queries and live tracks with unclaimed IDs, CPIA computes P, excluding the unmatched column from pair selection without using its score for rejection. Pairs are greedily accepted in descending score order, requiring unused rows and columns and $\|\tilde{p}_j^{\mathrm{coop}} - \hat{p}_h\|_{\mathrm{BEV}} \le \tau_{\mathrm{id}}$, where $\hat{p}_h$ is the final ego-track center. Association is skipped if either set is empty or $\tau_{\mathrm{id}} = 0$. Associated hypotheses inherit matched track IDs, those still without an identity receive new IDs. On ID transfer, the previous slot relinquishes the ID and stops output and propagation under it. Assignments update M, retaining one owner per ID in the next-frame bank.

3.5 Training Objectives

Initial cooperative training starts from pretrained single-agent checkpoints with randomly initialized cooperative modules. On both datasets, cooperative modules and all six decoder layers are trained, with the backbone and FPN frozen. The overall objective is

$$\mathcal{L} = \sum_{\ell=1}^{L} \left[\mathcal{L}_{\mathrm{cls}}^{(\ell)} + \mathcal{L}_{\mathrm{reg}}^{(\ell)} + \lambda_{\mathrm{dn}} \left(\mathcal{L}_{\mathrm{cls,dn}}^{(\ell)} + \mathcal{L}_{\mathrm{reg,dn}}^{(\ell)} \right) \right] + \lambda_{\mathrm{id}} \mathcal{L}_{\mathrm{id}}, \tag{17}$$

where classification, box regression, and denoising follow the single-agent recipe. The identity loss in Eq. 16 is applied once per frame with $\lambda_{\mathrm{id}} = 1.0$.

4 Experiments

We evaluate FFBL-Coop in vehicle-to-infrastructure (V2I) and vehicle-to-drone (V2D) cooperation. Experiments examine overall performance, module contributions, gains by camera-projection coverage, communication efficiency, and robustness to calibration errors and latency.

Datasets and Metrics. We use V2X-Seq-SPD (Yu et al., 2023) under its official 2 Hz validation protocol (Hao et al., 2025) and the Griffin-25M subset (Wang et al., 2026a) for V2D evaluation. Both benchmarks use nuScenes-style evaluation (Caesar et al., 2020). We report car-class AMOTA and AP, with AP averaged over center-distance thresholds $\{0.5, 1, 2, 4\}$ m. Ablations additionally report identity switches (IDS), and communication cost is measured in bytes per second (B/s). Evaluation and latency-injection protocols are detailed in Secs. G and F.1.

Implementation Details. FFBL-Coop uses MMDetection3D with a ResNet-50-FPN backbone and a six-layer decoder of hidden dimension $C = 256$. Training uses AdamW with learning rate 3×10^{-4}, weight decay 0.01, and batch size 8. Roadside features use a 0.4 m BEV grid, while Drone image features use FPN strides $\{4, 8, 16\}$. Main results use codebook-compressed features. Module ablations follow the initialization and training protocol specified in Sec. G. All six decoder layers are trained on both datasets, with the backbone and FPN frozen. Replacement rules are specified in Sec. 4.2. Baseline results in Table 1 are taken from Zhong et al. (2025); Wang et al. (2026b) under their original system configurations. No Fusion denotes the published non-cooperative baseline. Implementation details are provided in Sec. G.

4.1 Main Results

Table 1 compares complete systems in detection, tracking, and communication. Baseline results use their published architectures and training recipes. Sec. 4.2 assesses component contributions through retrained variants sharing perception initialization and a common training protocol.

Table 1: System-level comparison on V2X-Seq and Griffin-25M. AP and AMOTA are in percent, TC is transmission cost in B/s. FFBL-Coop uses $K_{\text{code}} = 256/512$, respectively. Baselines retain their published architectures and results. †Long-SCOPE reports the 0–50 m subset of its long-range evaluation and is included for reference. Bold marks the best results excluding this reference row.

Method	V2X-Seq			Griffin-25M		
	AP↑	AMOTA↑	TC↓	AP↑	AMOTA↑	TC↓
No Fusion	16.6	13.0	0	37.5	36.5	0
Late Fusion	19.6	26.3	6.60×10^2	37.8	37.7	1.56×10^3
Early Fusion	24.3	20.9	8.19×10^7	60.7	67.0	3.11×10^8
V2X-ViT (ECCV 22)	26.8	28.7	2.56×10^6	46.5	50.8	8.00×10^5
UniV2X (AAAI 25)	29.5	23.9	6.96×10^4	41.9	45.6	5.58×10^4
CoopTrack (ICCV 25)	39.0	32.8	5.64×10^4	47.9	48.8	1.17×10^5
SparseCoop (AAAI 26)	53.0	42.1	3.17×10^4	55.9	50.9	9.73×10^4
Long-SCOPE (CVPR 26)†	53.2	57.6	–	59.7	56.6	–
FFBL-Coop (Ours)	**54.8**	**61.1**	8.48×10^3	**65.3**	**68.8**	3.37×10^4

On V2X-Seq, FFBL-Coop achieves AP/AMOTA of $54.8/61.1$, exceeding SparseCoop's reported results by $1.8/19.0$ points. On Griffin-25M, it reaches $65.3/68.8$, with gains of $9.4/17.9$ points. These results demonstrate improved detection and aggregate tracking performance on both evaluated benchmarks. With codebook compression, FFBL-Coop has estimated payload rates of $8.48 \times 10^3/3.37 \times 10^4$ B/s on V2X-Seq/Griffin-25M, respectively, corresponding to approximately $26.8\%/34.6\%$ of SparseCoop's reported rates. On Griffin-25M, it exceeds Early Fusion's AMOTA of 67.0 with substantially less communication. Sec. 4.3 examines the effect of compression.

4.2 Ablation Study

All variants in Table 2 follow Sec. G. Cooperative modules and all six decoder layers are trained on both datasets, with the backbone and FPN frozen. Without CSA, allocation uses BEV-distance-based Hungarian matching under the same budget and gate. Without URA, mean feature aggregation replaces in-place refinement. Without CPIA, slot-ID inheritance replaces learned association and persistent mapping. Base combines these replacements while retaining cooperative input.

Table 2: Module ablations. Disabled CSA, URA, and CPIA use Hungarian allocation, mean aggregation, and slot-ID inheritance, respectively. URA-only uses one Hungarian assignment for admission and identity inheritance. Base retains cooperative input. Bold marks the best displayed results.

Base	CSA	URA	CPIA	V2X-Seq			Griffin-25M		
				AP↑	AMOTA↑	IDS↓	AP↑	AMOTA↑	IDS↓
✓				39.9	47.7	123	48.7	48.4	162
✓		✓		53.2	50.4	116	64.0	55.3	86
✓		✓	✓	54.2	60.1	**69**	64.1	66.1	57
✓	✓		✓	53.2	59.0	72	60.6	64.9	**48**
✓	✓	✓		54.3	59.3	78	62.9	67.3	124
✓	✓	✓	✓	**54.8**	**61.1**	70	**65.3**	**68.8**	50

Module Contributions and Identity Association. Full achieves the highest AP and AMOTA among the displayed configurations on both datasets, improving over Base by $14.9/16.6$ AP and $13.4/20.4$ AMOTA points on V2X-Seq/Griffin-25M. Removing URA causes the largest leave-one-out drops: $1.6/4.7$ AP and $2.1/3.9$ AMOTA points. Removing CPIA lowers AMOTA by $1.8/1.5$ points and increases IDS from $70/50$ to $78/124$. Under Hungarian allocation, adding CPIA to URA-only raises AMOTA from $50.4/55.3$ to $60.1/66.1$ and reduces IDS from $116/86$ to $69/57$, while AP changes from $53.2/64.0$ to $54.2/64.1$. With allocation and URA architecture fixed, this supports CPIA's combined correspondence establishment and persistent reuse beyond slot-ID inheritance. Replacing Hungarian allocation with CSA yields Full, adding $1.0/2.7$ AMOTA points. CPIA improves tracking under both allocation rules, although Full does not minimize IDS. Fixed-checkpoint interventions on V2X-Seq yield AMOTA of 60.5 for Map-only, 61.1 for Full, and 32.5 for Association-only (Sec. E.3). These results support retaining established correspondences across frames, with learned association providing an additional gain under this protocol.

Table 3 compares Full and Concat-only using the same fixed model. Positive-depth projections define ego-only, coop-only, and both-agent bins. Both-agent AP rises from $39.2/42.2$ to $53.5/72.1$ on V2X-Seq/Griffin-25M, while AP in the other bins remains nearly unchanged. The strongest gains thus occur in shared camera coverage, which accounts for approximately 22% of evaluated objects on V2X-Seq versus 74% on Griffin-25M. This smaller fraction, together with the smaller within-bin gain, helps explain the lower overall detection improvement on V2X-Seq.

4.3 Robustness and Communication Efficiency

Calibration Perturbations. Table 4 shows that at 1.0° yaw error, FFBL-Coop loses 5.4 AP points but only 0.7 AMOTA points, while IDS remains close to the unperturbed level. Translation errors show a similar pattern. Across the tested conditions, FFBL-Coop retains higher AP and AMOTA and fewer IDS than SparseCoop. These results show that degraded spatial alignment does not cause a comparable drop in tracking performance in the full pipeline.

Table 3: AP by camera-projection coverage. N: ground-truth count. Concat-only: no in-region fusion.

Bin	V2X-Seq			Griffin-25M		
	N	Concat-only	Full	N	Concat-only	Full
ego-only	1673	34.9	**35.0**	178	**18.6**	17.7
coop-only	2699	62.9	**63.1**	1569	54.9	**55.6**
both	1227	39.2	**53.5**	4959	42.2	**72.1**

Table 4: Calibration sensitivity on V2X-Seq.

Perturbation	FFBL-Coop			SparseCoop		
	AP↑	AMOTA↑	IDS↓	AP↑	AMOTA↑	IDS↓
None	54.8	61.1	70	53.0	42.1	74
Yaw 0.5°	53.3	61.0	72	48.8	41.8	87
Yaw 1.0°	49.4	60.4	67	45.7	40.0	88
Trans. $0.5\,\mathrm{m}$	50.1	59.9	68	46.6	40.9	85

Communication Latency. Table 5 evaluates Griffin-25M with velocity compensation enabled. At $200\,\mathrm{ms}$ delay, FFBL-Coop loses 5.9 AP points but only 0.4 AMOTA points. Thus, stale observations reduce detection accuracy, while aggregate tracking performance remains comparatively stable. FFBL-Coop retains higher AP and AMOTA than the displayed baselines at all tested delays. Its smaller AMOTA drop than SparseCoop, despite a larger AP drop, further illustrates this distinction.

Table 5: Communication latency on Griffin-25M.

Method	AP↑			AMOTA↑		
	0 ms	100 ms	200 ms	0 ms	100 ms	200 ms
Early Fusion	60.7	56.4	50.8	67.0	64.0	58.0
SparseCoop	55.9	55.3	54.0	50.9	49.0	48.6
Ours	**65.3**	**61.6**	**59.4**	**68.8**	**68.6**	**68.4**

Table 6: Feature-compression trade-offs on V2X-Seq. B/s denotes estimated query payload rate.

Codebook	B/s↓	AP↑	AMOTA↑	IDS↓
Uncompressed	1.72×10^5	54.9	61.2	49
$K_{\mathrm{code}} = 1024$	8.64×10^3	54.7	59.4	59
$\boldsymbol{K_{\mathrm{code}} = 256}$	8.48×10^3	54.8	61.1	70
$K_{\mathrm{code}} = 16$	8.48×10^3	54.8	59.2	63
Box only	8.32×10^3	50.6	50.2	199

Communication Efficiency. On the fixed Full V2X-Seq model (Table 6), $K_{\mathrm{code}} = 256$ reduces estimated payload from 1.72×10^5 to $8.48 \times 10^3\,\mathrm{B/s}$, with 0.1-point drops in AP and AMOTA and IDS increasing from 49 to 70. Even 16 codewords yield AP/AMOTA of $54.8/59.2$, versus $50.6/50.2$ without feature transmission, supporting the value of cooperative semantic features for cooperative aggregation. Both 16 and 256 codewords use one-byte indices, giving equal steady-state payloads, while 256 achieves higher AMOTA. Accounting includes all transmitted queries before receiver-side filtering.

5 Conclusion

In this work, we present FFBL-Coop, a *fuse first, bind later* framework that separates cooperative evidence admission from identity inheritance. CSA allocates cooperative evidence, URA guides ego-feature retrieval with cooperative semantics, and CPIA combines learned association with persistent ID mappings to bind identities after refinement. Together, these stages integrate complementary observations while supporting tracking continuity across changes in query placement. On V2X-Seq and Griffin-25M, FFBL-Coop improves detection and tracking with compact feature transmission. Ablations support decoupled identity management, while perturbation studies show comparatively stable tracking despite reduced detection accuracy. Future work will explore uncertainty-aware association and identity coordination across multiple agents under changing visibility and connectivity.

AI USE STATEMENT

Generative AI tools were used to aid or polish writing, for retrieval and discovery (e.g., finding related work), and for research ideation or execution (e.g., reviewer-style feedback). Details are described in the paper. The authors verified all citations, technical descriptions, analyses, and revisions, and take responsibility for the final content of this work.

REFERENCES

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631, 2020.

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 *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, pp. 88–100, 2019.

Shuxiao Ding, Lukas Schneider, Marius Cordts, and Juergen Gall. ADA-Track: End-to-end multi-camera 3D multi-object tracking with alternating detection and association. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15184–15194, 2024.

Ruiyang Hao, Haibao Yu, Jiaru Zhong, Chuanye Wang, Jiahao Wang, Yiming Kan, Wenxian Yang, Siqi Fan, Huilin Yin, Jianing Qiu, Yao Mu, Jiankai Sun, Li Chen, Walter Zimmer, Dandan Zhang, Shanghang Zhang, Mac Schwager, Ping Luo, and Zaiqing Nie. Research challenges and progress in the end-to-end V2X cooperative autonomous driving competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 1849–1860, 2025.

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*, volume 35, pp. 4874–4886, 2022.

Yue Hu, Juntong Peng, Sifei Liu, Junhao Ge, Si Liu, and Siheng Chen. 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.

Yanwei Li, Zhiding Yu, Jonah Philion, Anima Anandkumar, Sanja Fidler, Jiaya Jia, and Jose Alvarez. End-to-end 3D tracking with decoupled queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18302–18311, 2023.

Xuewu Lin, Zixiang Pei, Tianwei Lin, Lichao Huang, and Zhizhong Su. Sparse4D v3: Advancing end-to-end 3D detection and tracking. *arXiv preprint arXiv:2311.11722*, 2023.

Yoseph Linde, Andres Buzo, and Robert M. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28(1):84–95, 1980. doi: 10.1109/TCOM.1980.1094577.

Yen-Cheng Liu, Junjiao Tian, Nathaniel Glaser, and Zsolt Kira. When2com: Multi-agent perception via communication graph grouping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4106–4115, 2020.

Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETR: Position embedding transformation for multi-view 3D object detection. In *European Conference on Computer Vision*, pp. 531–548, 2022.

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*, pp. 4812–4818, 2023.

Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. In *European Conference on Computer Vision*, pp. 194–210, 2020.

Jiahao Wang, Xiangyu Cao, Jiaru Zhong, Yuner Zhang, Zeyu Han, Haibao Yu, Chuang Zhang, Lei He, Shaobing Xu, and Jianqiang Wang. Griffin: Aerial-ground cooperative detection and tracking dataset and benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9867–9875, 2026a.

Jiahao Wang, Zhongwei Jiang, Wenchao Sun, Jiaru Zhong, Haibao Yu, Yuner Zhang, Chenyang Lu, Chuang Zhang, Lei He, Shaobing Xu, and Jianqiang Wang. SparseCoop: Cooperative perception with kinematic-grounded queries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9876–9884, 2026b.

Jiahao Wang, Zikun Xu, Yuner Zhang, Zhongwei Jiang, Chenyang Lu, Shuocheng Yang, Yuxuan Wang, Jiaru Zhong, Chuang Zhang, Shaobing Xu, and Jianqiang Wang. Long-SCOPE: Fully sparse long-range cooperative 3D perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11599–11609, 2026c.

Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3D object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3621–3631, 2023.

Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. DETR3D: 3D object detection from multi-view images via 3D-to-2D queries. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 180–191, 2022.

Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 3D multi-object tracking: A baseline and new evaluation metrics. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 10359–10366, 2020.

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*, pp. 107–124, 2022.

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 *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 989–1000, 2023.

Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3D object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11784–11793, 2021.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5486–5495, 2023.

Haibao Yu, Wenxian Yang, Jiaru Zhong, Zhenwei Yang, Siqi Fan, Ping Luo, and Zaiqing Nie. End-to-end autonomous driving through V2X cooperation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9598–9606, 2025.

Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. MOTR: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision*, pp. 659–675, 2022.

Jiaru Zhong, Jiahao Wang, Jiahui Xu, Xiaofan Li, Zaiqing Nie, and Haibao Yu. CoopTrack: Exploring end-to-end learning for efficient cooperative sequential perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26954–26965, 2025.

APPENDIX

A OVERVIEW

This appendix provides detailed evaluation outputs, qualitative examples, additional ablations, and reproducibility settings for FFBL-Coop. These analyses complement the main results by examining module contributions, operating conditions, and communication efficiency.

Unless stated, results are reported for the `car` class using the benchmark evaluators. Occlusion and range analyses reuse predictions with centre-distance thresholds $\{0.5, 1, 2, 4\}\,\mathrm{m}$. Ground-truth objects outside the evaluated bin are ignored, and predictions matched to ignored objects are excluded from scoring. Metric scales and directions are specified in the corresponding tables.

The supplement is organized as follows:

- Section B reports detailed evaluation outputs on V2X-Seq-SPD and Griffin-25M, including threshold-specific AP and MOTA across recall levels.
- Section C presents qualitative comparisons on V2X-Seq and Griffin-25M.
- Section D examines performance by occlusion severity and object range.
- Section E provides complete module ablations and additional controlled analyses.
- Section F reports robustness and communication-efficiency analyses.
- Section G specifies training, evaluation, and communication-accounting protocols.
- Section H discusses methodological distinctions and limitations.

B OFFICIAL EVALUATION RECORDS

Tables 7 and 8 provide detailed evaluator outputs for the Full checkpoints reported in Table 1. Both use codebook-compressed features, with $K_{\mathrm{code}} = 256$ on V2X-Seq and 512 on Griffin-25M. All results concern the car class. AP, MOTA, and AMOTA are reported in percent.

Table 7: Detection results at centre-distance thresholds of 0.5, 1, 2, and $4\,\mathrm{m}$. Overall AP averages the four threshold-specific scores. ATE and AVE are measured in m and $\mathrm{m/s}$, respectively.

Dataset	AP@0.5	AP@1	AP@2	AP@4	AP↑	ATE↓	AVE↓
V2X-Seq	19.95	49.95	70.05	79.25	54.8	0.504	1.093
Griffin-25M	44.15	66.35	73.05	77.65	65.3	0.291	1.661

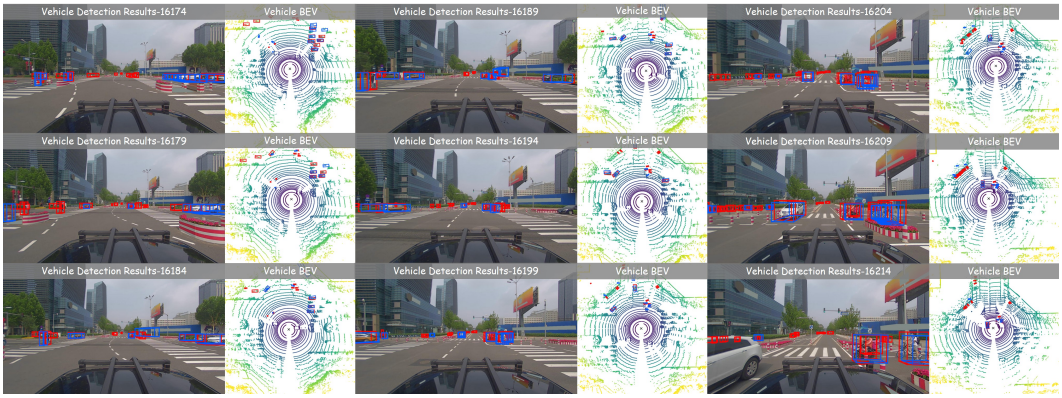
Table 8: Tracking results and transmitted payload rates. TP, FP, FN, and IDS use a $2\,\mathrm{m}$ matching threshold. MOTA subscripts denote recall levels. Communication accounting follows Table 1.

Dataset	B/s	TP	FP	FN	IDS	$\mathrm{MOTA}_{0.3}$	$\mathrm{MOTA}_{0.5}$	$\mathrm{MOTA}_{0.7}$	AMOTA↑
V2X-Seq	8.48×10^3	3510	551	2000	70	27.7	45.5	52.9	61.1
Griffin-25M	3.37×10^4	4982	669	1790	50	30.9	46.2	60.8	68.8

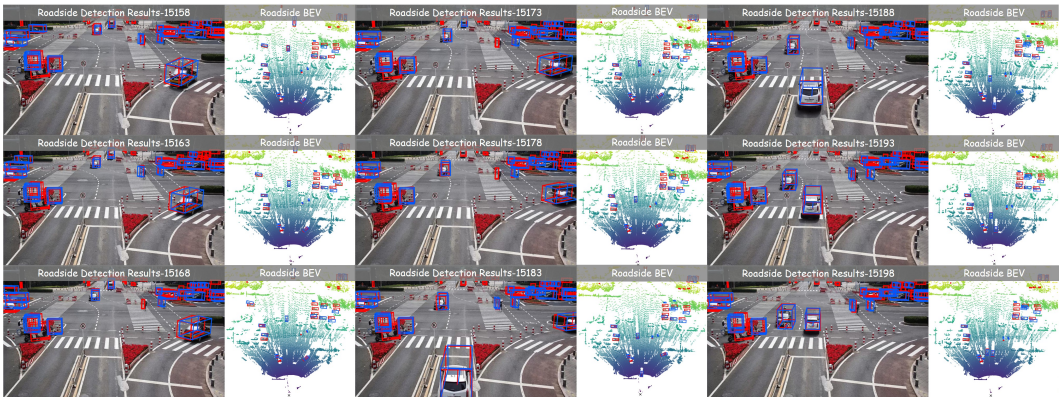
Detection and Localization. Table 7 reports detection accuracy at different localization tolerances. On V2X-Seq, AP rises from 19.95 at $0.5\,\mathrm{m}$ to 79.25 at $4\,\mathrm{m}$. On Griffin-25M, it rises from 44.15 to 77.65. The gains with relaxed thresholds indicate that precise localization remains a limitation. ATE and AVE complement overall AP by separately characterizing center and velocity estimation errors.

Tracking Errors. Table 8 shows that false negatives exceed FP and IDS combined on each dataset. Missed objects therefore remain a substantial source of error at the reported operating points, highlighting the importance of object recovery alongside identity continuity. MOTA at the selected recall levels complements AMOTA by showing performance at individual operating points. Table 2 further examines the contributions of admission, representation refinement, and identity management.

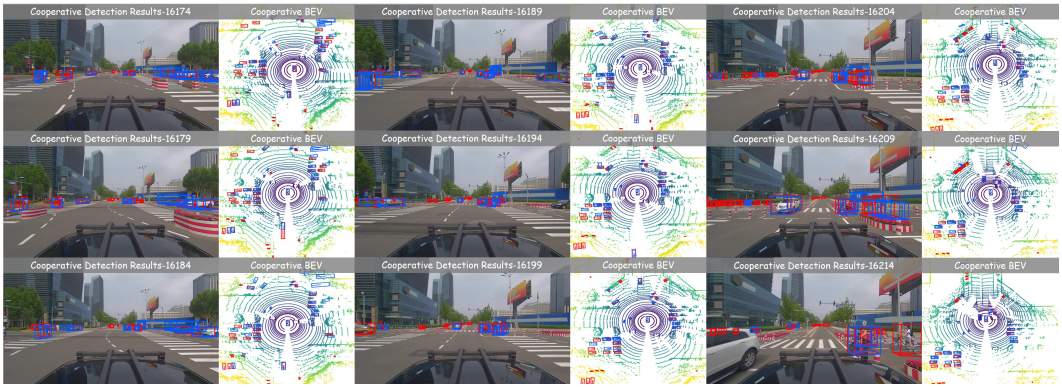
C Additional Qualitative Evidence



(a) Ego-only nine-frame sequence (vehicle tokens 016174–016214). Red: ground truth, blue: detections.



(b) Roadside-only nine-frame sequence (infrastructure tokens 015158–015198, aligned with vehicle tokens).



(c) Cooperative nine-frame sequence on the same vehicle tokens.

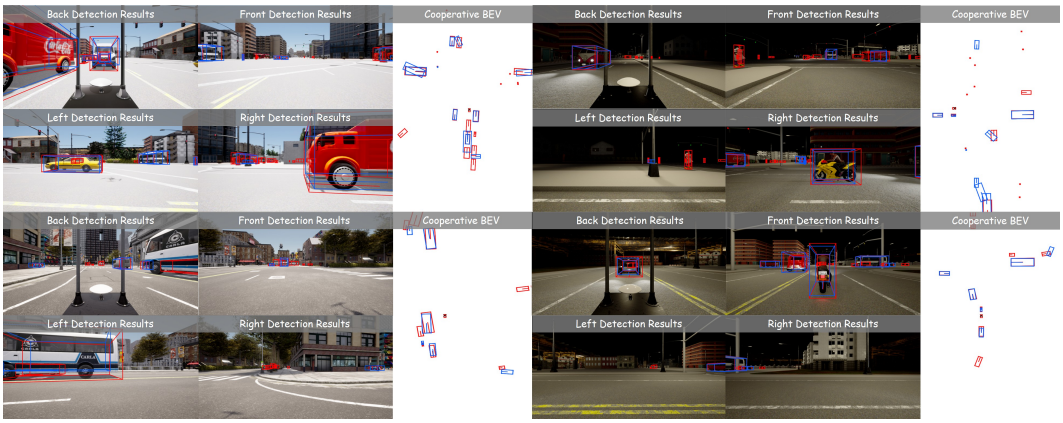
Figure 6: Comparison of ego-only, roadside-only, and cooperative detection sequences on V2X-Seq.

Vehicle–Infrastructure Cooperation. Figure 6 compares ego-only, roadside-only, and cooperative predictions over nine aligned V2X-Seq frames at 2 Hz. Vehicle tokens span 016174–016214, with infrastructure tokens 015158–015198. Roadside predictions include cars near the crosswalk and far-right region that are absent from the ego-only outputs. The cooperative sequence illustrates how these complementary observations extend detection coverage at the intersection.

Vehicle–Drone Cooperation. Figure 7 presents four Griffin-25M validation samples from drone and ego-vehicle perspectives. Each sample pairs a camera view with a BEV visualization, with ground truth in red and detections in blue. The paired views illustrate how viewpoint differences



(a) Drone-side detections on Griffin-25M. Red: ground truth, blue: detections.



(b) Vehicle-side cooperative detections on the same Griffin-25M samples.

Figure 7: Cooperative detections on Griffin-25M from both drone and ego-vehicle perspectives.

change object appearance and visibility. Cooperative outputs show detections of objects that appear small or partially occluded in ego images.

D Occlusion and Range Stratified Evaluation

We examine detection performance across occlusion levels and object ranges using stored predictions without retraining. Base uses the Concat-only configuration: it disables in-region cooperative fusion while retaining out-of-region cooperative box concatenation. Full enables the complete cooperative pipeline. These comparisons measure the benefit of in-region cooperative integration over box concatenation. Ground-truth objects outside the active bin are ignored, and predictions matched to ignored objects are excluded from scoring. AP and recall are reported in percent.

D.1 Occlusion Protocol

For a ground-truth object i, let $\mathcal{V}_i^{\text{ego}}$ denote the ego cameras in which its box has a valid projection with positive center depth. For each camera c, let P_i^c denote the image-clipped convex hull of the projected box and z_i^c its center depth. For objects with nonzero projected area, we compute the projected-occlusion ratio as

$$o_i = \frac{\sum_{c \in \mathcal{V}_i^{\text{ego}}} \left| P_i^c \cap \left(\bigcup_{j : z_j^c < z_i^c} P_j^c \right) \right|}{\sum_{c \in \mathcal{V}_i^{\text{ego}}} |P_i^c|}. \tag{18}$$

Table 9: Car detection by occlusion severity. N counts ground-truth objects. Base uses concat-only input. AP and recall are in percent, and Δ denotes the AP gain in percentage points. R@2m is Full recall at a 2 m matching threshold.

Dataset / bin	N	Base AP	Full AP	Δ	Full R@2m
V2X-Seq / light	2981	64.6	68.8	+4.2	95.2
V2X-Seq / occluded	2618	33.3	38.4	+5.1	84.0
Griffin-25M / light	6528	35.5	56.4	+20.9	70.5
Griffin-25M / occluded	1683	0.9	39.2	+38.3	75.0

Table 10: Car detection by ego-frame range. N counts ground-truth objects. AP is in percent, and Δ denotes the AP gain over concat-only Base in percentage points. Full@0.5 reports AP at a 0.5 m centre-distance threshold.

Dataset / range	N	Base AP	Full AP	Δ	Full@0.5
V2X-Seq / 0–20 m	1193	58.5	59.8	+1.3	29.6
V2X-Seq / 20–40 m	2751	53.4	57.6	+4.2	24.2
V2X-Seq / 40–50 m	1655	33.6	40.3	+6.7	10.9
Griffin-25M / 0–20 m	3742	31.6	47.3	+15.7	33.7
Griffin-25M / 20–40 m	3136	22.3	61.1	+38.8	40.3
Griffin-25M / 40–50 m	1333	3.2	27.6	+24.4	9.6

Only closer ground-truth boxes contribute to the overlap, with all areas measured in pixels. Objects are grouped as light ($o_i < 0.2$) or occluded ($o_i \geq 0.2$). This geometric measure approximates occlusion through projected box overlap.

Occlusion-Stratified Results. Table 9 shows AP improvements in both occlusion bins on both datasets. On V2X-Seq, Full improves AP by 4.2 points for light occlusion and 5.1 points for occluded objects. On Griffin-25M, the corresponding gains are 20.9 and 38.3 points. The larger gain in Griffin's occluded bin indicates a substantial benefit from in-region cooperative integration beyond the retained box-concatenation path.

D.2 Range-Stratified Results

Table 10 evaluates Base and Full predictions across ego-frame distance bins, examining where cooperative input improves detection and where precise localization remains challenging.

Full improves AP over Base in every evaluated range bin, showing that the detection benefits extend across nearby and distant objects. At 40–50 m, AP increases from 33.6 to 40.3 on V2X-Seq and from 3.2 to 27.6 on Griffin-25M. Nevertheless, Full@0.5 remains at 10.9 and 9.6, respectively, indicating that precise localization remains challenging in this range. Figure 8 illustrates the complementary coverage available from roadside observations, which reveal objects obscured in the ego view. This provides a qualitative example of useful cooperative evidence, while the range-stratified results establish detection gains within each bin without isolating their causes.

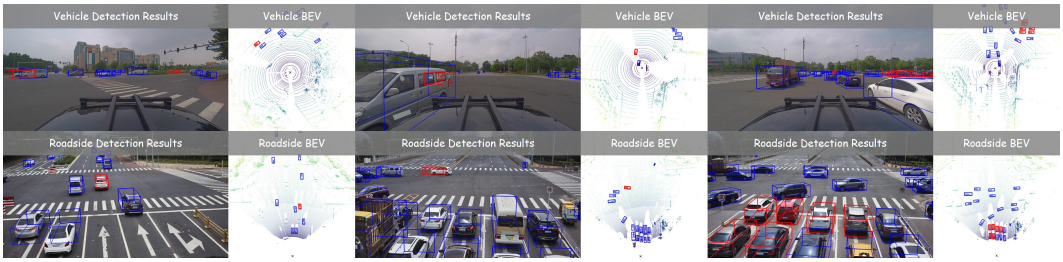


Figure 8: Paired ego and roadside views on V2X-Seq. Nearby vehicles obscure objects in the ego view, while the roadside perspective provides complementary observations.

E Additional Ablation Studies

We extend the main ablations with complete module combinations and additional analyses of admission, refinement, and identity management. Module variants follow the initialization and training protocol in Sec. G. Fixed-checkpoint diagnostics are identified separately.

E.1 Complete Module Ablations

Table 11 reports eight combinations sharing the codebook and training schedule. Cooperative modules and all six decoder layers are trained on both datasets, with the backbone and FPN frozen. Disabled CSA, URA, and CPIA use Hungarian allocation, mean aggregation without local readout, and slot-ID inheritance, respectively. Base retains cooperative input with these replacements.

Table 11: Complete module ablations under a shared training protocol. Base uses the stated replacements. Bold marks the best results.AP and AMOTA are reported in percent.

Base	CSA	URA	CPIA	V2X-Seq			Griffin-25M		
				AP↑	AMOTA↑	IDS↓	AP↑	AMOTA↑	IDS↓
✓				39.9	47.7	123	48.7	48.4	162
✓	✓			52.6	56.5	73	59.2	63.4	**43**
✓		✓		53.2	50.4	116	64.0	55.3	86
✓			✓	53.0	58.9	75	60.5	63.8	66
✓		✓	✓	54.2	60.1	**69**	64.1	66.1	57
✓	✓		✓	53.2	59.0	72	60.6	64.9	48
✓	✓	✓		54.3	59.3	78	62.9	67.3	124
✓	✓	✓	✓	**54.8**	**61.1**	70	**65.3**	**68.8**	50

Individual Contributions. The single-module controls reveal different effects on detection and tracking. URA yields the highest AP among these variants, whereas CPIA yields the highest AMOTA on each dataset. CSA alone also improves both metrics over Base. These patterns suggest that representation quality and identity handling contribute to aggregate tracking performance.

Combined Effects. The complete combinations show that module gains depend on the surrounding allocation and identity configuration. On Griffin-25M, adding URA to CSA raises AMOTA from 63.4 to 67.3 but increases IDS from 43 to 124, with CPIA also enabled, AMOTA reaches 68.8 and IDS decreases to 50. Thus, improved aggregate tracking accuracy can coexist with more identity switches, motivating separate evaluation of overall tracking performance and identity continuity. The combinations also show non-monotonic detection effects: adding CSA to URA-only lowers AP from 64.0 to 62.9, whereas adding it to URA+CPIA raises AP from 64.1 to 65.3. The modules provide complementary benefits without uniformly additive improvements across configurations.

E.2 Hyperparameter Sensitivity

Table 12 varies admission budget K and CPIA association radius τ_{id} individually on fixed Full checkpoints for V2X-Seq and Griffin-25M, with all other inference settings unchanged. The default budgets are 35 and 50, respectively. The default CPIA radii and CSA thresholds are $2.2\,\mathrm{m}$ on V2X-Seq and $1.5\,\mathrm{m}$ on Griffin-25M. The CSA thresholds remain fixed throughout these sweeps to distinguish association-radius sensitivity from allocation gating.

Table 12: Inference-time sensitivity on V2X-Seq. All removes the budget cap. $\tau_{\mathrm{id}} = 0$ disables spatial adoption. AP and AMOTA are in percent. Bold marks the defaults.

Admission budget				CPIA adoption radius			
K	AP↑	AMOTA↑	IDS↓	τ_{id}	AP↑	AMOTA↑	IDS↓
5	53.6	58.7	37	$0\,\mathrm{m}$	54.7	60.5	76
20	54.6	61.2	51	$1.0\,\mathrm{m}$	54.8	61.1	72
35	**54.8**	**61.1**	**70**	$\mathbf{2.2}\,\mathrm{m}$	**54.8**	**61.1**	**70**
All	54.7	61.0	79	$3.0\,\mathrm{m}$	54.6	59.3	59

Table 13: Inference-time sensitivity on Griffin-25M. All removes the budget cap. $\tau_{\mathrm{id}} = 0$ disables spatial adoption. AP and AMOTA are in percent. Bold marks the defaults.

Admission budget				CPIA adoption radius			
K	AP↑	AMOTA↑	IDS↓	τ_{id}	AP↑	AMOTA↑	IDS↓
10	64.3	67.9	55	$0\,\mathrm{m}$	65.2	68.7	63
25	64.9	68.3	53	$0.75\,\mathrm{m}$	65.3	68.8	59
50	**65.3**	**68.8**	**50**	$\mathbf{1.5}\,\mathrm{m}$	**65.3**	**68.8**	**50**
All	65.3	68.8	51	$3.0\,\mathrm{m}$	65.3	67.9	53

Table 14: Detailed ablations on V2X-Seq covering admission, refinement, and CPIA components.

Component	Intervention	AP↑	AMOTA↑	IDS↓
Reference	Full	54.8	61.1	70
CSA	Early prefix admission	50.5	50.9	67
CSA	$\tau_{\mathrm{CSA}} = 3.0\,\mathrm{m}$	54.5	60.0	74
URA	w/o local readout	52.3	54.7	97
URA	Mean-vector feature replacement	51.2	52.9	163
CPIA	Sticky occupancy	54.3	55.3	103
CPIA	Map only	54.7	60.5	76
CPIA	Association only	54.6	32.5	167

Admission Budget. On V2X-Seq, increasing K from 5 to 20 improves AP/AMOTA from $53.6/58.7$ to $54.6/61.2$. Further expansion offers little aggregate benefit: the default $K = 35$ gives slightly higher AP, while $K = 20$ achieves slightly higher AMOTA and fewer IDS. On Griffin-25M, AP and AMOTA improve as K increases from 10 to the default 50, with IDS decreasing from 55 to 50. Removing the cap provides no further AP or AMOTA gain over the default on either dataset. These results support a bounded admission budget: admitting more candidates does not necessarily improve detection or tracking.

Identity Association Radius. AP and AMOTA vary little across the tested radii up to the default on each dataset. Disabling learned association retains map reuse and new-track initialization, yielding AMOTA of $60.5/68.7$ on V2X-Seq/Griffin-25M, compared with $61.1/68.8$ at the defaults. Enabling association also reduces IDS from $76/63$ to $70/50$. Increasing the radius to $3.0\,\mathrm{m}$ lowers AMOTA to $59.3/67.9$, providing no aggregate tracking benefit. IDS nevertheless decreases on V2X-Seq at this radius, so its variation alone does not determine the preferred setting. These sweeps show modest additional gains from learned association alongside persistent mapping, while a wider spatial gate does not consistently improve the metrics.

E.3 Detailed Design Ablations

Table 14 examines admission, feature integration, and identity management on V2X-Seq. CSA and CPIA interventions use same checkpoint. URA variants are retrained under the shared protocol.

Admission Timing and Spatial Gating. Moving admission before the temporal update reduces AP/AMOTA by $4.3/10.2$ points, whereas increasing τ_{CSA} from 2.2 to $3.0\,\mathrm{m}$ reduces them by $0.3/1.1$ points. Under these interventions, the fixed model is more sensitive to admission timing than to the tested spatial-threshold change.

Cooperative Feature Integration. Removing local readout reduces AP/AMOTA from $54.8/61.1$ to $52.3/54.7$ and increases IDS from 70 to 97, supporting its contribution beyond in-place decoding. Replacing instance features with a mean vector yields AP/AMOTA of $51.2/52.9$ and 163 IDS. Together, these retrained controls support the use of instance-specific cooperative semantics and local context for cooperative refinement.

Correspondence Establishment and Temporal Reuse. Map-only yields AP/AMOTA of $54.7/60.5$ and 76 IDS, close to Full's $54.8/61.1$ and 70 IDS. Association-only retains AP at 54.6 but lowers AMOTA to 32.5 and increases IDS to 167. Sticky occupancy yields $54.3/55.3$ and 103 IDS. Map-only initializes mappings through ego-ID assignment during slot occupancy and uses them.

Full additionally associates unmapped queries with existing ego tracks and records accepted correspondences, adding 0.6 AMOTA points under this protocol. These results support persistent use of established ID assignments, with learned association providing an additional benefit. Figure 9 illustrates the identity behavior of Full and sticky-ID across consecutive frames.

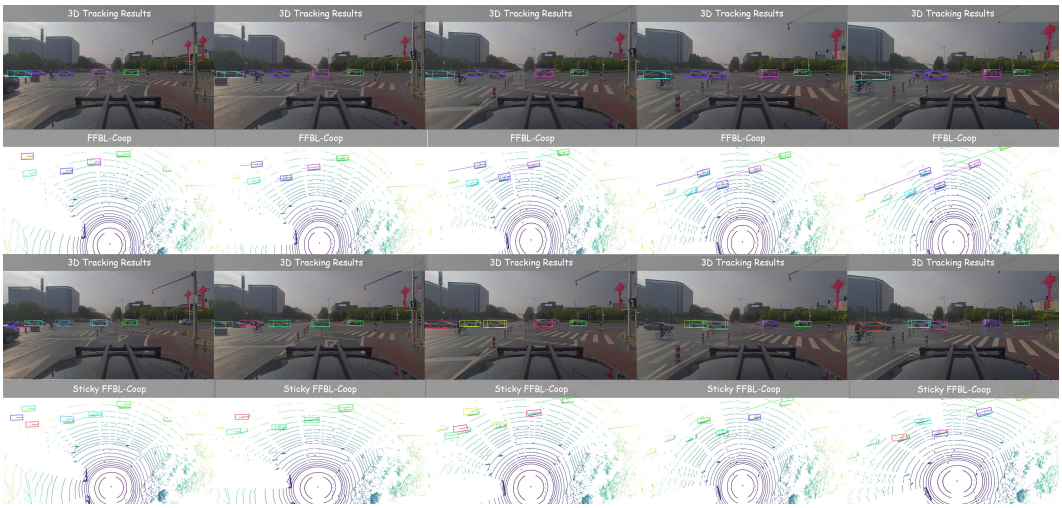


Figure 9: Identity continuity on five consecutive V2X-Seq frames (tokens 010562–010582, 2 Hz). Top: Full (CPIA). Bottom: CSA+URA with sticky IDs. Different colors denote different track IDs.

F Robustness and Communication Efficiency

F.1 Latency Injection Protocol

On Griffin-25M, communication delay is simulated by selecting drone observations d frames earlier from the native 10 Hz stream, with $d \in \{0, 1, 2, 4\}$ corresponding to $\{0, 100, 200, 400\}$ ms. Table 5 reports the 0, 100, and 200 ms settings. Delayed observations are supplied to the 2 Hz ego evaluator.

Before coordinate alignment, each cooperative center is compensated in the drone frame using

$$p_j^{+} = p_j^{\text{coop}} + v_j^{\text{coop}} \Delta t, \qquad \Delta t = t^{\text{ego}} - t^{\text{coop}}. \tag{19}$$

The compensated object state is then transformed into the ego frame before cooperative integration. All delay settings use the same fixed checkpoint with velocity compensation enabled.

On V2X-Seq, we retain the provided timestamps and evaluate calibration perturbations as a separate robustness condition. Yaw and translation are perturbed separately by fixed offsets at the magnitudes reported in Table 4, applied consistently across frames to $T_{\text{coop}\to\text{ego}}$. Model weights and cooperative processing rules remain fixed. SparseCoop is evaluated locally under the same perturbation protocol using fixed model weights.

F.2 Codebook Compression

We evaluate feature compression on a fixed Griffin-25M Full checkpoint. Codebooks are constructed offline by applying k-means to training-set instance features. Each 256-dimensional drone feature is encoded by its nearest codeword index and reconstructed by table lookup at the receiver, following Eq. 5. The accompanying object states and metadata are unchanged.

Initialization and Synchronization. The codebook is constructed offline from training-set instance features and held fixed throughout evaluation. Both agents use identical codebook entries; no online codebook updates are performed. If the codebook is transferred once before query streaming, its float32 payload is $4K_{\text{code}}C$ bytes, amounting to 524,288 bytes for $K_{\text{code}} = 512$ and $C = 256$. The reported C_{tx} measures steady-state query payload and excludes this one-time transfer. Includ-

ing the initial codebook payload, the average rate over a session of D seconds is $C_{\mathrm{tx}} + 4K_{\mathrm{code}}C/D$, excluding packet headers and synchronization metadata.

Per-Query Payload. Under the accounting convention in Sec. G, an uncompressed query uses 1076 bytes, comprising a 1024-byte feature and a 52-byte non-feature record: a 44-byte state, a 4-byte confidence score, and a 4-byte source track ID. Codeword indices use one byte for $K_{\mathrm{code}} \leq 256$ and two bytes for the larger evaluated codebooks, giving 53 and 54 bytes per query, respectively. The one-byte format is retained for $K_{\mathrm{code}} = 1$. All transmitted queries are counted before receiver-side region filtering and admission.

Table 15: Feature compression on a fixed Griffin-25M Full checkpoint. B/s measures steady-state query transmission, excluding one-time codebook synchronization. AP and AMOTA are in percent. Bold marks the default configuration.

Codebook	B/s↓	AP↑	AMOTA↑	IDS↓
Uncompressed	6.71×10^5	65.5	68.9	50
$K_{\mathrm{code}} = 1024$	3.37×10^4	64.7	68.3	53
$\boldsymbol{K_{\mathrm{code}} = 512}$	$\mathbf{3.37 \times 10^4}$	**65.3**	**68.8**	**50**
$K_{\mathrm{code}} = 256$	3.31×10^4	64.3	66.2	68
$K_{\mathrm{code}} = 64$	3.31×10^4	64.1	66.2	64
$K_{\mathrm{code}} = 1$	3.31×10^4	63.3	65.7	63

Accuracy and Communication Trade-off. Table 15 shows that the default codebook reduces estimated steady-state payload by approximately 95.0%, with AP/AMOTA decreasing by $0.2/0.1$ points and IDS remaining at 50. Reducing K_{code} to 256 lowers the payload further but reduces AP/AMOTA to $64.3/66.2$. With a single codeword, all cooperative features reconstruct to the same vector, yielding AP/AMOTA of $63.3/65.7$. These results support compact instance-specific semantic cues for ego-feature aggregation, with the default codebook retaining accuracy close to uncompressed transmission.

G Reproducibility and Evaluation Protocol

Evaluation Protocols. We use official validation splits and benchmark evaluators. V2X-Seq-SPD follows the 2 Hz protocol of Hao et al. (2025); Yu et al. (2023). Griffin-25M (Wang et al., 2026a) uses timestamp-aware cooperative inputs. Unless stated otherwise, evaluation concerns the `car` class within $50\,\mathrm{m}$ using nuScenes-style metrics (Caesar et al., 2020). AP averages centre-distance thresholds $\{0.5, 1, 2, 4\}\,\mathrm{m}$. AP, AMOTA, and MOTA are in percent, IDS in counts, and communication rates in B/s. Table 1 reports published baseline results (Zhong et al., 2025; Wang et al., 2026b). No Fusion denotes the published non-cooperative baseline. Long-SCOPE (Wang et al., 2026c) is a reference using its 0–$50\,\mathrm{m}$ evaluation subset. Stratified analyses reuse stored predictions, excluding out-of-bin ground truth and predictions matched to it. Projection bins require positive camera depth. Full and Concat-only share model weights and retain out-of-region box concatenation. The camera-projection, occlusion-, and range-stratified analyses compare these configurations. In the occlusion and range tables, Concat-only is denoted as Base.

Architecture and Training. We use MMDetection3D with a ResNet-50 backbone, four-level FPN, and six-layer decoder with feature dimension $C = 256$. Images are resized to 544×960 on V2X-Seq and 448×800 on Griffin-25M. Roadside BEV resolution is $0.4\,\mathrm{m}$, aerial image readout uses FPN strides $\{4, 8, 16\}$. Training uses camera-only inputs and the classes $\{\texttt{car}, \texttt{bicycle}, \texttt{pedestrian}\}$. The backbone and FPN remain frozen, cooperative modules and all six decoder layers are trained. We use AdamW with learning rate 3×10^{-4}, weight decay 0.01, batch size 8, cosine annealing, and mixed precision on four GPUs. Single-agent pretraining lasts $800/100$ epochs on V2X-Seq/Griffin-25M, followed by initial cooperative training and $4/8$ epochs of soft-admission fine-tuning. Griffin-25M uses a 24-epoch initial cooperative identity-binding stage. Final fine-tuning continues each preceding cooperative model.

Query Banks and Cooperative Updates. V2X-Seq uses $N = 360$ slots, 120 ego-reserved slots, and admission budget $K = 35$, Griffin-25M uses $N = 900$, 300 ego-reserved slots, and

$K = 50$. CSA follows the first decoder layer and temporal update, with $\tau_{\mathrm{CSA}} = 2.2/1.5\,\mathrm{m}$ and $T_{\mathrm{CSA}} = 0.5/0.3$, respectively. Both Full configurations enable local readout with $k_{\mathrm{nbr}} = 8$, radius $r = 8\,\mathrm{m}$, and Gaussian factor $\exp(-d^2/(2\sigma^2_{\mathrm{URA}}))$, where $\sigma_{\mathrm{URA}} = 4\,\mathrm{m}$. CSA follows Eqs. 10–11 with $\epsilon_\rho = 10^{-6}$. This dimensionless threshold gates initialization, normalization by ρ_i does not attenuate accepted updates. CPIA is supervised between cooperative queries and final decoded states of temporally retained live ego tracks. Ground-truth object identities determine the matched-track or unmatched target y_j. Softly aggregated slots receive detection supervision, their mixture weights do not assign identities. Hard slot indices, identity binding, and map updates are used only at inference.

Output Routing and Identity Management. Aligned queries are filtered in the ego frame using $\sqrt{x^2+y^2} \le 60\,\mathrm{m}$ and $z > -3\,\mathrm{m}$, excluding the ego footprint $x \in [-2.04, 2.04]\,\mathrm{m}$, $y \in [-0.92, 0.92]\,\mathrm{m}$. Valid queries with positive-depth center projections inside at least one ego image enter $\widetilde{Q}^{\mathrm{in}}_t$ for CSA/URA. Other valid queries follow the separate box-output path, invalid queries are discarded. Projection coverage does not imply absence of occlusion, and the routing region differs from the evaluation range. Both paths reuse valid mappings in M, preserving the mapped ego ID across path changes. The merged output contains at most one hypothesis per ego ID. Mapping reuse requires confidence at least 0.1, a live and currently unclaimed ego ID, and $t - t_{\mathrm{last}} \le T_{\mathrm{expire}}/\alpha$. The aging factor $\alpha \in [1, 2]$ depends on confidence and update frequency. Base expiration times are $30/16\,\mathrm{s}$ on V2X-Seq/Griffin-25M, with reference frequencies of $2/10\,\mathrm{Hz}$. Expired entries fail lookup and are periodically removed. Before each sequence, the evaluator resets the temporal bank, identity map, and ID counter, independently of batch-size changes.

Ablation Protocols. Within each dataset, module variants share single-agent initialization, the codebook, and the training schedule. Cooperative modules are randomly initialized, and each variant completes all cooperative stages, fine-tuning from its own preceding model. Each variant is trained three times, selecting checkpoints by validation AMOTA within each run. All reported metrics come from the run with median AMOTA. Hyperparameter, compression, and CSA/CPIA interventions use the Full weights reported in Table 1, changing only the specified inference conditions. The local-readout and mean-vector controls in Table 14 are retrained.

Communication and Codebook Synchronization. Transmitted query payloads are counted before receiver-side filtering or admission:

$$C_{\mathrm{tx}} = f_{\mathrm{tx}} \bar{N}_{\mathrm{tx}} b_q \qquad [\mathrm{B/s}], \tag{20}$$

where f_{tx} is transmission frequency, $\bar{N}_{\mathrm{tx}}$ is mean transmitted query count, and b_q is bytes per query. Each query contains 44 bytes of float32 state, 4 bytes of confidence, and a 4-byte source ID. Uncompressed features require 1024 bytes, codeword indices use one byte for $K_{\mathrm{code}} \le 256$ and two bytes for the tested larger codebooks. Total payloads are therefore 1076, 53, or 54 bytes per query, box-only transmission uses 52 bytes. Packet headers are excluded, and published baseline rates retain their original accounting conventions. V2X-Seq transmits approximately 80 queries per frame at $2\,\mathrm{Hz}$, yielding $8480\,\mathrm{B/s}$ with $K_{\mathrm{code}} = 256$. Griffin-25M transmits approximately 62.4 queries per drone frame at $10\,\mathrm{Hz}$, yielding $3.37 \times 10^4\,\mathrm{B/s}$ with $K_{\mathrm{code}} = 512$. These counts differ from bank capacity and admission budget. Codebooks are constructed offline from training-set features and remain fixed throughout evaluation. Reported C_{tx} includes query payload only. If an initial codebook transfer is required, including its float32 payload gives an average rate of $C_{\mathrm{tx}} + 4K_{\mathrm{code}}C/D$ over a session of D seconds, excluding packet headers and synchronization metadata. No subsequent codebook-update traffic is incurred in the evaluated setting.

H Method Notes and Limitations

Relationship to Cooperative Query Trackers. CoopTrack (Zhong et al., 2025), SparseCoop (Wang et al., 2026b), and Long-SCOPE (Wang et al., 2026c) develop cross-agent association and feature refinement using graph affinities, geometric and appearance cues, or contextual matching. FFBL-Coop builds on these mechanisms through *fuse first, bind later*: CSA allocates remote evidence, URA refines hypotheses with ego observations, and CPIA resolves identities using persistent mappings, learned association, and new-track initialization. Its distinction lies in separating

query placement from identity inheritance within the recurrent pipeline, rather than introducing unmatched-query retention or ego-image refinement as new capabilities.

Dependence on Cooperative Evidence. Performance depends on cooperative coverage, detection quality, and calibration. In-region refinement relies on useful ego observations, while out-of-region concatenation retains cooperative localization errors. Persistent mappings depend on consistent source IDs and may propagate accepted identity errors. Evaluation primarily concerns cars, broader category coverage remains to be validated.

Robustness and Communication. Calibration and latency remain limitations despite comparatively stable AMOTA under the tested conditions. A 1° yaw perturbation reduces V2X-Seq AP/AMOTA from $54.8/61.1$ to $49.4/60.4$. At $200\,\mathrm{ms}$ latency, Griffin-25M AP decreases from 65.3 to 59.4, while AMOTA changes from 68.8 to 68.4 with velocity compensation. These evaluations assess the complete pipeline and do not isolate individual robustness mechanisms. Reported query payload rates exclude packet headers and codebook synchronization, so deployment traffic depends additionally on update frequency and protocol overhead. Less reliable calibration, variable delays, packet loss, and identity coordination across multiple agents remain directions for future work.