COOPERTRACE: COOPERATIVE MULTI-MODAL END-TO-END MOTION PREDICTION VIA V2X COMMUNICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern autonomous driving stacks transition from modular pipelines to end-to-end learning frameworks. Equipped with diverse sensing modalities, autonomous vehicles are increasingly robust to sensor degradation and long-tail events. In the meantime, leveraging Vehicle-to-everything (V2X) communication, recent cooperative autonomy research also expands from perception to downstream tasks, demonstrating the benefits of situational awareness beyond individual sensing range. However, it is unclear how multi-modal sensing would interact with the end-to-end learning paradigm in the V2X setting. In this paper, we present COOPERTRACE, a *cooperative, multi-modal, and end-to-end* motion prediction framework to explore the feasibility and efficacy of this three-fold combination. COOPERTRACE fuses LiDAR and camera at every agent, attends to the features of connected agents, and trains the forecaster on tracking results from previous frames. Extensive evaluations on both simulated and real-world cooperative datasets show that COOPERTRACE achieves up to 68% lower motion prediction error, 134% higher MOTA, and 18% higher 3D detection AP than the compared state-of-the-art methods. We further stress-test its cooperative gains under emulated sensor degradation. COOPERTRACE's multi-modal multi-agent fusion maintains stable performance across these settings and degrades gracefully when one modality becomes unavailable, demonstrating its potential for harsh real-world deployment conditions.

1 INTRODUCTION

To navigate complex and dynamic environments safely, autonomous vehicles require precise motion forecasting to predict the future trajectories of surrounding agents. Yet the observations available to the forecaster are often incomplete or unreliable. Occlusions and limited sensing range restrict what a single vehicle can observe, while individual sensors may degrade under adverse conditions. V2X cooperation and multi-modal fusion address these issues in complementary ways: cooperation provides observations from additional viewpoints, while multi-modal sensing provides complementary information from different sensor modalities. However, even when these additional observations can improve perception, errors in perception and tracking may still propagate into downstream forecasting in modular pipelines, motivating end-to-end optimization of these stages. Existing end-to-end motion forecasting methods exploit V2X cooperation and multi-modal sensing separately: multi-modal approaches remain single-agent (Ye et al., 2023), while cooperative end-to-end approaches remain single-modality (Yu et al., 2025; Song et al., 2026). Thus, as summarized in Table 1, their integration within an end-to-end motion forecasting framework remains unexplored.

Table 1: Comparison across cooperative and end-to-end driving models. K is the number of forecast modes.

Method	V2X	Multi-modal	E2E	Track	Forecast
FaF	×	×	✓	✓	✓
UniAD	×	×	✓	✓	✓
FusionAD	×	✓	✓	✓	✓
CMP	✓	×	×	✓	✓
V2XPnP	✓	×	×	×	K=1
UniV2X	✓	×	✓	✓	✓
UniMM-V2X	✓	×	✓	✓	✓
CoopTrack	✓	×	✓	✓	×
Ours	✓	✓	✓	✓	K=6

Further, even with cooperative and multi-modal observations fused end-to-end, the forecaster still depends on a history produced by its own tracker, and that history carries coasted frames, localization

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.

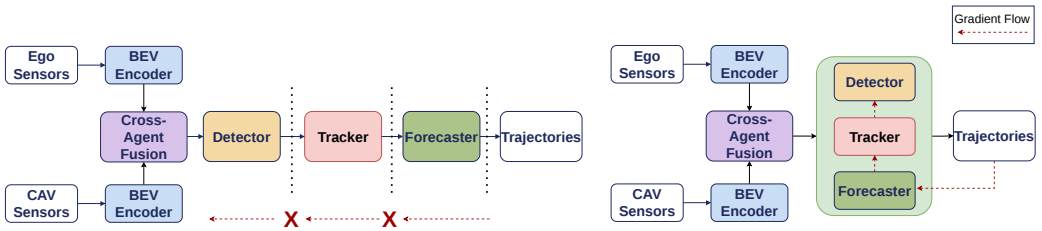

Figure 1: Example illustration of modular (*e.g.,* CMP (Wang et al., 2025)) and end-to-end (*e.g.,* UniV2X (Yu et al., 2025), UniMM (Song et al., 2026)) cooperative pipelines. Left: A modular cooperative pipeline passes boxes between detector, tracker, and forecaster, so no gradients are back-propagated. Right: single end-to-end graph from multi-agent input to trajectories.

drift and fragmented identities. Existing end-to-end systems keep no explicit trajectory: the history lives implicitly inside the track query, shaped by an auxiliary loss that is supervised by annotation, so the forecaster never sees the positions the tracker actually produced and the query can only be taught the annotated past. Modular pipelines do pass an explicit trajectory, but train the forecaster only on the tracks their tracker completes. In neither case does the forecaster learn from the histories it will actually receive in deployment.

Also, robust multi-modal forecasting requires that each sensing modality remain useful when another becomes unreliable. If the camera branch depends on LiDAR-derived inputs or features at inference, degradation of the LiDAR sensor can also impair camera-based predictions, reducing the redundancy provided by multi-modal sensing.

Motivated by these challenges, we introduce CooperTrace, an end-to-end cooperative motion forecasting framework that integrates both V2X cooperation and LiDAR-camera fusion. CooperTrace unifies detection, tracking, and forecasting within a differentiable framework and trains the forecaster on tracker-generated histories to match deployment conditions. To preserve cross-modal redundancy, LiDAR is used to supervise camera depth estimation during training but is not required by the camera branch at inference. Figure 1 contrasts an illustrative modular cooperative perception–tracking–forecasting pipeline with an end-to-end pipeline, which maintains a differentiable path from cooperative multi-modal observations to tracked representations and trajectory forecasts.

Our contributions are summarized as follows:

- **Cooperative multi-modal E2E forecasting.** To our knowledge, CooperTrace is the first framework to integrate V2X cooperation and LiDAR-camera fusion in an end-to-end detection, tracking, and trajectory forecasting system trained as one differentiable graph.
- **Closed-loop learning with explicit cooperative tracking histories.** In addition to propagated track queries, CooperTrace feeds an explicit one-second cooperative tracking history into motion prediction heads, enhancing generalizability to unknown situations at inference time.
- **Balancing modality independence and cross-modal fusion.** CooperTrace exploits projected LiDAR as depth supervision for the camera only during training to enhance sensor fusion, but keeps the camera branch independent at inference time to minimize the impact of sensor degradation.

We evaluate CooperTrace on both simulated (OPV2V (Xu et al., 2022c)) and real-world (V2X-Seq (Yu et al., 2023) and CooperScene (Wu et al., 2026)) datasets under a common deployment protocol (Section 4.1). Compared to CMP (Wang et al., 2025) and UniV2X (Yu et al., 2025), CooperTrace reduces minADE@5s by 26% and 62% on OPV2V, respectively. On V2X-Seq, where UniV2X operates at a different temporal sampling rate, endpoint errors are directly comparable at matched horizons; CooperTrace reduces minFDE@5s by 66% relative to UniV2X. Relative to the non-converged CMP† baseline on V2X-Seq, the minADE@5s reduction is 77%. On CooperScene, CooperTrace attains 0.964 m minADE@5s versus 0.870 m for CMP, while reducing the error by 68% relative to UniV2X. Across all three datasets, CooperTrace achieves 114% higher mean forecasting coverage than the mean coverage of CMP and UniV2X. Ablation studies on OPV2V show that training the forecaster on cooperative tracking histories lowers minADE@5s from 2.01 m to 1.78 m. Under severe LiDAR degradation (beams reduced from 64 to 8), a LiDAR-only cooperative agent forecasts 16.0 % of eligible vehicles, while CooperTrace forecasts 61.7 %.

2 Related work

Cooperative and end-to-end autonomy. Cooperative methods extend sensing through shared representations, including Bird-eye-view (BEV) fusion (Wang et al., 2020; Xu et al., 2022b;a), communication-efficient fusion (Hu et al., 2022), and pose-robust fusion (Lu et al., 2023). In parallel, end-to-end systems jointly learn perception, tracking, and prediction from single-agent LiDAR (Luo et al., 2018; Liang et al., 2020), cameras (Gu et al., 2023; Hu et al., 2023), or LiDAR-camera fusion (Ye et al., 2023). Recent work combines these directions: CMP (Wang et al., 2025) forecasts from cooperative detections through a non-differentiable Kalman-filter tracker (Weng et al., 2020); V2XPnP (Zhou et al., 2025) uses staged LiDAR fusion for single-trajectory prediction; UniV2X (Yu et al., 2025) and UniMM-V2X (Song et al., 2026) jointly model cooperative camera perception and prediction; and CoopTrack (Zhong et al., 2025) extends end-to-end cooperation to tracking without forecasting. Cooperative forecasting has also been studied with ground-truth tracks, as in V2X-Graph (Ruan et al., 2024) and Co-MTP (Zhang et al., 2025), so their forecasters receive cleaner track histories than a deployed tracker would provide. To our knowledge, CooperTrace is the first framework to integrate V2X cooperation and LiDAR-camera fusion in an end-to-end learning pipeline.

Multi-modal fusion and sensor degradation. LiDAR and cameras exhibit complementary failure modes: LiDAR degrades with range and adverse weather, whereas cameras retain dense appearance cues but require depth estimation. BEVFusion (Liu et al., 2023) fuses both modalities in BEV space, TransFusion (Bai et al., 2022) combines them in a query-based detector, and FusionAD (Ye et al., 2023) extends LiDAR-camera fusion to end-to-end prediction. The camera depth formulation is critical to preserving redundancy: using projected LiDAR as an inference-time input couples the camera representation to LiDAR quality, whereas camera-only lifting (Philion & Fidler, 2020) with LiDAR-based depth supervision during training (Li et al., 2023) preserves an independent camera branch at inference. Robustness to sensor degradation has primarily been studied at the detection level under corruptions such as adverse weather, LiDAR sparsification, camera degradation, and sensor misalignment (Beemelmanns et al., 2024). We instead study whether these effects persist through perception, tracking, and forecasting, and whether cross-agent and cross-modal redundancy complement or substitute for one another (Section 4.3).

3 Method

3.1 Overview

CooperTrace is a cooperative multi-modal end-to-end motion prediction framework, which maps the raw sensor streams of several connected vehicles to vehicles' future trajectories inside one differentiable graph. Figure 2 shows an overview of CooperTrace. Each agent encodes its LiDAR and camera data into BEV representations; partner BEVs are warped into the ego frame and fused by cross-agent attention, so information from different agents and modalities meets in one tensor. Object queries decoded from the fused BEV are carried to the next frame as track queries without detachment, which allows the forecasting loss to reach the upstream features. The motion head learns from a combination of BEV features, track queries, and cooperative tracking histories.

We consider an ego vehicle and up to $N_{\max} - 1$ connected partners within communication range. Each agent provides LiDAR sweeps at 10 Hz, camera images and an estimated pose. From these observations the model detects and tracks the surrounding vehicles and forecasts their motion. For each tracked vehicle it predicts $K = 6$ future trajectories over 5 s, 50 steps at 10 Hz. We write *multi-modal sensing* for the use of LiDAR and camera together and *multimodal futures* for the K trajectory hypotheses. At deployment, the history read by the forecaster comes from the system's own tracks, never from annotations; the resulting coverage-aware protocol is given in Section 4.1.

In the following subsections, we describe the perception (Section 3.2), tracking (Section 3.3), and forecasting (Section 3.4) modules, followed by the training procedure (Section 3.5).

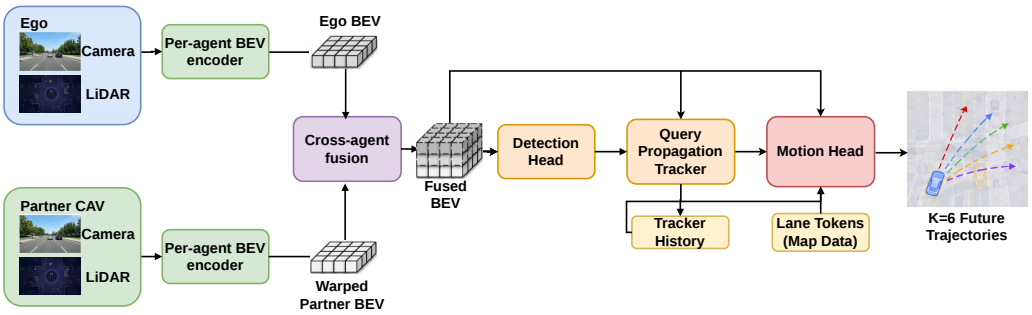

Figure 2: COOPERTRACE Overview. Per-agent LiDAR-camera BEVs are fused across agents in the ego frame; detection queries propagate across frames as track queries, and the motion head decodes K=6 futures from each query, its tracker-generated history and the fused BEV.

3.2 COOPERATIVE MULTI-MODAL PERCEPTION

Each agent encodes its own sensors in its local BEV frame. LiDAR goes through voxelization, a 3D sparse encoder and a 2D BEV backbone; images go through a Swin backbone and are lifted to BEV in the lift-splat form used by BEVFusion (Liu et al., 2023). A convolutional fuser combines the two into one BEV map $\mathbf{B}_i$. Every partner transmits its BEV feature to the ego vehicle, where it is warped into the ego frame with the relative pose, and the ego and warped partner maps are stacked along an agent axis. Alternating local-window and global attention then exchanges information across agents,

$$\mathbf{F} = \mathrm{Fuse}\left(\mathbf{B}_0,\, \mathcal{W}_{1\to 0}(\mathbf{B}_1),\, \ldots,\, \mathcal{W}_{N\to 0}(\mathbf{B}_N)\right),$$

where $\mathbf{B}_0$ is the ego feature and $\mathcal{W}_{i\to 0}$ the warp of partner i into the ego frame. The BEV map is the only thing exchanged between agents; The bandwidth is given in Appendix J. A TransFusion (Bai et al., 2022) head with 200 object queries of dimension 128 decodes detections from $\mathbf{F}$.

Camera depth. The value of a second modality depends on whether it remains functional when the primary one degrades. Some multi-modal detectors inject projected LiDAR depth into the camera pathway, creating an inference-time dependence between the branches. In contrast, BEVDepth-style methods (Li et al., 2023) use projected LiDAR only to supervise camera depth during training. We adopt this formulation within a fused detector: LiDAR provides a training target for the camera depth distribution but is not an inference-time input, making the camera branch structurally independent of LiDAR. Section 4.3 evaluates the benefit of this independence under LiDAR degradation.

3.3 QUERY-PROPAGATION TRACKING

Detection queries enter a track-query decoder of self-attention, cross-attention to the fused BEV and a feed-forward network. Queries matched to an object are carried to the next frame as a live tensor,

$$\mathbf{q}_t^{trk} = \mathrm{Dec}(\mathbf{q}_{t-1}^{trk}, \mathbf{F}_t),$$

and are not detached between frames, so gradient crosses the whole clip and identity is represented by the query vector itself rather than by an assignment. During training a clip-level matcher assigns ground-truth identities to the propagated queries, and the tracking loss

$$\mathcal{L}_{trk} = \mathcal{L}_{centre} + \mathcal{L}_{exist}$$

is added to the detection loss $\mathcal{L}_{det}$, where $\mathcal{L}_{centre}$ is the L_1 error between the centre a track query regresses and the centre of its assigned ground-truth object, and $\mathcal{L}_{exist}$ is a binary cross-entropy on each query's existence score. At inference each frame's boxes are taken from the detection head, keeping those with confidence above threshold, while the identity of each box is the propagated query that decoded it. The tracker fixes the input distribution of the forecaster: the propagated queries decide which vehicles receive a history at all and how noisy that history is. Appendix F compares this tracker with a modular hard-matching tracker on both metrics.

3.4 MOTION FORECASTING

For every track the motion head decodes $K = 6$ trajectories of 50 steps at 10 Hz and a probability for each. Existing query-based end-to-end systems pass the track query and the current box to the forecaster and integrate the history into the query, while modular pipelines pass a trajectory only. COOPERTRACE passes both: the query keeps the forecasting loss connected to detection and fusion, and an explicit history carries the kinematics the query does not. Before decoding, every input is rotated into the track's own heading frame, with the vehicle's current direction of travel along the x-axis, so that a maneuver such as lane following looks the same whichever way the vehicle is facing and the head learns one pattern per maneuver rather than one per heading; the decoded trajectories are rotated back into the ego frame afterwards. The head then reads four inputs: two of track itself, its query and its history, and two that describe the scene around it, the fused BEV and the lanes.

- **Track query.** The propagated query $\mathbf{q}_t^{trk}$ is used directly as the per-agent context, so the forecaster reads the same representation that carries identity.
- **Tracker-generated history.** The last 11 tracker positions are differenced into ten displacements, clamped at 3 m per step to bound the effect of a tracker outlier, and encoded with a GRU. During training these histories come from the deployed tracker rather than from annotations (Section 3.5).
- **Scene context.** Mode queries cross-attend to the fused BEV $\mathbf{F}_t$ around the current track position, with learned encodings of the sampled coordinates.
- **Lane context.** When a map is available, the eight nearest lane centrelines are encoded pointwise as polylines by a shared MLP and attended to by the mode queries. Missing lane sets are fully masked, so the same head runs on map-free datasets. Table 6 ablates this input and Appendix E compares polylines with the BEV lane raster used by the CMP family.

Each mode starts from a learned anchor embedding, and the anchor trajectories are obtained by k-means over training futures. Let $\hat{\boldsymbol{\tau}}_k \in \mathbb{R}^{50\times 2}$ be the trajectory decoded for mode k, $p \in \mathbb{R}^K$ the predicted mode probabilities, and $\boldsymbol{\tau}^*$ the ground-truth future of the track. The objective is the usual best-of-K loss,

$$\mathcal{L}_{mot} = \left\|\hat{\boldsymbol{\tau}}_{k^*} - \boldsymbol{\tau}^*\right\|_{\text{smooth-}L_1} + \tfrac{1}{2}\,\mathrm{CE}(p, k^*) + \lambda\,\mathcal{L}_{lane}, \qquad k^* = \arg\min_k \left\|\hat{\boldsymbol{\tau}}_k - \boldsymbol{\tau}^*\right\|, \quad (1)$$

where k^* is the mode closest to the ground truth: only that mode is regressed, the cross-entropy term teaches p to identify it, and $\mathcal{L}_{lane}$ pulls it toward the nearest lane where a map exists ($\lambda = 0.5$; $\lambda = 0$ on map-free datasets).

Coverage. A trajectory is produced only for tracks confirmed by at least three detections, matched within the last three frames, and holding a complete 11-position history, that is the current frame and the ten before it, one second at 10 Hz. A vehicle the tracker drops receives no forecast at all. We report *coverage*, the fraction of eligible vehicles that receive a forecast, next to every error, and Section 4.3 reads robustness mainly on this axis.

3.5 TRAINING

Stages. Training has four stages: the detector; the tracker with the detector frozen; the motion head with the tracker frozen; and an optional joint stage that fine-tunes everything above the frozen per-agent trunk with one backward pass per clip, under

$$\mathcal{L} = \mathcal{L}_{det} + \mathcal{L}_{trk} + \mathcal{L}_{mot}, \quad (2)$$

with $\mathcal{L}_{mot}$ from Equation 1 and learning-rate multipliers of 0.05 for the fusion blocks and detection head and 0.1 for the tracker. Staging is an optimization schedule, not a modular pipeline: every stage exchanges live tensors, no interface is frozen into decoded boxes, and a backward pass from the forecasting loss alone reaches the detector trunk with gradients, across frames as well as within them (Appendix C). At deployment, forecasting histories contain the drift, coasting, and fragmentation produced by the tracker; we therefore train the motion head on tracker-generated histories rather than annotations, using ground truth only to associate each history with its future target. Section 4.4 compares alternative history sources.

Initialisation. On OPV2V and CooperScene the camera model starts from its LiDAR parent: the LiDAR trunk and detection head are frozen and the camera fuser is zero-initialised, so step zero

is exactly the LiDAR model and the camera's contribution is attributable. For V2X-Seq, we train the multi-modal configuration without this LiDAR-parent initialisation. During joint training the 7.4 M-parameter per-agent trunk stays frozen and everything above it is trained; the memory budget behind this choice is in Appendix C.

4 EXPERIMENTS

4.1 SETUP

Datasets. We evaluate motion prediction on three cooperative datasets: OPV2V (Xu et al., 2022c), CooperScene (Wu et al., 2026), and V2X-Seq (Yu et al., 2023), spanning simulated and real-world settings. We use three cooperating agents by default where supported and five in the scaling ablation; V2X-Seq uses its native vehicle–infrastructure pair.

Baselines. For motion prediction, the baselines include CMP (Wang et al., 2025) and UniV2X (Yu et al., 2025). For detection, we compare with CoBEVT (Xu et al., 2022a), ERMVP (Zhang et al., 2024), and CoSDH (Xu et al., 2025). Additional datasets and baselines are reported in Appendix A.

Metrics and Evaluation Protocol. Ground truth is the union of all agent labels within the ego vehicle's detection range. A prediction is eligible when its ground-truth box exists and a complete 5 s future is available. Scoring further requires at least three tracker detections, a match within the last three frames, and localization within 2 m of ground truth. The motion head receives an 11-frame position history (1 s at 10 Hz), represented as ten displacements.

For motion prediction, we report *minADE* and *minFDE* at 1, 3, and 5 s with $K{=}6$, together with *coverage* and *EPA* (Gu et al., 2023). Coverage is the fraction of eligible agents that receive a forecast. Following ViP3D, EPA penalizes false positive forecasts and is computed over our eligible population as $(N_{\text{hit}} - 0.5N_{\text{FP}})/N_{\text{eligible}}$. Predicted and ground-truth agents are matched using a 2 m center-distance threshold at the current frame. A matched prediction is counted as a hit if its minimum final displacement error at 5 s is within 2 m, while unmatched forecast tracks are counted as false positives. For tracking, we follow the AB3DMOT (Weng et al., 2020) protocol and CLEAR-MOT (Bernardin & Stiefelhagen, 2008), using a 2 m center-distance gate consistent with the forecaster's position-based input. For detection, we report AP_{BEV} and AP_{3D} at IoU thresholds of 0.3, 0.5, and 0.7. The detection evaluation ranges are $\pm 140.8 \times \pm 40$ m for OPV2V, $\pm 72 \times \pm 72$ m for CooperScene, and $\pm 100.8 \times \pm 40$ m for V2X-Seq.

Unified Evaluation. Published cooperative-prediction results are not directly comparable due to different evaluation protocols. CMP filters detections with ground truth before tracking, removing false detections and identity-recovery errors. UniV2X reports a 0.69 accuracy score despite a 2.86 m absolute error. To level the ground, therefore, we re-evaluate all baselines under a unified protocol and validate each conversion against published results (see details in Appendix G).

4.2 END-TO-END RESULTS

Motion prediction. Table 2 compares three cooperative motion prediction systems. On OPV2V, COOPERTRACE reduces minADE@5s by 26 % relative to CMP while forecasting 18 percentage points more of the eligible population (EPA 0.409 vs. 0.265). On V2X-Seq, it reduces minADE@5s by 67 % relative to UniV2X at 2.5× higher coverage. On CooperScene, CMP attains lower average error but over only 22 % of eligible agents, compared with 55 % for COOPERTRACE; EPA accounts for this population difference and favors COOPERTRACE, 0.338 vs. 0.190. For a fair comparison, OPV2V uses CMP's released checkpoint and UniV2X is retrained following its official four-stage pipeline; CooperScene uses our retrained CMP; and V2X-Seq uses UniV2X's released 2 Hz checkpoint, so endpoint errors are directly comparable at matched horizons, but averaged errors are not due to different temporal sampling. Training details are in Appendix I.

Effect of cooperation. Cooperation is most directly reflected in forecasting coverage, making coverage a key metric for assessing its benefit. The additional agents contributed by a partner are predominantly occluded or distant vehicles that are not observable from the ego vehicle alone (median distance 60 m from the ego; Appendix H); consequently, agents that are not retained by the tracking

Table 2: Cooperative Motion Prediction Results

Dataset	Method	minADE↓ @1s	minADE↓ @3s	minADE↓ @5s	minFDE↓ @1s	minFDE↓ @3s	minFDE↓ @5s	Cov.↑	EPA↑
OPV2V (Xu et al., 2022c)	CMP (Wang et al., 2025)	0.327	0.941	1.833	0.501	1.840	3.982	0.711	0.265
	UniV2X (Yu et al., 2025)	1.301	2.412	3.562	1.585	3.603	5.897	0.453	−0.165
	CooperTrace (Ours)	**0.186**	**0.616**	**1.361**	**0.287**	**1.281**	**3.228**	**0.888**	**0.409**
CooperScene (Wu et al., 2026)	CMP (Wang et al., 2025)	**0.330**	**0.563**	**0.870**	**0.407**	**0.876**	**1.677**	0.219	0.190
	UniV2X (Yu et al., 2025)	1.037	1.882	2.979	1.166	2.933	5.243	0.151	−0.091
	CooperTrace (Ours)	0.345	0.601	0.964	0.373	0.986	1.885	**0.547**	**0.338**
V2X-Seq-SPD (Yu et al., 2023)	CMP† (Wang et al., 2025)	0.924	2.515	4.281	1.572	4.915	8.732	0.222	0.094
	UniV2X (Yu et al., 2025)	1.010	1.931	3.047	1.176	3.041	5.496	0.318	−0.033
	CooperTrace (Ours)	**0.366**	**0.623**	**0.998**	**0.408**	**0.976**	**1.878**	**0.783**	**0.455**

Table 3: Cooperative Detection Performance

Dataset	Method	AP_{BEV} ↑ @0.3	AP_{BEV} ↑ @0.5	AP_{BEV} ↑ @0.7	AP_{BEV} ↑ mean	AP_{3D} ↑ @0.3	AP_{3D} ↑ @0.5	AP_{3D} ↑ @0.7	AP_{3D} ↑ mean
OPV2V (Xu et al., 2022c)	CoBEVT (Xu et al., 2022a)	0.884	0.869	0.685	0.813	0.836	0.783	0.499	0.706
	ERMVP (Zhang et al., 2024)	0.905	0.902	0.878	0.895	0.858	0.837	0.716	0.803
	CoSDH (Xu et al., 2025)	0.890	0.885	0.843	0.873	0.846	0.819	0.680	0.782
	CooperTrace (Ours)	**0.978**	**0.964**	**0.914**	**0.952**	**0.928**	**0.901**	**0.800**	**0.876**
CooperScene (Wu et al., 2026)	CoBEVT (Xu et al., 2022a)	0.767	0.680	0.348	0.598	0.737	0.554	0.150	0.480
	ERMVP (Zhang et al., 2024)	0.872	0.801	0.545	0.739	0.835	0.707	**0.332**	0.625
	CoSDH (Xu et al., 2025)	0.826	0.722	0.450	0.666	0.692	0.295	0.067	0.351
	CooperTrace (Ours)	**0.898**	**0.835**	**0.559**	**0.764**	**0.866**	**0.734**	0.301	**0.634**
V2X-Seq (Yu et al., 2023)	CoBEVT (Xu et al., 2022a)	0.541	0.492	0.344	0.459	0.525	0.410	0.132	0.356
	ERMVP (Zhang et al., 2024)	0.729	0.642	0.480	0.617	0.709	0.564	0.269	0.514
	CoSDH (Xu et al., 2025)	0.671	0.600	0.400	0.557	0.628	0.438	0.102	0.390
	CooperTrace (Ours)	**0.845**	**0.750**	**0.589**	**0.728**	**0.814**	**0.659**	**0.337**	**0.604**

pipeline receive no forecast regardless of the accuracy achieved on the remaining population. Each stack also has a non-cooperative counterpart (Appendix I).

CooperTrace improves coverage on all three datasets, from 0.878 to 0.911 on OPV2V, from 0.509 to 0.547 on CooperScene, and from 0.626 to 0.783 on V2X-Seq. On V2X-Seq, where the cooperative partner is an elevated roadside unit, cooperation also increases detection performance by 11 AP_{BEV} (0.614 to 0.728). CMP similarly improves coverage on OPV2V (0.575 to 0.711) and CooperScene (0.152 to 0.219), but decreases on V2X-Seq (0.315 to 0.222). In this setting, its cooperative CoBEVT detector performs 4 AP_{BEV} below its non-cooperative SinBEVT counterpart, its tracker retains 6 percentage points fewer agents, and the resulting forecaster is trained on the smallest and most biased history set among the three datasets; this configuration does not converge († in Table 2).

On CooperScene, the 22 % of agents retained by CMP's tracker are predominantly nearby and continuously observed, yielding a small but comparatively clean training population and correspondingly low forecasting error. On V2X-Seq, the roadside partner further reduces this population (4,227 training records per epoch versus 10,269) and alters its composition, after which the forecaster fails to converge. The same forecaster trained on ground-truth histories reaches 1.67 m (Appendix I), indicating that the failure is associated with the tracker-induced history distribution rather than the forecasting architecture alone. More generally, a modular forecaster can learn only from agents retained by its upstream tracker, whereas the end-to-end formulation propagates track queries directly to the forecasting stage. Section 4.4 isolates the effect of the resulting history quality.

Tracking. Cooperation improves tracking on every dataset and under both protocols, and tracking is what bounds forecasting coverage, since an agent the tracker drops is never forecast however accurate the remaining tracks are. Under the 2 m centre gate CooperTrace reaches 82.9 / 40.3 / 62.6 MOTA on OPV2V, CooperScene and V2X-Seq, against 21.2 for the CMP family on CooperScene and 6.4 for UniV2X on the frames it reports; under the AB3DMOT protocol the published baselines were scored with, it reaches sAMOTA 0.699 against 0.649 for the strongest of them. Against a modular hard-matching tracker on the same detections, the learned-query tracker hands the forecaster histories with a third of the noise and 3.4 to 3.8 times fewer identity switches per matched frame, which is where the end-to-end advantage in history quality comes from. Appendix F provides complete results under both protocols.

Detection. Table 3 evaluates several published cooperative detectors under our protocol. On OPV2V, we lead at every IoU threshold, with the margin widening at stricter overlap (AP_{3D}@0.7: 0.800 vs.

Table 4: Robustness under LiDAR degradation, single agent on OPV2V, camera raw: LiDAR alone (Single) against LiDAR + camera (Fusion). Left: perception, mAP_{3D} under low, moderate, and high corruption severity; right: forecasting under moderate severity. All nine corruptions are in Appendix D.

Corruption	Low		Moderate		High	
	Single	Fusion	Single	Fusion	Single	Fusion
None (clean)	**0.866**	0.812	**0.866**	0.812	**0.866**	0.812
Fog attenuation	**0.800**	0.749	0.599	**0.648**	0.268	**0.431**
Beam reduction	**0.768**	0.724	0.216	**0.443**	0.002	**0.210**
LiDAR missing	0.305	**0.528**	0.081	**0.409**	0.003	**0.326**

Corruption	minADE@5s↓		Cov.↑		EPA↑	
	Single	Fusion	Single	Fusion	Single	Fusion
None (clean)	**1.404**	1.473	**0.878**	0.868	**0.434**	0.418
Fog attenuation	**1.392**	1.493	0.660	**0.777**	0.324	**0.375**
Beam reduction	**1.546**	1.731	0.160	**0.617**	0.068	**0.266**
LiDAR missing	**1.399**	1.785	0.060	**0.609**	0.029	**0.249**

Table 5: Robustness under camera degradation, single agent on OPV2V, LiDAR raw: camera alone (Single) against LiDAR + camera (Fusion), same layout as Table 4.

Corruption	Low		Moderate		High	
	Single	Fusion	Single	Fusion	Single	Fusion
None (clean)	0.333	**0.812**	0.333	**0.812**	0.333	**0.812**
Darkening	0.316	**0.816**	0.226	**0.805**	0.000	**0.802**
Camera missing	0.219	**0.820**	0.080	**0.832**	0.000	**0.842**

Corruption	minADE@5s↓		Cov.↑		EPA↑	
	Single	Fusion	Single	Fusion	Single	Fusion
None (clean)	1.614	**1.473**	0.582	**0.868**	0.184	**0.418**
Darkening	1.680	**1.449**	0.443	**0.827**	0.126	**0.403**
Camera missing	2.534	**1.491**	0.184	**0.893**	−0.480	**0.422**

0.716 for ERMVP, the strongest published detector, and 0.499 for CoBEVT), indicating a localization advantage rather than recall. On CooperScene, ERMVP is within 2.5 points at loose thresholds and surpasses us at the strictest, suggesting that our gain there is primarily recall. The sim-to-real transition is also more robust: from OPV2V to V2X-Seq, our AP_{BEV}@0.5 drops by 22 % versus 43 % for the CoBEVT backbone used by CMP, and AP_{3D} drops by 27 % versus 48 %; ERMVP and CoSDH lose 29 % and 32 % of AP_{BEV}@0.5 on the same transition. Weaker boxes shorten tracks and degrade motion histories, so for an end-to-end system the sim-to-real gap propagates through the whole pipeline. Appendix A provides additional configurations and provenance.

4.3 Robustness to sensor degradation

Tables 4 and 5 evaluate whether complementary modalities compensate for sensor degradation under physics-driven emulated LiDAR and camera corruptions at low, moderate, and high severity.

Detection. The camera contribution increases as LiDAR quality degrades. On clean data, fusion trails LiDAR-only by five mAP_{3D} points (0.812 vs. 0.866), reaches parity under moderate fog, and performs better under high-severity fog (0.431 vs. 0.268). The gap widens under beam reduction, where fusion retains 0.443 and 0.210 at moderate and high severity, respectively, while LiDAR-only falls to 0.216 and 0.002. When nearly all LiDAR returns are removed, fusion still achieves 0.409 and 0.326 at moderate and high severity, compared with 0.081 and 0.003 for LiDAR-only. This robustness is reciprocal: darkening, blurring, or removing the camera drives a camera-only detector to near-zero performance under severe degradation, whereas the fusion model remains between 0.77 and 0.84 when the LiDAR path is intact (Table 5).

Forecasting. At the forecasting stage, the camera does not reduce minADE: it is higher in every row of Table 4. Under degraded LiDAR, this is a population effect rather than an accuracy effect. Under moderate fog, the two single-agent models are equally accurate on the agents both cover (1.386 vs. 1.381 at 5 s over 26,713 track-frames), while the fusion model additionally forecasts 5,066 track-frames that LiDAR-only drops, nine tenths of them at 30–60 m, with 2.08 m error; these harder agents raise its own-population mean to 1.493 (Appendix H). The benefit of the camera is therefore primarily in coverage and EPA rather than in per-agent error. Under moderate fog, the LiDAR-only model loses a quarter of its coverage (0.878 to 0.660) and EPA (0.434 to 0.324), whereas the fusion model retains 0.777 and 0.375. Under high-severity LiDAR degradation, the fusion model still forecasts 60.9 % of eligible agents, compared with 6.0 % for LiDAR-only, highlighting the robustness provided by multi-modal sensing in long-tail conditions.

4.4 Ablation study

Tables 6 to 8 ablate CooperTrace on the OPV2V exam. Table 6 changes one factor at a time from the cooperative reference configuration (three agents, LiDAR, lane priors, staged training); Table 7

Table 6: One factor changed at a time from the reference configuration (three agents, LiDAR, lane priors, staged training with the perception and tracking modules frozen) on the OPV2V exam.

Factor	Setting	minADE@5s↓	Cov.↑	EPA↑
Reference	3 agents, LiDAR, lane priors, staged training	1.365	0.911	0.417
Agents	1 agent (no cooperation)	1.404	0.878	**0.434**
Agents	5 agents	1.445	**0.937**	0.420
Sensors	LiDAR + camera	1.361	0.888	0.409
Lane priors	removed	1.857	0.911	0.294
Training	joint fine-tuning of the whole graph	**1.338**	0.906	0.430

Table 7: Inputs to the motion head: two heads trained from one random initialisation for five epochs on the same frozen tracker.

Inputs	minADE@5s↓	Cov.↑	EPA↑
track query only	4.423	0.878	0.172
track query + history	**2.020**	0.878	**0.210**

Table 8: Source of the training-time history for the query + history head (single agent, map-free, OPV2V); the last row is the deployed recipe.

Training-time history	minADE@5s↓	Cov.↑	EPA↑
current position repeated	3.990	**0.892**	0.140
in-clip rollout completed with ground truth	2.010	0.873	0.204
offline cache rolled by the deployed tracker	**1.783**	0.878	**0.264**

changes what the motion head reads and Table 8 what its training-time history contains, both with the single-agent map-free settings on OPV2V dataset.

Agents, sensors, lane priors and training. Each factor in Table 6 moves a different quantity. Cooperation moves coverage: 0.878 with a single agent, 0.911 with three and 0.937 with five, while minADE@5s rises at five agents because the added vehicles are distant and sparsely observed. For modality ablation, the camera is neutral on clean data, two points of coverage lower and minADE@5s unchanged (1.361 against 1.365); its value appears under sensor degradation (Section 4.3). Lane priors are the largest accuracy lever: removing them raises minADE@5s from 1.365 to 1.857 and lowers EPA from 0.417 to 0.294 at the same coverage (0.911). The lane encoding itself matters far less; Appendix E provides this comparison. Fine-tuning the whole graph jointly, instead of training the motion head on the frozen perception and tracking modules, improves minADE@5s by 0.027 m and EPA by 0.013.

Inputs to the motion head. Two motion heads trained from the same random initialization on the same frozen tracker (Table 7) isolate the explicit history: the head that reads the query alone reaches 4.423 m at 5 s, and adding the one-second history lowers it to 2.020 at unchanged coverage. Folding an annotated past into the query with a UniAD-style auxiliary regression does not recover it: 4.671 against 4.332 for the plain query from the same warm start (Appendix H).

Training-time history. What the history contains then matters (Table 8): repeating the current position gives 3.990, a ground-truth-completed in-clip rollout 2.010, and the cache rolled by the deployed tracker 1.783 with the highest EPA (0.264). The ordering follows alignment with deployment rather than information content, so we use tracker-generated histories throughout; this is the largest single effect in our ablations, and it is available because the tracker is part of the same graph.

5 Conclusion

We present CooperTrace, a cooperative multi-modal end-to-end framework for detection, tracking, and motion forecasting. Under a deployment-oriented protocol using tracker-generated histories, CooperTrace achieves 1.361 m minADE at 5 s on OPV2V and 0.998 m on V2X-Seq-SPD. Our ablations show that the main gains come from aligning the forecaster with the tracker state and history distribution encountered at deployment, rather than from joint backpropagation alone. Cooperation consistently improves coverage, while the benefit of camera sensing is strongest under LiDAR degradation, motivating reliability-aware fusion. Current limitations include memory-constrained joint training, evaluation with at most five agents, reliance on OPV2V for most ablation and robustness studies, and the exclusion of planning. Two findings point forward: the track query alone is a weak forecasting input, so cooperative end-to-end systems should carry trajectories explicitly; and the camera primarily improves coverage rather than conditional forecasting accuracy, which argues for fusion conditioned on measured sensor reliability.

AI USE STATEMENT

Generative AI tools were used to assist with software code, experiment orchestration, retrieval and discovery of relevant literature, and drafting and editing portions of the manuscript. All AI-assisted code was reviewed and tested by the authors, and all AI-assisted manuscript text and literature suggestions were reviewed and verified for accuracy and relevance. The experiments were designed and run by the authors, and every numerical result reported in the paper was traced to the corresponding evaluation log or experimental output. The authors take responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

All datasets are public research datasets containing no personal data. Motion prediction for driving is safety-relevant; the models here are research prototypes and are not validated for deployment.

REPRODUCIBILITY STATEMENT

We will release the code, model configurations, dataset converters, and evaluation files used in this work. Section 4.1 specifies the evaluation protocol and model-selection procedure; Appendix H reports the boundary checkpoints used in our analyses; Appendix G documents the metric implementations and procedures used to re-evaluate external baselines; and Appendix K provides hyperparameters, random seeds, container information, and hardware details. Released result artifacts will include the corresponding evaluation protocol in their identifiers to avoid ambiguity across evaluation settings.

REFERENCES

Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1090–1099, June 2022.

Till Beemelmanns, Quan Zhang, Christian Geller, and Lutz Eckstein. Multicorrupt: A multi-modal robustness dataset and benchmark of lidar-camera fusion for 3d object detection. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pp. 3255–3261, 2024. doi: 10.1109/IV55156.2024.10588664.

Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008, 2008. URL https://link.springer.com/article/10.1155/2008/246309.

Junru Gu, Chenxu Hu, Tianyuan Zhang, Xuanyao Chen, Yilun Wang, Yue Wang, and Hang Zhao. ViP3D: End-to-end visual trajectory prediction via 3D agent queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5496–5506, June 2023.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. In *Thirty-sixth Conference on Neural Information Processing Systems (Neurips)*, November 2022.

Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. BEVDepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1477–1485, 2023.

Ming Liang, Bin Yang, Wenyuan Zeng, Yun Chen, Rui Hu, Sergio Casas, and Raquel Urtasun. PnPNet: End-to-end perception and prediction with tracking in the loop. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L. Rus, and Song Han. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

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 *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4812–4818. IEEE, 2023.

Wenjie Luo, Bin Yang, and Raquel Urtasun. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Jonah Philion Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Proceedings of the European Conference on Computer Vision*, 2020.

Hongzhi Ruan, Haibao Yu, Wenxian Yang, Siqi Fan, and Zaiqing Nie. Learning cooperative trajectory representations for motion forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and local movement refinement. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 6531–6543, 2022.

Ziyi Song, Chen Xia, Chenbing Wang, Haibao Yu, Sheng Zhou, and Zhisheng Niu. Unimm-v2x: Moe-enhanced multi-level fusion for end-to-end cooperative autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40. AAAI Press, 2026. doi: 10.1609/aaai.v40i11.37870.

Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, James Tu, and Raquel Urtasun. V2VNet: Vehicle-to-vehicle communication for joint perception and prediction. In *Computer Vision – ECCV 2020*, pp. 605–621. Springer International Publishing, 2020. doi: 10.1007/978-3-030-58536-5_36.

Zehao Wang, Yuping Wang, Zhuoyuan Wu, Hengbo Ma, Zhaowei Li, Hang Qiu, and Jiachen Li. CMP: Cooperative motion prediction with multi-agent communication. *IEEE Robotics and Automation Letters*, 2025.

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

Bo Wu, Ruoshen Mo, Justin Yue, Yanyu Zhang, Janice Nguyen, Guoyuan Wu, Amit Roy-Chowdhury, Matthew J. Barth, and Hang Qiu. Cooperscene: Multi-modal cooperative autonomy benchmark with c-v2x communication characterization. In *European Conference on Computer Vision*, 2026.

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

Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers. In *Conference on Robot Learning (CoRL)*, 2022a.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022b.

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 IEEE International Conference on Robotics and Automation (ICRA)*, 2022c.

Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, Hongkai Yu, Bolei Zhou, and Jiaqi Ma. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *The IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2023.

Tengju Ye, Wei Jing, Chunyong Hu, Shikun Huang, Lingping Gao, Fangzhen Li, Jingke Wang, Ke Guo, Wencong Xiao, Weibo Mao, Hang Zheng, Kun Li, Junbo Chen, and Kaicheng Yu. FusionAD: Multi-modality fusion for prediction and planning tasks of autonomous driving. *arXiv preprint arXiv:2308.01006*, 2023.

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*, 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 *The 39th Annual AAAI Conference on Artificial Intelligence*, 2025.

Jingyu Zhang, Kun Yang, Yilei Wang, Hanqi Wang, Peng Sun, and Liang Song. ERMVP: Communication-efficient and collaboration-robust multi-vehicle perception in challenging environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12575–12584, 2024.

Xinyu Zhang, Zewei Zhou, Zhaoyi Wang, Yangjie Ji, Yanjun Huang, and Hong Chen. Co-MTP: A cooperative trajectory prediction framework with multi-temporal fusion for autonomous driving. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 801–807, 2025.

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 (ICCV)*, pp. 26954–26965, 2025.

Zewei Zhou, Hao Xiang, Zhaoliang Zheng, Seth Z. Zhao, Mingyue Lei, Yun Zhang, Tianhui Cai, Xinyi Liu, Johnson Liu, Maheswari Bajji, Xin Xia, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. V2xpnp: Vehicle-to-everything spatio-temporal fusion for multi-agent perception and prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25399–25409, October 2025.

A Full detection matrix and the union-label rule

Table 9: Detection on all six datasets (mean AP over IoU 0.3 / 0.5 / 0.7, union exam per dataset).

	Single L		Single L+C		Coop L		Coop L+C	
Dataset	mAP_{BEV}	mAP_{3D}	mAP_{BEV}	mAP_{3D}	mAP_{BEV}	mAP_{3D}	mAP_{BEV}	mAP_{3D}
OPV2V	0.942	0.866	0.921	0.838	0.952	0.876	**0.958**	**0.882**
CooperScene	0.476	0.386	0.497	0.403	**0.722**	**0.597**	0.714	0.590
V2V4Real	0.667	0.531	no camera		**0.820**	**0.623**	no camera	
DAIR-V2X-C	0.727	0.633	0.723	0.628	0.830	0.691	**0.830**	**0.698**
TUMTraf V2X	0.682	0.523	no camera calib.		**0.902**	**0.782**	no camera calib.	
V2X-Seq	0.614	0.489	0.644	0.535	0.728	**0.604**	0.709	0.582

Table 10: Two further published detectors on the OPV2V (Xu et al., 2022c) union exam ($\pm 140.8 \times \pm 40$ m), same scorer as Table 3.

	$\text{AP}_{BEV} \uparrow$				$\text{AP}_{3D} \uparrow$			
Method	@0.3	@0.5	@0.7	mean	@0.3	@0.5	@0.7	mean
CoAlign (Lu et al., 2023)	0.888	0.878	0.800	0.855	0.846	0.814	0.650	0.770
Where2comm (Hu et al., 2022)	0.858	0.855	0.822	0.845	0.816	0.798	0.702	0.772

Table 11: The union-label rule (mAP_{3D}, one exam per row): cooperative models trained on ego-visible labels score below the single-agent model.

Dataset	Single	Coop, ego-visible labels	Coop, union labels	Gain
V2V4Real	0.531	0.507	**0.623**	+9.2
DAIR-V2X-C	0.633	0.594	**0.691**	+5.8
CooperScene	0.386	0.230[k]	**0.597**	+21.1
OPV2V	0.866	natively union	0.876	+1.0

[k]scored on an earlier ±72 m exam, on which the single model scores 0.316.

B TRANSFER TO COOPERSCENE AND V2V4REAL

Table 12: Modular pipeline on two real cooperative datasets, minADE/minFDE in metres.

		@1s↓		@3s↓		@5s↓		
Dataset	Setting	minADE	minFDE	minADE	minFDE	minADE	minFDE	n
CooperScene (Wu et al., 2026)	Single L	0.729	0.992	1.364	2.157	2.001	3.441	31,985
	Single L+C	0.730	1.001	1.381	2.212	2.035	3.592	32,878
	Coop L	0.688	0.924	1.313	2.111	1.929	3.293	33,896
	Coop L+C	0.680	0.939	1.330	2.233	2.020	3.542	32,887
	Coop L, trajectory exchange	0.627	0.860	1.226	1.984	**1.868**	**3.342**	31,985
V2V4Real (Xu et al., 2023)	Single L	0.298	0.433	0.729	1.341	1.232	2.363	25,133
	Coop L	0.315	0.445	0.713	1.286	**1.203**	**2.361**	25,435
	SinBEVT (Xu et al., 2022a)	1.45	2.68	3.48	6.28	5.03	8.67	–
	V2VNet (Wang et al., 2020)	1.45	2.31	3.10	5.55	4.44	7.57	–
	CMP-Full (Wang et al., 2025)	1.14	2.03	2.66	4.82	3.91	7.02	–

Notes. The SinBEVT, V2VNet, and CMP-Full rows are reported by CMP (Wang et al., 2025). The CooperScene (Wu et al., 2026) single and cooperative rows were produced with detector input clips of different extent (±140.8 m and ±72 m, respectively), although both are scored on the same range-140 exam. Part of the cooperative gain may therefore reflect the input-range difference rather than cooperation alone.

C GRADIENT AUDIT AND MEMORY

Table 13: Gradient audit: norms after one backward pass of the joint loss on a 4-frame clip (top) and parameter tensors receiving gradient in the cooperative model (bottom).

Module (stand-in)	params	with grad	grad norm	finite
detector BEV trunk	24	18	36.83	yes
track propagation	64	64	25.94	yes
motion head	45	28	0.99	yes
forecasting loss only → detector trunk	24	16	2.27	yes
frame-1 loss → frame-0 BEV feature	–	–	0.0035	yes
Module (real cooperative model)	**tensors**	**with grad**		
per-agent trunk (frozen)	107	0		
cross-agent fusion	72	72		
detection head	64	64		
tracker	80	76		
motion head	75	53		

Table 14: Peak GPU memory under 4-frame back-propagation through time (48 GB cards).

Configuration	peak memory
single-agent joint stage, full detector trainable	10.9–11 GiB
cooperative joint stage, 3 agents, full trunk trainable	OOM (46.3 GiB at the first clip)
cooperative joint stage, 3 agents, frozen per-agent trunk	37.4 GiB
fusion graph alone, 3 agents: plain / checkpointed	30.5 / 2.9 GiB
fusion graph alone, 5 agents: plain / checkpointed	$>$ 46.4 GiB (OOM) / 4.9–7.8 GiB
cooperative detector fine-tuning, 5 agents, batch 2	35.9 GiB

D COOPERATIVE ROBUSTNESS

We group the corruption settings into three ordered strengths—low, moderate, and high—corresponding to progressively stronger sensor degradation.

Table 15: All nine single-agent corruptions, same layout as Table 4; the camera-only model is a stage-3 checkpoint.

	Perception, mAP_{3D} ↑						Forecasting at moderate severity					
	Low		Moderate		High		minADE@5s↓		Cov.↑		EPA↑	
Corruption	alone	L+C	alone	L+C	alone	L+C	alone	L+C	alone	L+C	alone	L+C
LiDAR degraded, camera raw ("alone = LiDAR only")												
none (clean)			**0.866** / 0.812				**1.404**	1.473	**0.878**	0.868	**0.434**	0.418
fog attenuation	**0.800**	0.749	0.599	**0.648**	0.268	**0.431**	**1.392**	1.493	0.660	**0.777**	0.324	**0.375**
beam reduction	**0.768**	0.724	0.216	**0.443**	0.002	**0.210**	**1.546**	1.731	0.160	**0.617**	0.068	**0.266**
point dropout	**0.858**	0.803	**0.769**	0.735	0.306	**0.526**	**1.435**	1.520	0.791	**0.830**	0.385	**0.393**
Gaussian noise	**0.866**	0.811	**0.837**	0.795	0.652	**0.727**	**1.420**	1.503	0.842	**0.853**	**0.408**	0.405
sector occlusion	**0.722**	0.720	0.659	**0.676**	0.567	**0.613**	**1.549**	1.646	0.795	**0.828**	0.353	**0.364**
LiDAR missing	0.305	**0.528**	0.081	**0.409**	0.003	**0.326**	**1.399**	1.785	0.060	**0.609**	0.029	**0.249**
Camera degraded, LiDAR raw ("alone = camera only")												
none (clean)			0.333 / **0.812**				1.614	**1.473**	0.582	**0.868**	0.184	**0.418**
darkening	0.316	**0.816**	0.226	**0.805**	0.000	**0.802**	1.680	**1.449**	0.443	**0.827**	0.126	**0.403**
motion blur	0.220	**0.805**	0.051	**0.767**	0.007	**0.811**	2.267	**1.466**	0.106	**0.799**	−0.031	**0.382**
camera missing	0.219	**0.820**	0.080	**0.832**	0.000	**0.842**	2.534	**1.491**	0.184	**0.893**	−0.480	**0.422**

Table 16: Cooperative counterparts of Table 4 at moderate severity (three agents, OPV2V (Xu et al., 2022c)). Both columns are the Table 2 checkpoints; – = not run.

	Coop, LiDAR only			Coop, LiDAR + camera		
Corruption (moderate severity)	minADE@5s↓	Cov.↑	EPA↑	minADE@5s↓	Cov.↑	EPA↑
none (clean reference)	1.365	0.911	0.417	1.361	0.888	0.409
fog attenuation	1.347	0.781	0.378	1.475	0.809	0.367
beam reduction	1.534	0.327	0.126	–	–	–
point dropout	1.386	0.864	0.402	–	–	–
Gaussian noise	1.367	0.897	0.400	–	–	–
snow	1.453	0.664	0.286	–	–	–
sector occlusion	1.506	0.814	0.331	1.607	0.851	0.350
LiDAR missing	1.633	0.187	0.065	1.820	0.579	0.211
motion blur (camera)	1.365	0.911	0.417	1.491	0.854	0.376
fog + motion blur	1.347	0.781	0.378	1.485	0.733	0.345

A LiDAR-only cooperative system is unaffected by camera corruptions, so its blur and fog+blur rows repeat its clean and fog rows.

Snow is the one corruption the camera cannot insure against: it adds phantom returns rather than removing real ones, so single-agent LiDAR + camera reaches 0.789 / 0.541 / 0.260 mAP_{3D} at low / moderate / high severity against 0.835 / 0.509 / 0.290 for LiDAR alone; training the fusion model on a mixture of corruption families lifts those cells to 0.797 / 0.662 / 0.317, above LiDAR-only at moderate and high severity, at a cost of 0.003 on clean data. Fog with motion blur is the compositional case: degrading both sensors at once leaves the single-agent fusion model at 1.441 / 0.647 / 0.314

against 1.392 / 0.660 / 0.324 for LiDAR alone, so the two redundancies do not compose when both are attacked.

E LANE ENCODING AND THE FIRST-EPOCH ANCHOR

Two ablations are held out of Tables 6 to 8 because they were run on the single-agent line only, and a table whose adjacent rows differ in more than one variable is worse than no table.

Lane encoding. Replacing our eight lane polylines with a BEV raster read by a small CNN — the form the CMP (Wang et al., 2025) family uses — gives minADE@5s 1.364 at coverage 0.878 and EPA 0.436, against 1.404 / 0.878 / 0.434 for polylines. The two encodings are within 0.04 m, so the map representation is not what separates the systems in Table 2. No cooperative raster run exists.

Holding the head to its first epoch. On the cooperative line the motion head drifts after its first epoch, trading its stationary-vehicle mode for straight-line endpoint hit rate. Distilling it towards a target taken from its own first-epoch weights — a training-time regulariser on this head alone, inert at inference — gives 1.742 / 0.911 / 0.273 cooperative without lane priors (against 1.857 / 0.911 / 0.294), 1.328 / 0.911 / 0.416 cooperative with them (against 1.365 / 0.911 / 0.417), and 1.331 / 0.878 / 0.387 single-agent with them (against 1.404 / 0.878 / 0.434). It buys mean error and sells endpoint hit rate: every cooperative row improves in minADE while EPA is flat or lower, and the single-agent row loses 0.047 of EPA for 0.073 m. The main tables therefore use the unanchored head.

F TRACKING UNDER BOTH PROTOCOLS

Table 17: Tracking under two protocols. Block (a): the AB3DMOT (Weng et al., 2020) evaluator with a 3D-IoU gate on the 16 OPV2V (Xu et al., 2022c) test sequences, matching the protocol used for the published baseline. Block (b): CLEAR-MOT with a 2 m centre gate on the full deployment exam. The two blocks are not directly comparable.

(a) AB3DMOT protocol, 3D-IoU gate — OPV2V test, 16 sequences

Tracker	Setting	sAMOTA↑	AMOTA↑	AMOTP↑	MOTA (%)↑	IDS↓
CooperTrace	Coop 3 CAV (L+C)	**0.699**	**0.471**	**0.380**	**67.3**	217
CoBEVT + AB3DMOT[a]	Coop	0.649	0.410	0.341	63.1	4[b]

(b) Deployment ledger, 2 m centre gate — what the forecaster consumes

Dataset	Tracker, setting	MOTA (%)	IDF1 (%)	MOTP (m)	Cov.
OPV2V	**CooperTrace**, Coop 3 CAV	**82.9**	**75.7**	0.146	**0.911**
CooperScene	**CooperTrace**, Coop V+2V	**40.3**	**48.2**	0.406	**0.547**
CooperScene	CoBEVT + AB3DMOT, Coop[c]	21.2	37.0	**0.259**	0.219
V2X-Seq	**CooperTrace**, Coop V+I	**62.6**	**58.8**	0.317	**0.783**
V2X-Seq	CoBEVT + AB3DMOT, Coop[c]	26.7	42.0	0.216	0.222
V2X-Seq, restricted to the 675 frames reported by UniV2X (Yu et al., 2025) at 2 Hz					
V2X-Seq	UniV2X, Coop V+I[d]	6.4	24.4	0.772	0.318
V2X-Seq	**CooperTrace**, Coop V+I	**55.0**	**58.1**	0.322	**0.783**

Notes. OPV2V (Xu et al., 2022c); CooperScene (Wu et al., 2026); V2X-Seq (Yu et al., 2023); CoBEVT (Xu et al., 2022a); AB3DMOT (Weng et al., 2020). [a]Outputs released with CMP (Wang et al., 2025), re-evaluated by us under the same evaluator. [b]The CMP (Wang et al., 2025) pipeline re-keys its tracker input using ground-truth identities upstream; its false positives are therefore zero by construction, and its identity-switch count is not comparable with ours. [c]Retrained by us and scored on the same frames and ground-truth pool as our rows. [d]Official UniV2X checkpoint at its native 2 Hz; our tracker runs at 10 Hz and is only *scored* on those frames, which is why its MOTA is 55.0 here versus 62.6 when scored on all frames.

Table 18: CMP (Wang et al., 2025)'s motion evaluation, audited from its code, against our deployment exam.

Aspect	CMP (Wang et al., 2025)	Ours
forecast grid	50 steps at 10 Hz	same
error at 5 s	mean over steps 1–49	steps 1–50
endpoint at 1 s / 3 s	step 11 / 31 (1.1 s / 3.1 s)	step 10 / 30
prediction-truth identity	oracle: detections filtered at IoU 0.7, tracks re-keyed to the majority ground-truth identity	greedy centre match within 2 m at evaluation time
false positives	dropped before tracking	stay in the tracker; unmatched tracks unscored
history validity	all 11 tracker states present, interpolation counts	all 11 present, at least 3 lifetime detections, matched within the last 3 frames
sample universe	every frame × every ego × every observer (130,491 samples)	every frame × every ego (one score per matched track-frame)

G Protocol audits

CMP reproduction. (Wang et al., 2025) Their released weights evaluated with their code give CoBEVT (Xu et al., 2022a) 1.852 / 4.143 (paper: 1.858 / 4.163) and SinBEVT (Xu et al., 2022a) 2.217 / 5.176 (paper: 2.222 / 5.185). Per sample, we match their uploaded results to within 0.001 m.

UniV2X convention. (Yu et al., 2025) Their evaluator on their cooperative weights reproduces the published 0.69 / 0.74 in its accuracy-style table (0.686 / 0.711, miss rate 0.069), while the same run's absolute-error table reports 2.859 / 5.311 at 5 s. Our scorer gives 2.798 / 4.907 on the matched population.

FusionAD common-range control. (Ye et al., 2023) On the common range ($|x| \leq 51.2$, $|y| \leq 40$ m; 34,817 of 40,874 eligible agents), FusionAD (Ye et al., 2023) obtains 3.763 / 7.432 at coverage 0.735, compared with 1.693 / 4.056 at 0.945 for our single-agent end-to-end model. On the complement, FusionAD (Ye et al., 2023) matches 76 of 6,057 vehicles, whereas ours forecasts 58.7 % at 2.430 / 5.536.

V2X-Seq common-subset construction. (Yu et al., 2023) UniV2X (Yu et al., 2025)'s true-positive set contains 1,180 samples at 2 Hz. The strict four-way intersection with all three of our lines contains 865 samples, while the phase-union subset admitting all five 2 Hz offsets contains 4,466. A naive 10 Hz variant is a cadence artifact and is therefore not reported.

H Per-boundary numbers and same-population analysis

Table 19: Every boundary checkpoint of the end-to-end runs (OPV2V (Xu et al., 2022c) test, minADE/minFDE at 5 s); the reported boundary is in bold. The from-scratch pair ran to epoch 8 (query only) and epoch 6 (query + history).

	iter 20,000		iter 40,000		last	
Run	minADE	minFDE	minADE	minFDE	minADE	minFDE
single, cache history	**1.765**	**4.200**	1.801	–	1.797	4.357
single, cache + lane priors	**1.329**	**2.965**	1.326	–	1.328	–
single, + curvature/distillation/augmentation	**1.458**	**3.340**	1.613	3.782	1.663	3.987
single, no history channel	**4.332**	**8.009**	4.566	8.415	4.664	8.536
single, query + box heading, no history	**5.285**	**10.052**	–	–	4.778	9.042
single, UniAD (Hu et al., 2023)-style past folded into the query	**4.671**	**8.719**	–	–	4.769	8.986
single, same, heading from own motion	**5.001**	**9.699**	–	–	–	–
coop, cache history, no priors	**1.835**	**4.478**	1.862	4.620	1.981	5.042
coop, cache + lane priors	**1.263**	**2.869**	1.273	2.909	1.301	3.003
From random initialisation (Table 7, top): epoch 5 / epoch 6 / last						
single, track query only	**4.423**	**8.253**	4.389	8.285	4.434	8.505
single, track query + history	**2.020**	**4.860**	2.105	5.060	2.105	5.060

The warm-started input arms all score below the plain query at the same 20,000-iteration budget: a head warm-started from a checkpoint trained with both inputs measures the distance from that checkpoint rather than the value of each input, which is why Table 7 reports the from-scratch pair and this table keeps the warm-started arms. The UniAD (Hu et al., 2023)-style arm adds an auxiliary head that regresses each live query's annotated past (ten steps at 10 Hz, L1 loss) so that the history can enter the query implicitly, with the agent frame taken from the box yaw or from the track's own motion; it reaches 4.671 and 5.001 at 20,000 iterations and 4.769 at the end, against 4.332 and 4.664 for the plain query, so the auxiliary regression does not put the kinematics into the query within this budget.

Table 20: Effect of cooperation by ego-distance (OPV2V (Xu et al., 2022c), lane priors): same-population error (n=35,549), with coverage and endpoint recall on each model's own population.

		Error↓			Coverage↑		Endpoint recall↑	
Distance	n	Single	Coop	Gain	Single	Coop	Single	Coop
0–30 m	20,634	1.196	1.140	4.7 %	95.9 %	95.2 %	51.7 %	**54.1 %**
30–60 m	13,041	1.385	1.300	**6.2 %**	85.0 %	85.3 %	42.0 %	42.4 %
60–100 m	1,859	1.945	1.918	1.4 %	58.2 %	59.4 %	**17.1 %**	14.0 %
100–140 m	15	5.085	5.066	0.4 %	18.8 %	**40.0 %**	0 %	0 %
all	35,549	1.306	1.241	5.0 %	–	–	–	–

Exclusive sets: cooperation forecasts 1,264 vehicle-frames over 195 tracks the single-agent model never forecasts, at 1.886 / 4.021 and a median distance of 60.1 m; the reverse set is similar in size but distinctly nearer, consistent with extra identity churn near the ego under cooperation. Without lane priors the recall advantage of cooperation grows monotonically with range, peaking at +49 % relative at 60–100 m.

LiDAR against LiDAR + camera on the same population. Single agent with lane priors on the OPV2V test exam, the two rows of Table 4. Clean: 34,580 track-frames are covered by both models, at 1.390 (LiDAR) against 1.455 (LiDAR + camera) minADE@5s, and the sets each model covers alone are of similar size (1,326 and 909 track-frames). Fog at severity 3: 26,713 common track-frames at 1.386 against 1.381; the fusion model alone covers a further 5,066 track-frames (346 tracks, 92 % of them at 30–60 m) at 2.084 m, while LiDAR alone covers 275 that the fusion model drops. The fusion model's higher own-population mean under fog (1.493 against 1.392) is therefore the price of forecasting the agents LiDAR alone loses, not lower accuracy on the agents both see. (Xu et al., 2022c)

I External baselines

UniV2X on V2X-Seq-SPD. (Yu et al., 2025; 2023) Using the official code and weights, lane IoU reproduces the published 16.2 % and 17.8 % for the two stages exactly, validating our data preparation. Its tracking metric under our re-evaluation is 0.11–0.14, compared with 0.24–0.30 published, indicating a recall ceiling in the pipeline rather than a preparation defect. On our union detection exam, its head scores 0.131 / 0.013.

UniV2X on OPV2V. (Yu et al., 2025; Xu et al., 2022c) Using our implementation of the official four-stage pipeline, with geometry verified pixel-accurately, detection fails in both the single-agent and cooperative settings. A train-set overfitting probe reaches only 3–7 % recall on frames seen during 20 epochs of training, indicating that the failure arises during training rather than evaluation. The resulting predictions are also poorly conditioned on the image: 73–77 % of confident detections fall behind the ego despite the use of front-facing cameras only. One plausible contributing factor is the limited training set, which contains 1,292 cooperative frames sampled at 2 Hz, with a single 800×600 camera per agent.

CMP on V2X-Seq-SPD. (Wang et al., 2025; Yu et al., 2023) Its three detectors are retrained by us and its own AB3DMOT (Weng et al., 2020) and MTR (Shi et al., 2022) code are used, so the row in Table 2 is a property of the modular pipeline, not of the dataset. Funnel on the 48,202 ground-truth boxes: CoBEVT (Xu et al., 2022a) detects 52 % at IoU 0.5 (AP 0.492; CoSDH (Xu et al., 2025) 66 %

/ 0.600 and ERMVP (Zhang et al., 2024) 71 % / 0.617 on the same exam), AB3DMOT (Weng et al., 2020) keeps 27 % (precision 99 %), and 22 % of eligible agents reach the forecaster with a complete history. The forecaster is then trained on 4,227 records per epoch against 10,269 on CooperScene (Wu et al., 2026), and its own evaluator never settles: minADE@5s per epoch 9.7, 9.3, 10.4, 4.1, 8.0, 9.2, 7.4, 6.3, 9.8, 8.3, 4.4, 8.2 (final 6.67; CooperScene (Wu et al., 2026) converges to 1.0), with an initial loss 80 times that of the CooperScene (Wu et al., 2026) run. Input conventions (heading in degrees, half extents, world coordinates) are identical to the CooperScene (Wu et al., 2026) run that converged, and the rare identity jumps in the trajectories (0.02 % of steps) are present in our own annotations as well. The control that isolates the cause: the same MTR (Shi et al., 2022), same data and schedule, trained and evaluated on ground-truth histories instead of AB3DMOT (Weng et al., 2020) output, converges from the third epoch (1.75, 2.32, 1.75, 1.68, 1.70, 1.85, 1.68 m), scores 1.67 m at its best epoch under CMP (Wang et al., 2025)'s own selection rule and drifts to 2.84 m by the last, so even with oracle histories the schedule is not fully stable at batch size 1, but it never returns to the 4–10 m band of the tracker-history run. Cooperation itself is not the failure: CMP (Wang et al., 2025)'s non-cooperative SinBEVT (Xu et al., 2022a) scores 0.499 AP and 0.315 coverage on V2X-Seq (Yu et al., 2023) and its cooperative CoBEVT (Xu et al., 2022a) 0.459 and 0.222, whereas our single-agent to cooperative step moves 0.614 to 0.728 AP and 0.626 to 0.783 coverage on the same roadside unit.

FusionAD on OPV2V. (Ye et al., 2023; Xu et al., 2022c) FusionAD (Ye et al., 2023) is retrained on OPV2V (Xu et al., 2022c) with checkpoint selection by validation error. Forecasts are interpolated from its native 2 Hz output to 10 Hz, with endpoints preserved exactly. One caveat remains: its training pipeline scales images to quarter resolution while its test pipeline feeds full resolution, so the camera branch may contribute little or nothing at test time and our reported row may be pessimistic.

Where2comm and CoAlign. (Hu et al., 2022; Lu et al., 2023) Union-label retrains give CoAlign (Lu et al., 2023) 0.855 / 0.770 at σ=0, 0.830 / 0.741 at σ=0.4, and 0.811 / 0.723 at σ=0.6 in the pose-noise sweep.

Emission ablation. Emitting an under-trained per-query box head gives 0.014–0.029. Emitting the detection head's boxes while propagating identity through the queries gives 0.439 / 0.427 / 0.433 at three epochs.

UniMM-V2X and V2XPnP. (Song et al., 2026; Zhou et al., 2025) UniMM-V2X (Song et al., 2026) has released code but no weights. The only public reproduction emits 12 tracked cars over a 675-frame validation split under its own evaluator, so no quantitative result can be attributed to the method without retraining. V2XPnP (Zhou et al., 2025) provides no model, training code, or inference code as of September 2026.

Why the CMP rows omit its trajectory aggregator. (Wang et al., 2025) The full CMP (Wang et al., 2025) configuration did not train to completion on CooperScene (Wu et al., 2026) in four attempts: with the authors' `GRAD_NORM_CLIP` of 1000 it diverged at epoch 4, at a clip of 10 at epoch 15 and again over a full 30-epoch run, and resuming the last finite checkpoint at a clip of 1.0 with the learning rate reduced to 10^{-6} it diverged again at epoch 21. Clipping and learning-rate reduction postpone the divergence rather than remove it; immediately before the fourth one the loss oscillated through a negative value, which points at a collapsed variance in the objective rather than an exploding gradient. Every non-finite checkpoint was archived rather than deleted. Reporting the aggregator-free configuration is the charitable choice, not a handicap: scoring the best finite full-model checkpoint we obtained — epoch 21 of 30, verified finite in all 1,059 tensors — gives minADE 0.621 / 1.366 / 2.185 and EPA 0.172 at coverage 0.219, against 0.330 / 0.563 / 0.870 and EPA 0.190 at the same coverage for the aggregator-free model trained to completion. Where the full model does train, it is the weaker of the two on this dataset.

UniV2X box convention on our detection exam. (Yu et al., 2025) Its predictions arrive in our ego frame already, but two conventions differ and both were identified against exam ground truth on centre-matched pairs rather than fitted to any metric. Its (d_x, d_y) are (width, length) where ours are (length, width), and its heading is $-\psi + \pi/2$ in our frame: that choice gives a median $|\Delta\psi|$ of 4.1° with 86.2 % of pairs inside 15°, against 80.3° and 10.2 % for the raw angle. Correcting it raises its

AP_{BEV} at IoU 0.3 from 0.182 to 0.219. A convention-free check agrees with the resulting level: at a score threshold of 0.2 it emits 9,334 boxes against 9,288 ground-truth vehicles and 41 % of those vehicles have a prediction within 2 m, so its centres are roughly right while its extents and headings are not. The converter is `tools/univ2x_det/univ2x_to_det_dump.py`.

J COMMUNICATION COST

Table 21: Fusion at four stages of the pipeline on CooperScene (Wu et al., 2026), reporting minADE and minFDE at 1, 3, and 5 s together with measured payload.

	@1s		@3s		@5s			
Fusion stage	minADE	minFDE	minADE	minFDE	minADE	minFDE	n	KB/frame
none (single agent)	0.729	0.992	1.364	2.157	2.001	3.441	31,985	0
BEV features	0.688	0.924	1.313	2.111	1.929	3.293	33,896	10^2–10^3
track states	0.662	0.895	1.234	1.951	**1.820**	**3.146**	48,438	3.1
predicted trajectories	**0.627**	**0.860**	1.226	1.984	1.868	3.342	31,985	3.4

Exchanging predicted trajectories reaches most of the benefit of exchanging BEV features at roughly 1/300 of the payload, and beats it outright on V2V4Real (Xu et al., 2023). Track exchange is a coverage mechanism and not an accuracy one. An adoption-disabled control returns its population to the baseline's and its error advantage disappears entirely – 1.231 against 1.232, one millimetre, inside the harness noise floor and with its history splices still applied – while at full population it is actually worse than the baseline. The apparent accuracy gain is composition: the adopted tracks must average about 1.47 m at five seconds to produce the reported mean, markedly easier than the population they join. This is the coverage bias we identify in Section 4.1 appearing inside one of our own methods rather than in a baseline, which is the reason we report it here rather than quietly restating the row. In-graph feature fusion is the accuracy end of this curve and trajectory exchange the bandwidth end; a deployment can pick a point on it without retraining perception. Query-level fusion at the detection stage helps where the single detector is weak ($+20.6$ mAP_{3D} on CooperScene (Wu et al., 2026)) and hurts where it is strong.

K IMPLEMENTATION DETAILS

- Perception: LiDAR trunk on a $0.1 \times 0.1 \times 0.2$ m voxel grid over $[-140.8, 140.8] \times [-40, 40]$ m, TransFusion (Bai et al., 2022) head with 200 queries of dimension 128, 30 epochs cosine; cooperative variant with fused-axial cross-agent attention, three agents within 70 m (five in the scaling study), union labels; camera branch Swin-T with depth-supervised lift-splat at 256×512, frozen-LiDAR and zero-init recipe; camera-only detector at 384×800 with LiDAR depth as a training-time label only.
- Tracker: 4-frame clips, existence thresholds 0.3, deployment emission from the detection head at 0.3, matching 2 m, minimum 3 hits, at most 3 coasted frames.
- Motion head: K=6, 50 steps at 10 Hz, history length 11, displacement clamp 3 m, K-means anchors; lane priors from polylines extracted from the map rasters (skeletonised, 20 points per line, after correcting an axis transpose in the released converter), eight nearest within 30 m, 10 % line dropout at training.
- Schedules: motion stage 8 configured epochs at $2 \cdot 10^{-4}$; joint stage 3 epochs at 10^{-4} cosine with the multipliers of Section 3.5, checkpoints every 2,500 iterations. History caches: 241,637 entries single (roll quality MOTA 88.3, IDF1 77.5, noise 0.158 m per step, 86 % of complete-future rows covered) and 256,840 cooperative (89.8, 78.1, 0.138 m, 95 %).
- Hardware: 48 GB GPUs; joint stage 0.5 s per iteration single-agent and 1.4 s cooperative.