CETM: COOPERATIVE EVIDENCE TRUST MANAGEMENT FOR AERIAL-GROUND COOPERATIVE PERCEPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Aerial-ground cooperative 3D perception fuses complementary observations from ground vehicles and unmanned aerial vehicles, offering a promising paradigm for enhancing automated-driving perception, extending sensing coverage, and alleviating occlusion. However, aerial-ground deployment is constrained by two key bottlenecks, namely cross-platform observation heterogeneity that leads to inconsistent spatial evidence among candidates and communication temporal perturbations that tend to cause unstable tracking identities. To overcome these limitations, we propose Cooperative Evidence Trust Management (CETM), an online receiver-side management framework for aerial-ground cooperative perception. Specifically, the framework contains two complementary modules, where Spatial Reliability Calibration (SRC) calibrates candidate ranking using cross-route spatial evidence, while Lineage-Guided Memory (LGM) maintains stable scene-local identities using spatiotemporal lineage evidence. Extensive experiments on the Griffin dataset demonstrate that CETM achieves state-of-the-art cooperative detection and tracking performance, outperforming existing methods by a large margin and exhibiting strong robustness to communication perturbations. The code and model will be publicly available.

1 INTRODUCTION

Although automated-driving perception has made substantial progress (Zhou et al., 2023; Zhang et al., 2022a; 2023), a single ground platform remains fundamentally limited by a restricted field of view, occlusion, and insufficient observation range. Cooperative perception (Caillot et al., 2022; Xu et al., 2022a;b; 2025) has therefore emerged as an important paradigm for overcoming these single-platform limitations. By allowing vehicles and roadside infrastructure (Gong et al., 2026; Wang et al., 2025) to exchange information, cooperative perception can integrate complementary observations to enhance the overall perception performance of the multi-agent system.

However, conventional vehicle-to-vehicle (Li et al., 2026) and vehicle-to-infrastructure (Sekaran et al., 2025) cooperation remains constrained by platform placement, viewpoint coverage, and communication conditions, making it difficult to provide consistently complementary observations. Aerial-ground cooperation combines ground vehicles with unmanned aerial vehicles, exploiting the elevated, wide-area, and flexibly deployable observations of aerial platforms to supplement environmental information that ground platforms cannot readily obtain. Griffin (Wang et al., 2026a) is the first to establish this setting as a cooperative benchmark, positioning reliable online detection and tracking for aerial-ground cooperation as a critical evaluation task for practical deployment.

This goal presents two major challenges, as shown in Fig. 1(a). The first is how to assess the spatial reliability of current cooperative candidates. Aerial and ground platforms differ substantially in altitude, viewing direction, occlusion pattern, and localization quality; consequently, the same target may produce spatially displaced or duplicated candidates with different confidence rankings across platforms or cooperative routes. If candidates are selected directly according to the detection score of a single route, the receiver cannot distinguish a high-response candidate lacking spatial support from one jointly supported by multiple observations.

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.

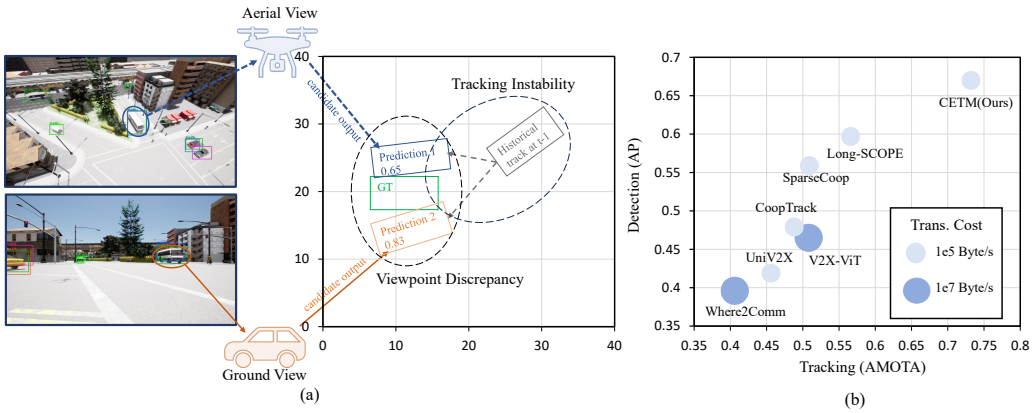

Figure 1: (a) Core challenges in aerial-ground cooperative perception and (b) performance comparison on Griffin dataset.

The second is how to maintain stable target identities across consecutive frames. Communication delay, intermittent packet loss, and candidate replacement can change candidate arrival times, counts, and source relations, preventing current candidates from being directly matched to historical observations. SORT (Bewley et al., 2016) associates detections using motion and geometric proximity, ByteTrack (Zhang et al., 2022b) preserves low-confidence detections during matching, and CoopTrack (Zhong et al., 2025) extends association to cooperative observations. However, these frame-wise strategies can still produce identity switches, fragmented tracks, and erroneous identity inheritance under aerial-ground temporal asynchrony.

To address these issues, we propose Cooperative Evidence Trust Management (CETM), an online receiver-side management framework for aerial-ground cooperative perception. Specifically, CETM contains two complementary modules, where Spatial Reliability Calibration (SRC) reassesses candidate reliability using cross-route spatial consistency, while Lineage-Guided Memory (LGM) uses spatiotemporal lineage evidence to maintain stable scene-local identities. Extensive experiments on the Griffin dataset (Wang et al., 2026a) demonstrate that CETM achieves state-of-the-art cooperative detection and tracking performance as shown in Fig. 1(b), outperforming existing methods by a large margin and exhibiting strong robustness to communication perturbations.

In summary, the main contributions of this paper are as follows:

- We propose CETM, an online aerial-ground cooperative perception framework that organizes cooperative candidates into a continuous evidence-to-identity management process.
- We develop Spatial Reliability Calibration (SRC), which uses cross-route spatial evidence to recalibrate candidate ranking.
- We design Lineage-Guided Memory (LGM), which uses spatiotemporal lineage evidence to maintain stable scene-local identities.
- Extensive experiments on the Griffin dataset demonstrate that CETM achieves SOTA performance, outperforming existing methods by a large margin and exhibiting strong robustness to communication perturbations.

2 Related Work

2.1 Cooperative 3D Perception

Multi-agent 3D cooperative perception uses complementary observations to extend sensing coverage and compensate for incomplete target visibility (Caillot et al., 2022). Existing approaches are commonly organized by exchange level. Early cooperation (Chen et al., 2019; Arnold et al., 2020; Yu et al., 2022) exchanges raw data or low-level features, intermediate cooperation (Liu et al., 2020b; Wang et al., 2020; Xu et al., 2023) exchanges BEV or compressed representations, and

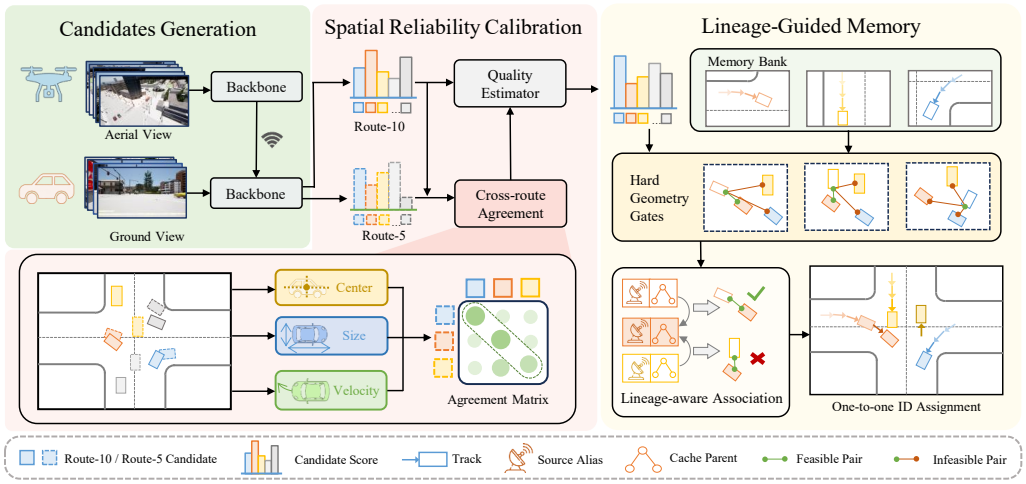

Figure 2: Overview of Cooperative Evidence Trust Management. Spatial Reliability Calibration uses cross-route evidence to recalibrate primary candidates, while Lineage-Guided Memory associates calibrated candidates with scene-local tracks to produce reliable detections and temporally stable identities.

late cooperation (Li et al., 2022; Xu et al., 2022b) exchanges object-level results such as boxes or queries. Some studies (Liu et al., 2020a; Li et al., 2021) further introduce communication graph grouping, cross-agent matching, and motion modeling to improve cooperative perception. Despite these advances, most vehicle-vehicle and vehicle-infrastructure methods assume relatively compatible viewpoints and stable motion relationships. In aerial-ground cooperation, platform heterogeneity and communication disturbances can lead to spatially inconsistent candidates and disrupted temporal association (Lei et al., 2024).

2.2 Aerial-Ground Cooperative Perception

Aerial-ground cooperative perception combines near-field vehicle observations with elevated unmanned-aerial-vehicle views and requires evaluation across both detection and temporal tracking. The Griffin (Wang et al., 2026a) dataset establishes a unified benchmark with continuous observations, occlusion, aerial-platform motion, communication delay, data loss, and localization noise. Where2Comm (Hu et al., 2022) selects spatial regions with high cooperative value, CoopTrack (Zhong et al., 2025) associates cross-agent observations with temporal trajectories, and SparseCoop (Wang et al., 2026b) uses instance-level queries with explicit geometric and motion information to reduce reliance on dense BEV representations. However, aerial-ground cooperation introduces reliability and continuity issues beyond single-frame detection or fusion. Viewpoint, occlusion, and localization differences can produce candidates with inconsistent positions (Lu et al., 2022), scores (Guo et al., 2017), or duplicate hypotheses, while communication delay and data loss (Wang et al., 2026a) can make candidates delayed, replaced, or temporarily absent (Lei et al., 2024). Existing methods provide limited explicit management of these reliability differences and rarely offer a causal, auditable mechanism for tracking candidate sources and identity continuation (Zhong et al., 2025). In this paper, we propose Cooperative Evidence Trust Management for managing spatial reliability and temporal identity continuity.

3 Method

3.1 Task Definition and Pipeline Overview

In this paper, we focus on online 3D detection and tracking in aerial-ground cooperative perception, following the benchmark setting established by Griffin (Wang et al., 2026a). A ground vehicle and one or more aerial agents observe a shared scene from complementary viewpoints. The ground vehicle receives its local observations together with asynchronous cooperative information, which

may correspond to an earlier timestamp because of communication delay. At time t, the receiver obtains a candidate set $\mathcal{D}_t = \{d_i^t\}_{i=1}^{N_t}$, where each candidate is represented as:

$$d_i^t = \big(b_i^t,\, c_i^t,\, s_i^t,\, m_i^t,\, \ell_i^t\big), \tag{1}$$

where b_i^t, c_i^t, s_i^t, m_i^t, and ℓ_i^t denote its 3D box, class label, confidence, motion attributes, and source or lineage metadata. Our goal is to recover reliable objects within the ego vehicle's region of interest from these heterogeneous, temporally misaligned observations and to associate valid observations across frames. The resulting predictions form temporally consistent 3D detections and tracks with calibrated confidence, motion states, and stable scene-local identities.

As illustrated in Fig. 2, the frozen sparse cooperative detector first processes the ground and aerial inputs and produces Route-10 primary candidates together with Route-5 supporting candidates. After all candidates are expressed in the ego coordinate system, Spatial Reliability Calibration uses cross-route agreement in center, size, and velocity to recalibrate the ranking of Route-10 candidates while preserving their boxes and lineage metadata. The calibrated candidates then enter Lineage-Guided Memory, which predicts active scene-local tracks from committed historical states and removes incompatible candidate-track pairs through hard geometric gates. Source aliases and cache parents are used only among the remaining feasible relations, after which deterministic one-to-one assignment resolves competition and updates the scene memory. Accepted matches retain stable tracking IDs, valid unmatched candidates initialize new tracks, and expired tracks are removed. Memory and lineage state are reset at scene boundaries. The complete receiver-side procedure is summarized in Algorithm 1 of Appendix A.3.

3.2 Spatial Reliability Calibration

In aerial-ground cooperative scenarios, observations from vehicles and unmanned aerial vehicles have different viewpoints, occlusion patterns, and spatial localization quality. Even when they correspond to the same object, they may produce spatially displaced, differently scored, or duplicated candidates in the current frame. Similarly, the same detector may produce different candidate rankings under different cooperative interaction routes, so a high-scoring candidate from one route may not have sufficient support across routes. Because conventional outputs rely directly on single-frame detector scores, the CETM framework has difficulty distinguishing a high response from more reliable spatial evidence, a concern related to confidence calibration in modern neural predictors, including the approaches of Guo et al. (Guo et al., 2017) and Kendall and Gal (Kendall & Gal, 2017). We therefore introduce Spatial Reliability Calibration (SRC), which uses cross-route spatial consistency to reassess the reliability of current candidates.

SRC first treats Route-10 candidates as primary candidates and searches for Route-5 support. Let p_i^t, z_i^t, and v_i^t denote candidate center, size, and velocity. The feasible cross-route pair set is

$$\mathcal{A}_t = \big\{(i,j) \,\big|\, c_i^t = c_j^t,\ \|p_i^t - p_j^t\| \le \delta_p,\ \rho(z_i^t, z_j^t) \le \delta_z,\ \|v_i^t - v_j^t\| \le \delta_v\big\}, \tag{2}$$

where $i \in \mathcal{D}_t^{10}$, $j \in \mathcal{D}_t^5$, ρ is the maximum per-axis size ratio, and the fixed thresholds are $\delta_p = 1.0$ m, $\delta_z = 2.0$, and $\delta_v = 4.0$ m/s. Deterministic one-to-one matching over $\mathcal{A}_t$ selects at most one supporting candidate $j(i)$ for each primary candidate. If a match exists, the two route scores form $s_i^{\mathrm{cons},t}$; otherwise, $s_i^{\mathrm{cons},t} = s_i^t$. Route-5 is thus a reliability source rather than an independent output stream, so candidate membership is neither expanded nor pruned.

More specifically, let $j(i)$ denote the matched Route-5 candidate for d_i^t. The cross-route consensus score is defined as

$$s_i^{\mathrm{cons},t} = \begin{cases} 1 - (1 - s_i^t)(1 - s_{j(i)}^{5,t}), & \text{if a valid match exists,} \\ s_i^t, & \text{otherwise.} \end{cases} \tag{3}$$

This formulation increases the support for a primary candidate when an independently routed candidate provides consistent evidence, while leaving unsupported candidates at their original score.

After obtaining the inter-route correspondences, SRC constructs a spatial descriptor x_i^t for each primary candidate. The descriptor summarizes the scores from both routes, center and size agreement, velocity compatibility, target range and bearing, local candidate density, object dimensions, available

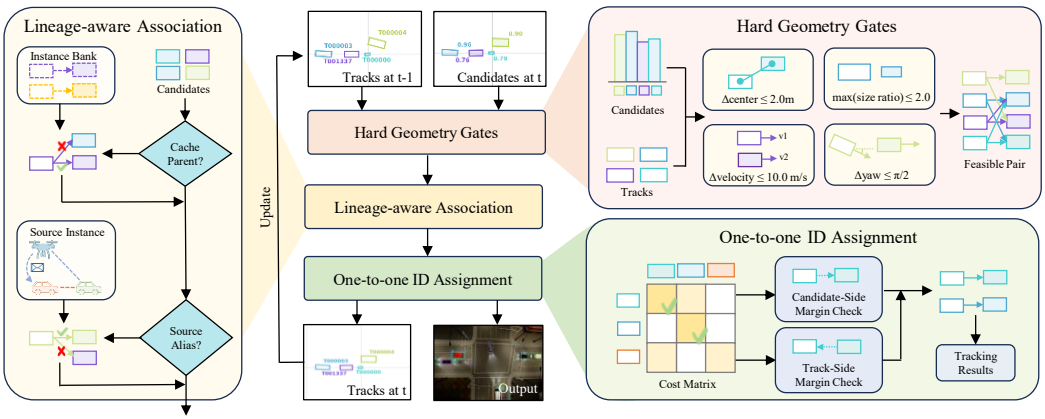

Figure 3: Architecture of Lineage-Guided Memory. Hard Geometry Gates retain feasible candidate-track pairs; Lineage-aware Association prioritizes them using source aliases and cache parents; One-to-one ID Assignment performs competition-aware matching for stable identity continuation.

attributes, and the route-matching state. A frozen quality estimator produces the candidate quality estimate

$$q_i^t = Q\big(x_i^t\big)\,, \tag{4}$$

where Q is fitted only on the Train37 OOF protocol described in Appendix B.3. SRC then geometrically blends this estimate with the cross-route consensus score as follows.

$$\widehat{s}_i^t = (q_i^t)^{0.25}(s_i^{\mathrm{cons},t})^{0.75}. \tag{5}$$

Here, $s_i^{\mathrm{cons},t}$ retains the detector's original response and cross-route support, while q_i^t calibrates the candidate's spatial reliability. The final score $\widehat{s}_i^t$ ranks candidates, giving spatially corroborated candidates more stable preference than unsupported single-route responses.

SRC operates on current-frame reliability. Calibration preserves candidate boxes, class labels, membership, and lineage metadata; a score-independent key recovers the native query, source alias, cache parent, timestamp, and cooperative source information for subsequent association. SRC is therefore a score-calibration module that forwards candidates to LGM for identity management.Additional SRC details are provided in Appendix A.1.

3.3 Lineage-Guided Memory

Single-frame ranking cannot determine whether a current observation should continue a stable trajectory. In aerial-ground cooperation, delay, packet loss, and candidate replacement can change candidate counts, query indices, and score ordering across frames. Geometry-only association, as used in SORT (Bewley et al., 2016), DeepSORT (Wojke et al., 2017), and ByteTrack (Zhang et al., 2022b), can therefore create identity switches or track fragments after short observation gaps. The CETM framework maintains scene-local tracks at the receiver and conditionally transfers identity using causal motion, geometry, and lineage evidence.

For each active track T_k, the scene memory stores its stable ID, most recent box, velocity, timestamp, lifecycle state, native-alias set $\mathcal{N}_k$, and source-alias set $\mathcal{A}_k$. This state is scene-local and reset at scene boundaries. Fig. 3 summarizes the association stages. Let p_k and v_k be the center and velocity of the most recent observation. Its predicted center at the current time is

$$\widetilde{p}_k^t = p_k + \Delta t\, v_k. \tag{6}$$

LGM uses only track states committed at the previous or earlier time steps for causal motion prediction and does not read future frames or ground-truth annotations. This causal state propagation follows the motion-prediction principle of Kalman filtering (Kalman, 1960) and the tracking-by-detection

formulation used by SORT (Bewley et al., 2016). The prediction provides a temporal reference for candidate–track compatibility screening, allowing CETM to continue checking potential associations when a candidate is temporarily missing or cooperative information arrives late.

Before association, LGM forms the following hard-geometry-compatible pair set:

$$\mathcal{G}_t = \left\{ (i,k) \,\middle|\, \|p_i^t - \widetilde{p}_k^t\| \le \delta_p^t,\ \rho(z_i^t, z_k) \le \delta_z^T,\ \|v_i^t - v_k\| \le \delta_v^T,\ \Delta\psi_{ik} \le \delta_\psi \right\}. \tag{7}$$

Only pairs in $\mathcal{G}_t$ enter lineage-guided association. The fixed gates are $\delta_p^t = 2.0 + 2.0\Delta t$ m, $\delta_z^T = 2.0$, $\delta_v^T = 10$ m/s, and $\delta_\psi = \pi/2$ for moving objects. Geometry has the highest priority. Lineage cues cannot override an implausible spatial or motion relation.

For candidate d_i^t, let its lineage metadata be $\ell_i^t = (r_i^t, a_i^t)$, where r_i^t is an optional cache-parent native identity and $a_i^t = (u_i^t, n_i^t)$ is an optional source alias containing the source agent and source-instance identity. Within $\mathcal{G}_t$, LGM forms cache-parent and source-alias proposals as:

$$\begin{aligned} \mathcal{P}_{\text{cache}} &= \{(i,k) \in \mathcal{G}_t \mid r_i^t \in \mathcal{N}_k\}, \\ \mathcal{P}_{\text{source}} &= \{(i,k) \in \mathcal{G}_t \mid a_i^t \in \mathcal{A}_k,\ \omega_k \in \{\text{confirmed}, \text{lost}\}\}. \end{aligned} \tag{8}$$

where ω_k is the track lifecycle state. Cache-parent proposals take precedence over source-alias proposals; both apply only to geometrically feasible continuity hypotheses.

CETM commits non-conflicting priority proposals and locks their candidate and track endpoints. For the remaining gated pairs, it uses normalized geometric costs:

$$C_{ik} = 0.45C_{ik}^{\text{ctr}} + 0.20C_{ik}^{\text{size}} + 0.20C_{ik}^{\text{vel}} + 0.15C_{ik}^{\text{yaw}}. \tag{9}$$

with

$$\begin{aligned} C_{ik}^{\text{ctr}} &= \frac{\|p_i^t - \widetilde{p}_k^t\|}{2 + 2\Delta t}, \quad C_{ik}^{\text{size}} = \frac{1}{3\log 2} \sum_{d \in \{x,y,z\}} \left| \log \frac{z_{i,d}^t}{z_{k,d}} \right|, \\ C_{ik}^{\text{vel}} &= \frac{\|v_i^t - v_k\|}{10}, \quad C_{ik}^{\text{yaw}} = \frac{\Delta\psi_{ik}}{\pi/2}. \end{aligned} \tag{10}$$

CETM then performs deterministic minimum-cost bipartite matching over the remaining gated pairs as follows:

$$\mathcal{M}_t^* = \arg\min_{\mathcal{M} \subseteq \mathcal{G}_t^{\text{rem}}} \sum_{(i,k) \in \mathcal{M}} C_{ik}, \quad \text{s.t.} \quad |\mathcal{M}(i)| \le 1,\ |\mathcal{M}(k)| \le 1. \tag{11}$$

For a residual match (i,k), let $\mathcal{K}_i$ and $\mathcal{I}_k$ be its feasible competing tracks and candidates. The two-sided ambiguity margin is

$$\Delta C_{ik} = \min\left\{ \min_{k' \in \mathcal{K}_i \setminus \{k\}} (C_{ik'} - C_{ik}),\ \min_{i' \in \mathcal{I}_k \setminus \{i\}} (C_{i'k} - C_{ik}) \right\}, \tag{12}$$

where an empty competitor set uses the cost threshold 1.0 as its alternative. Priority pairs are retained, and the accepted residual association set is

$$\widehat{\mathcal{M}}_t = \{(i,k) \in \mathcal{M}_t^* \,|\, C_{ik} \le 1.0,\ \Delta C_{ik} \ge 0.05\}, \tag{13}$$

where $\mathcal{G}_t^{\text{rem}}$ excludes endpoints locked by priority proposals. Accepted matches update the corresponding track; valid unmatched candidates can initialize new tracks, while unmatched historical tracks are retained for a finite lifetime.

Pairs with conflicting lineage cues or insufficient separation from competing matches remain unmatched. By transferring identities only through geometrically feasible and unambiguous associations, LGM prevents erroneous identity inheritance and maintains track continuity when candidates are delayed, replaced, or temporarily missing. Additional LGM details are provided in Appendix A.2.

4 Experiments

4.1 Datasets, Protocol, and Metrics

We evaluate CETM on Griffin (Wang et al., 2026a), a unified simulation benchmark for aerial-ground cooperative 3D detection and tracking. Griffin provides continuous sequences from complementary

Table 1: Quantitative comparison on Griffin. TC is transmitted bytes per second. The best results highlighted in bold and the second-best underlined.

Method	AP ↑	AMOTA ↑	TC (BPS) ↓	CE (FPS) ↑
No Fusion	37.5	36.5	0	8.10
Late Fusion	37.8	37.7	1.56×10^3	6.83
Early Fusion	<u>60.7</u>	<u>67.0</u>	3.11×10^8	5.17
V2X-ViT	46.5	50.8	2.56×10^6	7.56
Where2Comm	39.6	40.6	3.30×10^5	7.60
UniV2X	41.9	45.6	5.58×10^4	7.06
CoopTrack	47.9	48.8	8.00×10^5	6.23
SparseCoop	55.9	50.9	9.73×10^4	11.64
Long-SCOPE	59.7	56.6	1.90×10^5	7.68
CETM (**Ours**)	**67.0**	**73.2**	9.73×10^4	8.47

Table 2: Component-wise ablation of CETM. The rows compare the native detector, Spatial Reliability Calibration (SRC), Lineage-Guided Memory (LGM), and their combination.

Method	SRC	LGM	AP ↑	AMOTA ↑	IDS ↓	Recall ↑	FRAG ↓
Native	–	–	58.00	54.06	595	66.08	142
Calibration-only	✓	–	66.99	54.51	598	66.57	139
Memory-only	–	✓	58.00	72.97	29	76.91	10
Full CETM	✓	✓	**66.99**	**73.18**	39	**79.33**	16

ground-vehicle and unmanned-aerial-vehicle viewpoints. It also incorporates challenging conditions such as occlusion, aerial-platform motion, communication delay, packet loss, and localization noise. We follow the official data split and preprocessing protocol. All experiments use a frozen SparseCoop checkpoint, fixed dual-route settings, a fixed SRC manifest, and fixed association parameters. Frames are processed chronologically, with LGM memory committed before the next frame and reset at each scene boundary. Runtime is measured on one NVIDIA RTX 3090 and includes both detector routes and all receiver-side CETM processing. The fixed receiver-side settings are listed in Appendix A.4 and Table 5, while the reproducibility protocol is given in Appendix B.1.

Following the nuScenes evaluation protocol (Caesar et al., 2020), we measure detection using average precision (AP) and tracking using average multi-object tracking accuracy (AMOTA). To distinguish detection coverage from temporal identity stability, we additionally report identity switches (IDS), recall, and track fragmentations (FRAG). Higher AP, AMOTA, and recall indicate better performance, whereas lower IDS and FRAG indicate more stable trajectories.

4.2 Main Results

Table 1 compares detection and tracking performance on the Griffin dataset within the 0–50 m evaluation range. We compare CETM with No Fusion, Late Fusion, Early Fusion, V2X-ViT, Where2Comm, UniV2X, CoopTrack, SparseCoop (Wang et al., 2026b), and Long-SCOPE (Wang et al., 2026c), and report AP, AMOTA, transmission cost (TC), and computational efficiency (CE).

As shown in Table 1, CETM achieves new state-of-the-art detection and tracking performance, reaching 67.0 AP and 73.2 AMOTA. Compared with SparseCoop, CETM improves AP by 11.09% and AMOTA by 22.28% while retaining the same 9.73×10^4 BPS cooperative payload. It also surpasses Long-SCOPE by 7.29% AP and 16.58% AMOTA while using only 51.2% of its 1.90×10^5 BPS transmission cost. These results demonstrate the complementary effectiveness of Spatial Reliability Calibration and Lineage-Guided Memory under a fixed detector and cooperative payload. Communication accounting is detailed in Appendix B.6.

4.3 Visualization Results

Fig. 4 compares the tracking results of CETM, SparseCoop, and the baseline on Griffin. CETM produces accurate bounding boxes and maintains stable identities across consecutive frames. SparseCoop

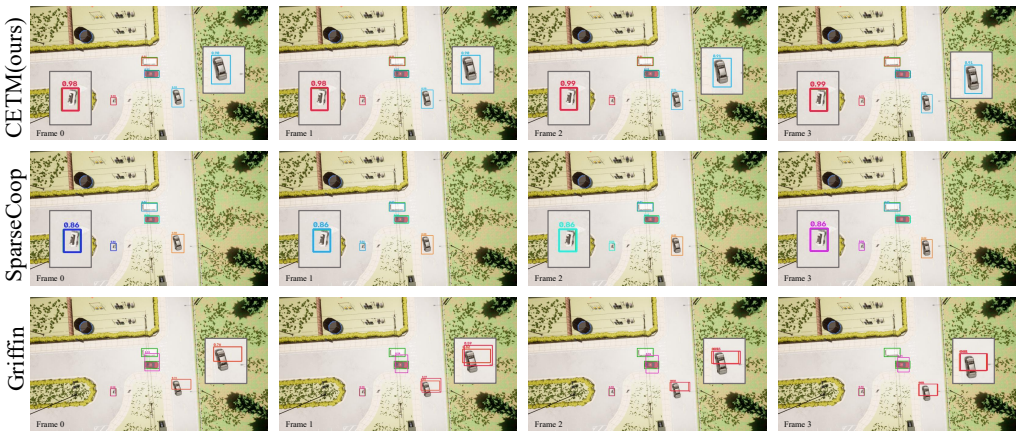

Figure 4: Visualization of tracking results on the Griffin dataset. From top to bottom: CETM, SparseCoop, and Griffin.

Table 3: Hard-geometry-gate ablation for LGM. All variants use identical SRC detections.

LGM gate	AMOTA ↑	IDS ↓	FRAG ↓	Recall ↑
No hard geometry	67.19	65	38	70.91
Center-only	67.18	58	37	73.08
Full: center + size + velocity + yaw	**73.18**	**39**	**16**	**79.33**

detects most targets, but frequent ID changes lead to fragmented trajectories. The Griffin baseline further exhibits inaccurate localization, missed detections, and false positives. These examples show that CETM improves both current-frame detection quality and temporal identity consistency. Additional examples are provided in Appendix B.7.

4.4 Ablation Studies

Components Ablation. Table 2 compares the native receiver with each module alone and with both modules enabled under the same frozen protocol. Spatial Reliability Calibration alone raises AP from 58.00 to 66.99, but changes AMOTA only from 54.06 to 54.51. Lineage-Guided Memory alone retains the native AP while raising AMOTA to 72.97 and reducing IDS from 595 to 29. Combining the modules yields 66.99 AP, the highest AMOTA (73.18), and the highest recall (79.33). The contrast separates SRC's contribution to detection quality from LGM's contribution to temporal continuity; the full configuration provides the strongest joint AP-AMOTA-recall result, although Memory-only has fewer identity switches and fragments.

Hard Geometry Gating. Table 3 tests whether lineage evidence remains reliable without the complete geometric gates. All three variants use identical SRC detections and retain the same association cost, solver, and lineage cues; only the gate definition changes. Without hard geometry, AMOTA falls to 67.19 and IDS rises to 65, compared with 73.18 and 39 for the full gate. A center-only gate reduces IDS to 58, but leaves AMOTA at 67.18 and FRAG at 37. Adding size, velocity, and yaw constraints raises recall from 73.08 to 79.33 and reduces FRAG to 16. Thus, lineage cues are most effective after implausible matches have been excluded geometrically.

Lineage Cue Ablation. Table 4 isolates source alias and cache parent while holding detections, thresholds, and geometric gates fixed. Without source alias, IDS rises from 39 to 210 on clean inputs and from 87 to 217 under 30% packet loss. Removing both lineage cues and relying on geometry alone further raises IDS to 253 and 241, respectively. Omitting cache parent slightly increases AMOTA, but raises clean-input IDS from 39 to 43 and FRAG from 16 to 24. These results identify source alias as the stronger continuity cue, while cache parent supplies a more conservative constraint. Neither cue replaces geometric feasibility: both are used only to distinguish candidate-track pairs that pass the gates.

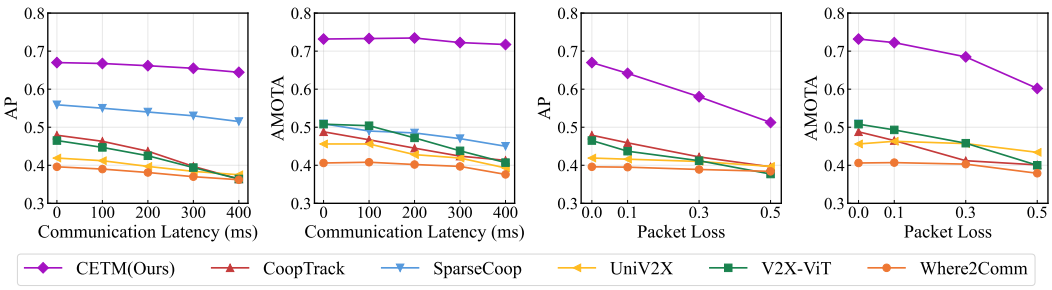

Figure 5: Robustness comparison on Griffin under communication delay and packet loss.

Table 4: Lineage-evidence ablation with frozen detections, thresholds, and geometric gates.

Condition	Variant	AMOTA ↑	IDS ↓	Recall ↑	FRAG ↓
Clean	Full	73.18	39	79.33	16
	w/o cache	73.40	43	79.65	24
	w/o alias	71.09	210	79.15	18
	Geometry	69.41	253	79.80	23
30% loss	Full	68.49	87	71.95	72
	w/o cache	69.06	87	70.83	74
	w/o alias	65.97	217	71.48	104
	Geometry	65.69	241	71.61	107

4.5 Robustness Experiments

Fig. 5 compares CETM with the Griffin baselines under packet loss and communication delay. The two disturbances affect different forms of cooperative evidence: packet loss removes observations, whereas delay makes them temporally misaligned. CETM remains above the displayed baselines in both AP and AMOTA across the evaluated disturbances.

Across delays of 100-400 ms (Table 8), CETM maintains AP above 64 and AMOTA above 71. At 400 ms, it obtains 64.44 AP and 71.74 AMOTA, compared with 66.99 AP and 73.18 AMOTA on clean inputs. Unlike missing packets, delayed observations remain available for validation against active tracks. LGM therefore prevents stale or inconsistent candidates from overwriting existing identities, producing a more gradual degradation under temporal asynchrony.

Under 10% packet loss (Table 9), CETM retains 64.16 AP and 72.24 AMOTA, while IDS remains close to the clean result (37 versus 39). At 30% loss, AP and AMOTA decrease to 58.01 and 68.49, and IDS rises to 87. Severe 50% loss further increases IDS to 257 and FRAG to 220. These results show that SRC and LGM tolerate moderate evidence loss, but continuity degrades when cooperative observations become too sparse to support reliable track updates.

5 Conclusion

This paper addresses insufficient spatial reliability of multi-source candidates and unstable cross-frame identities in aerial-ground cooperative perception by proposing the Cooperative Evidence Trust Management (CETM) framework. CETM consists of Spatial Reliability Calibration (SRC) and Lineage-Guided Memory (LGM), which respectively calibrate candidate ranking using cross-route spatial evidence and maintain stable scene-local identities by combining causal motion, hard geometric constraints, and lineage cues. Detection, tracking, and communication-perturbation experiments on the Griffin dataset show that CETM significantly improves detection accuracy and temporal tracking performance while maintaining robust performance under communication delay and packet loss; the results further show that SRC primarily improves candidate reliability ranking, whereas LGM primarily strengthens identity continuity.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing by improving the grammar, wording, and readability of the manuscript and providing critical feedback on the clarity of the methodology description and experimental presentation. The research ideas, methodology, implementation, experiments, data, and reported results were developed and produced by the authors. We reviewed, verified, and revised all AI-assisted textual edits. We take responsibility for the final content of this work, including its technical claims and experimental results.

REPRODUCIBILITY STATEMENT

To support reproducibility, Section 3 provides a detailed description of the overall methodological framework and each module, and Section 4 provides the execution details of the experiments as well as the datasets, data splits, and evaluation protocols used. Appendix A provides additional task and pipeline details, while Appendix B records sensitivity diagnostics, delay-level metrics, and qualitative diagnostics. Furthermore, after this paper is accepted, we will publicly release the source code and models.

REFERENCES

Eduardo Arnold, Mehrdad Dianati, Robert de Temple, and Saber Fallah. Cooperative perception for 3d object detection in driving scenarios using infrastructure sensors. *IEEE Transactions on Intelligent Transportation Systems*, 23(3):1852–1864, 2020.

Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pp. 3464–3468. Ieee, 2016.

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 *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11618–11628. IEEE, 2020.

Antoine Caillot, Safa Ouerghi, Pascal Vasseur, Rémi Boutteau, and Yohan Dupuis. Survey on cooperative perception in an automotive context. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):14204–14223, 2022.

Qi Chen, Sihai Tang, Qing Yang, and Song Fu. Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In *2019 IEEE 39th International Conference on distributed computing systems (ICDCS)*, pp. 514–524. IEEE, 2019.

Kexin Gong, Puyi Yao, Guiyang Luo, Quan Yuan, Tiange Fu, Hui Zhang, and Jinglin Li. From discriminative to generative: A diffusion-based paradigm for multi-agent collaborative perception. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4266–4274, 2026.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *Advances in neural information processing systems*, 35:4874–4886, 2022.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Zixing Lei, Zhenyang Ni, Ruize Han, Shuo Tang, Chen Feng, Siheng Chen, and Yanfeng Wang. Robust collaborative perception without external localization and clock devices. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7280–7286. IEEE, 2024.

Hangyu Li, Bofeng Cao, Zhaohui Liang, Wuzhen Li, Juyoung Oh, Yuxuan Chen, Shixiao Liang, Hang Zhou, Chengyuan Ma, Jiaxi Liu, et al. Cats-v2v: A real-world vehicle-to-vehicle cooperative perception dataset with complex adverse traffic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2294–2303, 2026.

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

Yiming Li, Dekun Ma, Ziyan An, Zixun Wang, Yiqi Zhong, Siheng Chen, and Chen Feng. V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. *IEEE robotics and automation letters*, 7(4):10914–10921, 2022.

Yen-Cheng Liu, Junjiao Tian, Nathaniel Glaser, and Zsolt Kira. When2com: Multi-agent perception via communication graph grouping. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4105–4114. IEEE, 2020a.

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

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. *arXiv preprint arXiv:2211.07214*, 2022.

Karthikeyan Chandra Sekaran, Markus Geisler, Dominik Rößle, Adithya Mohan, Daniel Cremers, Wolfgang Utschick, Michael Botsch, Werner Huber, and Torsten Schön. Urbaning-v2x: A large-scale multi-vehicle, multi-infrastructure dataset across multiple intersections for cooperative perception. *Sensors (I— V)*, 10:16, 2025.

Bowen Wang, Yafei Wang, Wei Gong, Siheng Chen, Genjia Liu, Minhao Xiong, and Chin Long Ng. V2xscenes: A multiple challenging traffic conditions dataset for large-range vehicle-infrastructure collaborative perception. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28385–28395. IEEE, 2025.

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, et al. 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, et al. Long-scope: Fully sparse long-range cooperative 3d perception. *arXiv preprint arXiv:2604.09206*, 2026c.

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*, pp. 605–621. Springer, 2020.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 3645–3649. IEEE, September 2017. doi: 10.1109/ICIP.2017.8296962.

Junhao Xu, Yanan Zhang, Zhi Cai, and Di Huang. Cosdh: Communication-efficient collaborative perception via supply-demand awareness and intermediate-late hybridization. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6834–6843. IEEE, 2025.

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. Springer, 2022a.

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

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*, pp. 989–1000. PMLR, 2023.

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

Jinqing Zhang, Yanan Zhang, Qingjie Liu, and Yunhong Wang. Sa-bev: Generating semantic-aware bird's-eye-view feature for multi-view 3d object detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3325–3334. IEEE, 2023.

Yanan Zhang, Jiaxin Chen, and Di Huang. Cat-det: Contrastively augmented transformer for multimodal 3d object detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 898–907. IEEE, 2022a.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, pp. 1–21. Springer, 2022b.

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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 26954–26965. IEEE, 2025.

Chao Zhou, Yanan Zhang, Jiaxin Chen, and Di Huang. Octr: Octree-based transformer for 3d object detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5166–5175. IEEE, 2023.

A Additional Method Details

A.1 Additional SRC Details

Route-10 remains the primary candidate stream, whereas Route-5 supplies supporting evidence only and never contributes independent detections. SRC therefore changes candidate reliability scores without expanding or pruning the candidate set. For each Route-10 candidate, deterministic same-class one-to-one matching checks center, size, and velocity consistency; unsupported candidates retain their native scores and all candidate metadata. The quality estimator uses the resulting spatial descriptor and is frozen before the official evaluation. Its out-of-fold training and cross-route support coverage are detailed in the SRC diagnostic subsection.

A.2 Additional LGM Details

Each scene-local track stores a stable ID, the latest box, velocity, timestamp, lifecycle state, and accumulated native and source aliases. A source alias identifies a source agent and source-instance lineage, whereas a cache parent records inherited query continuity; neither is treated as a physical object ID. LGM considers these cues only within the hard-geometry-compatible pair set. Cache-parent proposals precede source-alias proposals, and committed proposals lock both endpoints before residual one-to-one matching. Pairs that fail the cost or competition conditions remain unmatched; valid unmatched candidates may initialize tracks, while unmatched histories persist only for the fixed lifecycle horizon.

A.3 Receiver-Side Algorithm

Algorithm 1 CETM Evidence-to-Identity Management

Require: Frozen cooperative backbone $\mathcal{B}$, quality estimator Q, fixed gates and association parameters Θ, chronological observations $\{\mathcal{O}_t\}_{t=1}^{T}$
Ensure: Calibrated detections with stable scene-local tracking IDs
Initialize scene-local memory $\mathcal{M} \leftarrow \emptyset$
for each frame t in chronological order **do**
  Obtain and ego-align Route-10 candidates $\mathcal{D}_t^{10}$ and Route-5 candidates $\mathcal{D}_t^{5}$ using $\mathcal{B}$
  Find Route-5 support for each Route-10 candidate using center, size, and velocity consistency
  Compute $s_i^{\mathrm{cons},t}$ and $\widehat{s}_i^{t}$ using Eqs. 3–5, while preserving lineage metadata
  Causally predict active tracks in $\mathcal{M}$ using Eq. 6
  Form candidate–track pairs that pass the hard geometric gates
  Form cache-parent and source-alias priority proposals within the eligible pairs
  Commit non-conflicting priority proposals and lock their endpoints
  Compute association costs and obtain Hungarian matches for the remaining pairs under Θ
  Reject residual pairs that exceed the cost threshold or fail the two-sided margin
  Update matched tracks, initialize valid unmatched candidates, and retain unmatched tracks within their finite lifecycle
  Output calibrated detections with stable scene-local tracking IDs
  if the next frame belongs to a new scene **then**
    $\mathcal{M} \leftarrow \emptyset$
  end if
end for

A.4 Implementation Details and Fixed Parameters

SRC accepts cross-route support only when candidate centers, 3D box sizes, and velocity states satisfy fixed consistency conditions. Its score blending weight is fixed before evaluation. For LGM, hard gates are applied to predicted-center residual, box-size ratio, velocity change, and yaw difference before any lineage cue is considered. The association cost combines normalized center, size, velocity, and heading differences using the fixed weights in Eq. 9.

All gates, score-blending settings, association margins, and unmatched-track lifetimes are fixed throughout the reported evaluation. Source alias and cache parent act within geometrically feasible relations; proposals failing the gate and residual pairs failing the cost or competition margin remain unmatched.

B Additional Experimental Details

B.1 Reproducibility Protocol

All reported results use the frozen SparseCoop checkpoint, fixed dual-route settings, fixed SRC manifest, and fixed association parameters on the Griffin official 10-scene, 0–50 m Car protocol. The clean run contains 1,490 chronological frames. The receiver processes one frame at a time, commits memory before the next frame, resets state at scene boundaries, and does not read validation annotations during inference or parameter selection. The strict online runtime is measured on one NVIDIA RTX 3090 and includes both detector routes and receiver-side processing. The "Native" value in Table 2 is the locally reproduced receiver-side reference under this frozen protocol.

B.2 Hyperparameter Sensitivity

The default $(\alpha, g, m, c) = (0.25, 2, 0.05, 1.0)$ is the operating configuration used throughout the main evaluation. Table 6 characterizes nearby settings on the official 10-scene split. The first five rows sweep α, whereas the final three rows change one non-default setting while retaining the other defaults. Stricter margins increase AMOTA to 73.61 but increase IDS to 47, while longer memory reaches 73.38 AMOTA with 32 IDS and slightly lower recall. The default retains the highest reported AP and near-maximal recall while using the shorter memory horizon.

Table 5: Fixed receiver-side implementation details used in all reported evaluations.

Component	Fixed setting
Hardware and software	NVIDIA RTX 3090 (24,576 MiB), Linux 5.15, Python 3.8.20, PyTorch 1.13.1+cu117, CUDA 11.7, MMCV 1.7.2, NumPy 1.19.5, and SciPy 1.10.1.
Cooperative backbone	Frozen SparseCoop `iter_2064.pth`; deterministic deformable attention; batch size one; Route-10 supplies final candidate geometry and Route-5 supplies supporting evidence.
SRC	Same-class deterministic one-to-one support matching: center distance ≤ 1.0 m, maximum size ratio ≤ 2.0, and velocity distance ≤ 4.0 m/s. A frozen OOF histogram-gradient-boosting quality ensemble is blended with consensus using exponents 0.25 and 0.75.
LGM prediction and gates	Causal XY prediction $\hat{c}_t = c_{\text{last}} + v_{\text{last}}\Delta t$; center residual $\leq 2.0 + 2.0\Delta t$ m, maximum per-axis size ratio ≤ 2.0, velocity residual ≤ 10.0 m/s, and moving-object yaw difference $\leq \pi/2$.
Association and lifecycle	Priority lineage commitment followed by Hungarian assignment of remaining pairs; center/log-size/velocity/yaw weights 0.45/0.20/0.20/0.15; cost ≤ 1.0; two-sided margin ≥ 0.05; `max_gap_frames=2`; confirmation after two observations; reset at scene boundaries.
Timing scope	Both detector routes, receiver-side calibration, lineage resolution, association, scene management, data loading, and Python orchestration; 968.707 ms/frame, 918.355 MiB peak CUDA allocated, and 1,188 MiB peak CUDA reserved.

Table 6: Sensitivity analysis on the frozen official-10 protocol.

α	g	m	c	AP ↑	AMOTA ↑	IDS ↓	Recall ↑
0.000	2	0.05	1.0	66.08	72.58	19	72.76
0.125	2	0.05	1.0	66.83	73.25	35	78.35
0.250	2	0.05	1.0	66.99	73.18	39	79.33
0.375	2	0.05	1.0	66.93	73.34	49	79.83
0.500	2	0.05	1.0	66.77	73.28	30	76.29
0.250	5	0.05	1.0	66.99	73.38	32	79.31
0.250	2	0.10	1.0	66.99	73.61	47	79.27
0.250	2	0.05	1.5	66.99	73.18	39	79.33

B.3 SRC Cross-Route Evidence Diagnostic

We diagnose the evidence available to SRC using five-fold out-of-fold (OOF) predictions over 37 training scenes. Each Town-balanced fold holds out seven or eight scenes and fits the quality estimator on the remaining 29 or 30 scenes; the held-out candidate scores are therefore produced by a model that did not train on their scenes. The diagnostic contains 607,225 active Route-10 Car candidates from 5,510 frames. "Active" means that the candidate center lies within the receiver-centric 0–50 m range; no detector-score cutoff is applied. Train-scene ground truth supervises OOF quality-model fitting and labels the post-hoc diagnostic, but no ground truth is read during SRC inference or used to select the fixed gates, distance bins, or official-10 validation configuration.

For each Route-10 anchor, valid support requires a same-class Route-5 candidate under deterministic one-to-one matching, with center distance ≤ 1.0 m, maximum size ratio ≤ 2.0, and velocity distance ≤ 4.0 m/s. Table 7 reports the availability of this support by receiver-centric range. Coverage increases from 88.46% at 0–10 m to 93.21% at 40–50 m, showing that Route-5 provides broadly available supporting evidence across the evaluated range.

Table 7: Route-5 support coverage for active Route-10 Car candidates by receiver-centric range in the SRC OOF diagnostic.

Range	Candidates	Support coverage
0–10 m	59,881	88.46%
10–20 m	74,329	88.96%
20–30 m	109,838	91.45%
30–40 m	167,835	92.66%
40–50 m	195,342	93.21%

B.4 Delay Metrics

Table 8 gives the CETM results underlying the communication-delay panels in Fig. 5. Each delay condition uses its own valid chronological frame set with frozen weights and thresholds. The 400 ms condition contains 1,450 valid frames and yields AP 64.44, AMOTA 71.74, 33 IDS, 79.01 recall, and 34 fragments.

Table 8: Full CETM results under delayed cooperative inputs.

Delay	Frames	AP ↑	AMOTA ↑	IDS ↓	Recall ↑	FRAG ↓
100 ms	1480	66.75	73.31	53	80.24	19
200 ms	1470	66.17	73.45	41	79.60	18
300 ms	1460	65.49	72.25	43	79.71	28
400 ms	1450	64.44	71.74	33	79.01	34

B.5 Packet-Loss Metrics

Table 9 reports the CETM results shown in the packet-loss panels of Fig. 5. The packet-loss conditions use the same token-keyed corruption protocol with frozen detector weights and receiver-side parameters.

Table 9: Full CETM results under token-keyed packet loss.

Packet loss	AP ↑	AMOTA ↑	IDS ↓	Recall ↑	FRAG ↓
0%	66.99	73.18	39	79.33	16
10%	64.16	72.24	37	77.49	24
30%	58.01	68.49	87	71.95	72
50%	51.25	60.20	257	68.87	220

B.6 Communication Accounting

CETM retains the SparseCoop inter-agent communication protocol. Cooperative agents transmit the same kinematic queries under the same route settings. SRC and LGM execute solely at the receiver, using received candidate attributes and locally maintained scene memory. They do not transmit candidate packets, track states, or additional lineage messages. Consequently, the full CETM configuration retains the 9.73×10^4 BPS cooperative payload reported for SparseCoop in Table 1.

B.7 More Visualizations

To complement the main visualization in Fig. 4, we provide additional detection and tracking examples under fixed scenes and display settings. The detection views Fig. 6 highlight localization and coverage differences, whereas the tracking views Fig. 7 show how candidate changes propagate to cross-frame identities. These examples are qualitative diagnostics only; all quantitative claims are evaluated using the metrics in the main text.

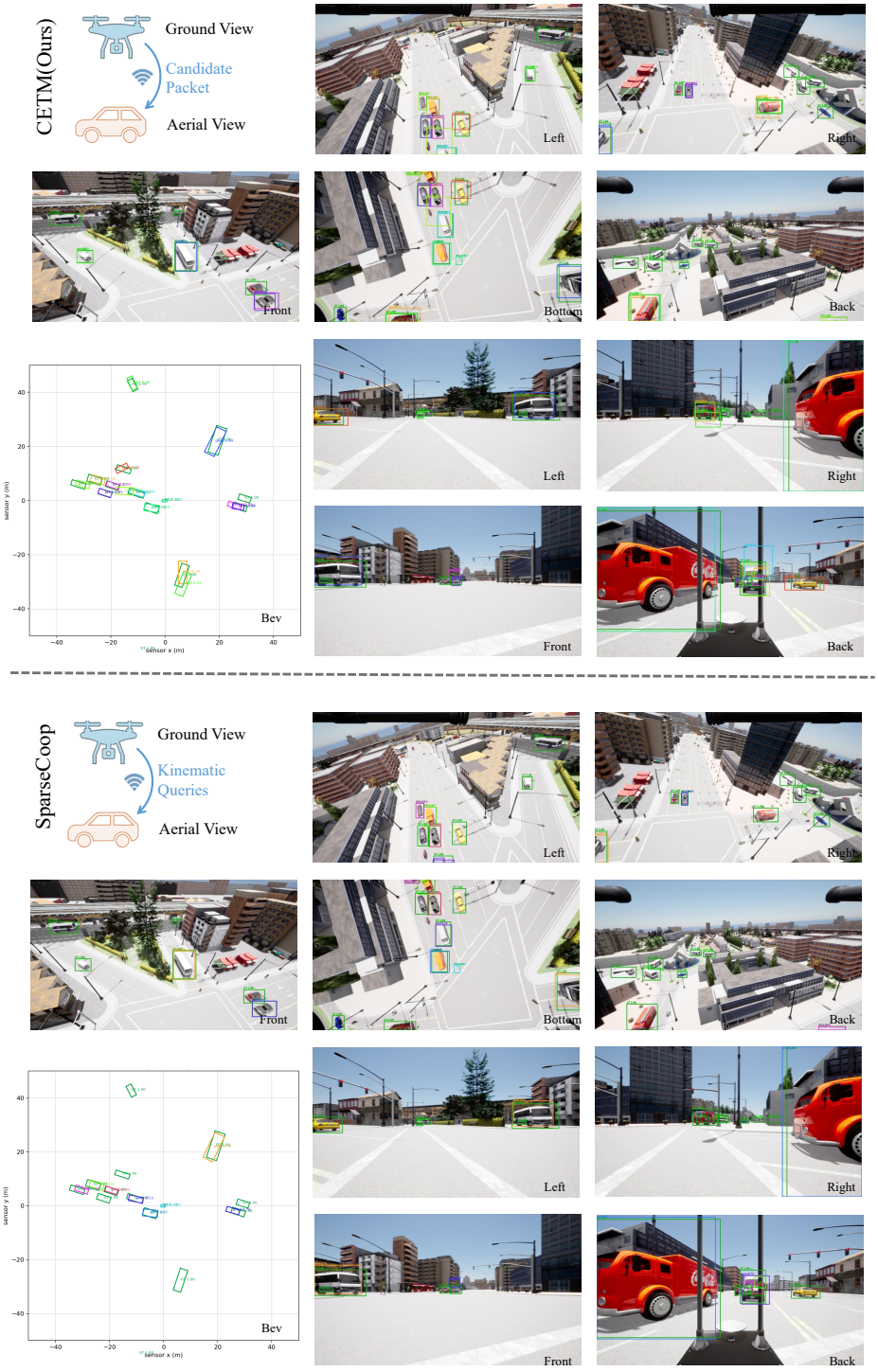

Figure 6: Detection visualization comparing CETM with SparseCoop.



Figure 7: Consecutive-frame tracking visualization comparing CETM with SparseCoop.