CO-REGISTERED WORLD-STATE SEQUENCES FOR COLLABORATIVE PERCEPTION AND PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Collaborative perception may rely on one fused snapshot to expand each agent's view, whereas prediction benefits from a sequence of shared world states to model how the world evolves. Each state integrates multi-agent observations from one timestamp in a common frame. Existing methods fall into two groups, neither explicitly constructing such a sequence. The first aggregates time before spatial alignment: each agent encodes its observation history and transmits the representation, which is aligned and fused across agents at the current timestamp. The second compensates for delay by extrapolating delayed messages to the current timestamp, again yielding a single fused snapshot. Both reduce historical information to a current-time representation without explicitly constructing shared states at past timestamps, and their temporal features can conflate object motion with observer motion without a common spatial reference across timestamps. A diagnostic comparison highlights room for improvement: constant-velocity extrapolation from ground-truth historical tracks matches published SOTA. This motivates co-registering and fusing multi-agent observations into a shared world state at each timestamp before temporal aggregation, so that scene dynamics can be modeled in a consistent spatial reference. Following this principle, we introduce CoReST (**Co**llaborative **Re**gistered **S**tates over **T**ime), which aligns observations from each past timestamp to the ego vehicle's coordinate frame at prediction time and fuses them into a shared world state before temporal aggregation. To support this pipeline, World-State Attention constructs each shared state through visibility-aware masking, source encoding, and pose-tolerant attention, while World-State Difference extracts motion cues from differences between consecutive states in the shared coordinate frame. On V2XPnP-Seq, CoReST achieves EPA scores of 52.8 and 52.2 in the vehicle-centric and V2V settings, respectively, surpassing the published SOTA scores of 48.2 and 40.6, while adding less than 2% to the baseline model's parameter count.

1 INTRODUCTION

Collaboration between vehicles and infrastructure expands each agent's field of view beyond its own occlusions by sharing observations. Extending this setting to prediction further allows agents to model how the shared world will evolve. Existing research has largely focused on what to transmit, when to communicate, and how to fuse the received information (Xu et al., 2022c; Hu et al., 2022; Xu et al., 2022b;a). Collaborative perception may rely on a single fused snapshot, whereas prediction benefits from a sequence of shared world states for modeling motion over time. This raises the question of how to organize observations across agents and timestamps to support motion modeling. This motivates co-registering and fusing multi-agent observations into a shared world state at each timestamp before temporal aggregation, so that scene dynamics can be modeled in a consistent spatial reference. Maintaining such states is consistent with the broader emphasis on verifiable shared representations (Hu et al., 2026). However, existing intermediate-fusion approaches do not explicitly construct this shared history.

Existing methods organize temporal collaboration in two main ways (Fig. 1a,b). One aggregates within each agent before spatial alignment: each agent encodes its history in its moving frame, and only the resulting representation is aligned across agents at the current time (Zhou et al., 2025a). The other treats time as communication delay and extrapolates delayed messages to the present (Lei et al.,

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.

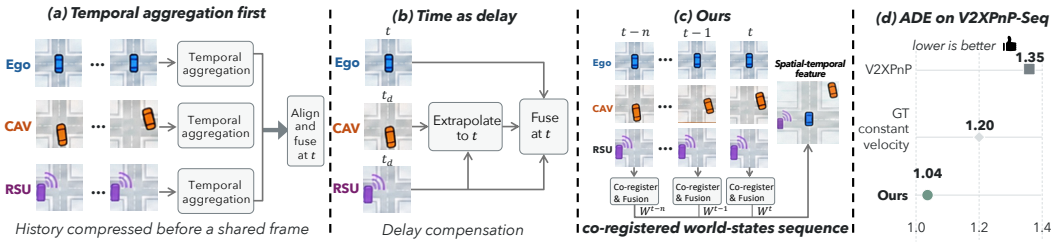

Figure 1: Temporal collaboration paradigms and ADE on V2XPnP-Seq. (a) Each agent's history is aggregated before current-time alignment and fusion. (b) Delay compensation extrapolates collaborators to t and still produces one fused snapshot. (c) CoReST co-registers and fuses observations at every timestamp into a shared world-state sequence, then aggregates it over time. (d) CoReST achieves 1.04 m ADE, versus 1.35 m for V2XPnP and 1.20 m for ground-truth constant velocity.

2022; Wei et al., 2023; Yu et al., 2023a). Aligning an agent's own historical point clouds before encoding (Yang et al., 2023) provides local temporal registration, but collaborators still contribute only to the current representation under intermediate fusion.

These approaches collapse history before or during fusion rather than construct shared historical states in a common spatial reference. Across timestamps, changes in moving frames can therefore mix object and observer motion and obscure the dynamics needed for prediction. A diagnostic comparison also highlights room for improvement: constant-velocity extrapolation from ground-truth historical tracks matches the published SOTA (Fig. 1d), with ADE/FDE of $1.20/2.44\,\mathrm{m}$ versus $1.35/2.36\,\mathrm{m}$. Although it assumes ground-truth history, this comparison suggests that collaborative models may not fully exploit temporal information and motivates constructing shared history explicitly.

Motivated by these observations, we introduce CoReST (Collaborative Registered States over Time) (Fig. 1c). At each historical timestamp, CoReST aligns all agents' observations to the ego vehicle's coordinate frame at prediction time and fuses them into a shared world state before temporal aggregation. Each state retains its original timestamp while changing only its coordinate reference. Fusion without this registration would leave successive states in different frames. This registration separates coordinate changes induced by observer motion from changes in scene content, allowing temporal reasoning over a co-registered shared history. Spatial alignment alone, however, does not resolve incomplete visibility or pose errors. Unobserved regions must be distinguished from observed free space, and fusion must tolerate residual misalignment.

To support reliable state construction and motion reasoning within this pipeline, CoReST uses *World-State Attention* (WSA) and *World-State Difference* (WSD). WSA builds each state through visibility-aware masking, source encoding, and pose-tolerant attention. The visibility mask restricts contributions to observed regions, source encoding distinguishes contributors while preserving agent exchangeability, and spatially tolerant attention accommodates pose errors. WSD computes differences between consecutive states in the current ego frame. With observer-induced coordinate changes removed, these differences encode motion cues without first estimating displacement. WSD supplies the resulting features to the trajectory head to support future motion prediction. Together, WSA and WSD add less than 2% to the parameter count.

Our contributions are as follows:

- **Temporal collaboration benefits from a co-registered shared history.** Co-registering and fusing all agents' observations at each timestamp gives temporal aggregation a sequence of shared world states in a common spatial reference, rather than separate histories in moving local frames.
- **Parameter-efficient shared-state construction and motion modeling.** WSA constructs visibility- and pose-aware shared states, while WSD extracts motion cues from consecutive aligned states. Together, they add less than 2% to the parameter count.
- **Evaluation across three datasets.** CoReST improves overall performance on real-world V2XPnP-Seq and V2X-Seq-SPD and simulated OPV2V. On V2XPnP-Seq, it reaches EPA scores of 52.8 vehicle-centric and 52.2 in V2V, surpassing the published SOTA scores of 48.2 and 40.6, respectively.

2 Related Work

Collaborative perception. Agents exchange raw points, intermediate features or decoded boxes, and the intermediate setting has held the field's attention because it trades accuracy against bandwidth most favourably (Chen et al., 2019; Xu et al., 2022c). Progress along this line is measured almost entirely by the fusion operator: a maximum over agents (Chen et al., 2019), attention over a multi-scale pyramid made robust to pose error (Lu et al., 2023), transformers that model the V2X graph and the heterogeneity of its nodes (Xu et al., 2022b;a), graph distillation (Li et al., 2021), mixtures of experts and geometry-aware attention (Kong et al., 2025; Li et al., 2025; Wang et al., 2026), confidence maps that decide what is worth sending (Hu et al., 2022; Liu et al., 2024), and alignment layers that keep the exchange working across sensors and backbones (Lu et al., 2024). These methods produce a fused view of the present, and the baseline rows of Tables 1 and 2 follow this line.

Temporal aggregation in collaborative perception. V2VNet studied joint perception and prediction over V2X (Wang et al., 2020), sequential datasets made the task measurable (Yu et al., 2023b; Zhou et al., 2025a), and subsequent work has largely come from the forecasting side: collaborative trajectory representations (Ruan et al., 2024), histories of unequal length fused across agents (Zhang et al., 2025), infrastructure knowledge at signalised intersections (Yin et al., 2026), cooperation carried through to risk and to planning (Lei et al., 2025; Yu et al., 2025), and the joint task itself with its baselines and its training schedule (Wan et al., 2026; Zhou et al., 2025b). Once prediction enters, methods must decide where to aggregate history relative to the exchange.

Representative perception-and-prediction systems organize temporal collaboration in two ways. The first aggregates history within each agent and transmits the result once, as in the one-step strategy of Zhou et al. (2025a). Time is collapsed in the agent's moving frame before agents meet at the current instant. The second treats time as communication delay. SyncNet compensates latency with a pyramid LSTM (Lei et al., 2022), FFNet transmits a feature flow the receiver extrapolates (Yu et al., 2023a), CoBEVFlow matches regions of interest across frames and warps them forward (Wei et al., 2023), and the idea now extends to two-stage motion compensation (Tian et al., 2025), to irregular sensor ticking times under sparse query models (Dao et al., 2024; Yuan & Sester, 2024), to object-level anchors carrying pose correction and temporal propagation together (Li et al., 2026), and to motion-aware or bandwidth-aware correspondence (Hong et al., 2024; Gao et al., 2025). What this group estimates is the displacement of content rather than the frame in which the grid is indexed, and its destination is typically the present. Its misalignment is caused by delay and disappears once sensors are synchronised; the one we address does not.

Methods that explicitly register history are closer to our setting. CoST projects historical features to the current frame and jointly fuses them with current agents for collaborative detection (Tang et al., 2025), without constructing a shared state sequence for prediction. SCOPE and its journal extension synchronise the *ego's own* previous point clouds to the current frame before encoding, while collaborators contribute only at the present (Yang et al., 2023; 2025). This route does not extend directly to collaborators under intermediate fusion, where historical observations arrive as features. Thus, these approaches do not construct shared historical states in a common spatial reference before temporal aggregation.

Shared representations and world states. Shared world models aim to maintain coherent representations of scene geometry, dynamics, visibility, uncertainty, and provenance across agents and time (Hu et al., 2026). Existing collaborative perception methods typically instantiate a shared representation as a fused BEV at one timestamp, leaving the construction of temporally co-registered states less explored. CoReST addresses this gap by constructing a co-registered sequence with explicit visibility and source information. This perspective is also related to generative multi-agent world modeling, which organizes information from multiple agents into shared representations (Liu et al., 2026).

3 Problem Formulation

Setting. We consider N agents indexed by $\mathcal{A} = \{1, \dots, N\}$ that cooperate via V2X communication, with agent 1 designated as the ego agent. At timestamp s, agent i has a pose $p_i^s \in \mathrm{SE}(2)$ and a sensor observation X_i^s. Let t denote the current timestamp and $\mathcal{H} = \{t-n, \dots, t-1, t\}$ the set containing

the current timestamp and the preceding n timestamps. Given the observations $\{X_i^s\}_{i\in\mathcal{A},\, s\in\mathcal{H}}$ and their associated poses, the task is to detect the objects present at t and predict a future trajectory for each object. We denote the ground-truth boxes at t by $\mathcal{B}^t$ and their associated future trajectories by $\mathcal{Y}$.

Spatiotemporal information. Observations across agents provide complementary spatial information, while observations across timestamps capture scene evolution. We represent their joint processing by a general fusion function without prescribing the order of spatial and temporal aggregation:

$$\begin{aligned} G &= \mathrm{Fusion}\big(\{X_i^s, p_i^s\}_{i\in\mathcal{A},\, s\in\mathcal{H}}\big), \\ (\hat{\mathcal{B}}^t, \hat{\mathcal{Y}}) &= \mathrm{Predict}(G), \end{aligned} \tag{1}$$

where G is the fused spatiotemporal representation, $\hat{\mathcal{B}}^t$ denotes the predicted bounding boxes at the current timestamp t, and $\hat{\mathcal{Y}}$ denotes the associated predicted future trajectories.

Objective. We learn the spatial and temporal operations jointly by minimizing the differences between the outputs and ground truth for perception and prediction:

$$\min_{\theta}\ \mathbb{E}_{\mathcal{D}}\left[\mathcal{L}_{\mathrm{perc}}\big(\hat{\mathcal{B}}^t, \mathcal{B}^t\big) + \lambda\, \mathcal{L}_{\mathrm{pred}}\big(\hat{\mathcal{Y}}, \mathcal{Y}\big)\right], \tag{2}$$

where θ denotes the model parameters, $\mathcal{D}$ is the training data distribution, and $\lambda \geq 0$ balances the two losses. The remaining question is how to design Fusion to organize information across agents and time for both perception and prediction.

4 METHOD

Overview. As shown in Fig. 2, CoReST instantiates the general fusion function in Eq. (1) as a sequence of shared world states. Its input contains observations from $n+1$ timestamps for every available agent. A shared encoder maps each one to a BEV feature F_i^s while retaining its agent and timestamp. At every timestamp, the features are transformed to the ego frame at prediction time t, and World-State Attention (WSA) fuses the co-registered observations into W^s. The sequence $W^{t-n}, \ldots, W^{t-1}, W^t$ thus describes different moments on the same spatial grid, so its variation can reflect scene dynamics without inheriting changes in the observer's coordinate frame.

A ConvGRU aggregates the shared states into the spatiotemporal feature G. The detection head reads G to localize objects at t, while the trajectory head samples features at the detected boxes to predict their future motion. In parallel, World-State Difference (WSD) projects the change between the two latest co-registered states into a residual D. Adding D only on the trajectory branch exposes recent state change without altering the feature used for detection. WSA constructs spatially consistent evidence at each time, the ConvGRU summarizes its evolution, and WSD supplements prediction with the most recent change. The complete pipeline is

$$W^s = \mathrm{WSA}\Big(\big\{\mathcal{W}_{\mathbf{T}_{(i,s)\to(1,t)}} F_i^s\big\}_{i\in\mathcal{A}}\Big), \quad s \in \mathcal{H}, \tag{3a}$$

$$G = \mathrm{ConvGRU}\big(W^{t-n}, \ldots, W^{t-1}, W^t\big), \tag{3b}$$

$$\hat{\mathcal{B}} = \mathrm{Det}(G), \tag{3c}$$

$$D = \mathrm{WSD}\big(W^t - W^{t-1}\big), \tag{3d}$$

$$\hat{\mathcal{Y}} = \mathrm{Traj}\Big(\mathrm{Sample}\Big(G + D; \hat{\mathcal{B}}\Big)\Big). \tag{3e}$$

Here $\mathcal{W}_{\mathbf{T}_{(i,s)\to(1,t)}}$ applies the pose-derived warp from agent i at s to the ego frame at t. W^s is the shared world state at timestamp s, G is the recurrent spatiotemporal representation, $\hat{\mathcal{B}}$ denotes the detected boxes, and $\hat{\mathcal{Y}}$ denotes their predicted trajectories. $\mathrm{Sample}(\cdot; \hat{\mathcal{B}})$ extracts object features at those detections; the detection path receives G, whereas the trajectory path receives $G + D$.

Co-register and fuse at each timestamp. Each F_i^s is initially expressed in the coordinate frame of agent i at timestamp s. Before cross-agent fusion, CoReST warps it with $\mathbf{T}_{(i,s)\to(1,t)}$, composed from the logged poses p_i^s and p_1^t. Consequently, every agent's observation at every historical timestamp

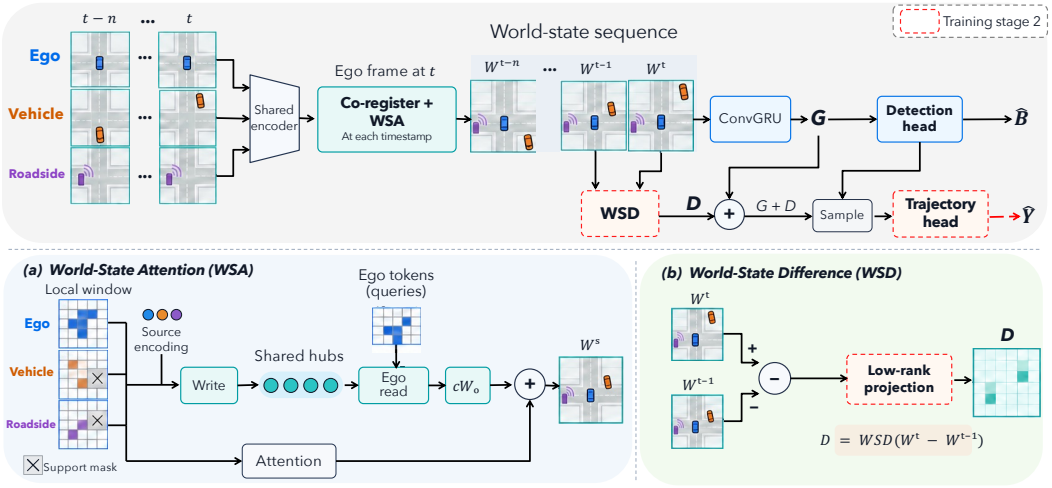

Figure 2: CoReST pipeline and two-stage training. Co-registered multi-agent features are fused by WSA into a world state W^s at every timestamp, and a ConvGRU aggregates the sequence into G for both task heads. WSD projects the change between W^{t-1} and W^t into D, which augments only the trajectory branch. Lower panels detail visibility- and source-aware WSA and low-rank WSD; dashed red outlines mark the modules optimized in Stage 2.

is expressed in the ego vehicle's coordinate frame at prediction time. The transformed feature still describes the scene at timestamp s; only its coordinate reference changes. Moving objects retain their displacement across timestamps in this common reference.

At each timestamp, WSA fuses the co-registered agent features into W^s. Repeating this operation gives $W^{t-n}, \ldots, W^{t-1}, W^t$ on one grid, where a cell refers to the same physical location throughout the sequence. This makes temporal aggregation and the difference $W^t - W^{t-1}$ spatially meaningful. The implementation replaces the current relative pose $\mathbf{T}_{(i,t)\to(1,t)}$ in the existing warp with $\mathbf{T}_{(i,s)\to(1,t)}$, adding neither parameters nor another resampling operation. The remaining step at each timestamp is to combine the aligned observations into a shared state.

World-State Attention. Co-registration gives observations a common spatial reference, but their valid coverage and residual pose error still affect fusion. WSA constructs each state by pooling valid evidence within local windows and retaining its agent source in the attention keys. For a fixed timestamp, write $x_i = \mathcal{W}_{\mathbf{T}_{(i,s)\to(1,t)}} F_i^s$ for the co-registered feature of agent i. The base fusion path applies parameter-free attention over agents at each cell and reads its ego row. Its softmax can assign weight to zero padding outside an agent's valid support, and its cell-wise comparison is sensitive to residual pose error. WSA augments this path with a residual that uses support masks and nearby spatial evidence.

Warping an all-ones map with the same transform produces the support mask M_i of agent i. Within each window, learnable hub tokens Z collect the multi-agent evidence before the ego reads it. The agent tokens X write into these hubs as

$$H = \mathrm{Attn}(Z, K_{\mathrm{src}}(X), V(X); M)\,. \tag{4}$$

Here Attn denotes scaled dot-product attention with queries, keys, and values in that order; M collects the support masks and excludes unsupported tokens before softmax. K_{src} projects the tokens to keys and applies a fixed source rotation $R_{\sigma(i)}$ to their first half (Liu et al., 2026), while V projects their values. The rotation uses equidistant regular-simplex vertices to distinguish contributors without learned agent slots. The ego uses vertex 0; collaborators receive a random injective assignment during training and their index order at test time.

The ego tokens X_1 then read the aggregated hubs H to form the state within the window:

$$W^s = \mathrm{Att}(x) + cW_o\,\mathrm{Attn}(Q(X_1), K(H), H)\,, \tag{5}$$

where Q and K are query and key projections, Att is the base fusion path, W_o is zero-initialised, and the binary gate c is active only when a collaborator has support in the window. Pooling within a

window lets the ego cell read evidence written at nearby positions, providing tolerance to residual pose error. Applying this operation across windows forms W^s; repeating it at each timestamp produces the state sequence for temporal aggregation.

We use four hubs with attention width $C/4$ and window sizes $8, 4, 2$ across the three pyramid scales, each spanning 6.4 m. The gate preserves the base output in single-agent scenes, and zero initialization preserves it at the start of training. The projections and hub tokens add 216k parameters, or 1.8% of the model.

World-State Difference. The ConvGRU in Eq. (3b) aggregates the co-registered state sequence into G, which supports both detection and trajectory prediction. This recurrent representation is jointly shaped by the two task objectives. The common spatial reference also allows the latest states to be compared directly: $W^t - W^{t-1}$ describes feature changes at the same physical locations. WSD uses this difference to provide the trajectory head with an explicit state-change cue alongside the aggregated history.

We implement WSD as the rank-32 projection $\mathrm{WSD}(\Delta W) = \mathbf{U}\mathbf{V}\Delta W$, with $\mathbf{V} \in \mathbb{R}^{32\times 256}$, $\mathbf{U} \in \mathbb{R}^{256\times 32}$, and $\mathbf{U}_0 = 0$. The implementation also contains a 32×32 mixing layer folded into $\mathbf{V}$, for 17.7k parameters in total. The detection head predicts $\hat{\mathcal{B}}$ from G. The trajectory head samples $G + D$ at these detected boxes, as shown in Eq. (3e), combining the aggregated history with the latest state change. The residual enters after the ConvGRU and is routed only to the trajectory head, so it directly supplements prediction while detection uses G.

Training. The detection objective combines focal classification, smooth-ℓ_1 box regression, and direction classification. The trajectory head predicts three modes $\{\hat{\mathcal{Y}}_m\}_{m=1}^3$ with scores $\boldsymbol{\pi}$. We select $m^* = \arg\min_m \mathcal{L}_{\mathrm{reg}}(\hat{\mathcal{Y}}_m, \mathcal{Y})$ and train that mode together with a soft scoring target $\mathbf{q}$. The task and stage objectives are

$$\mathcal{L}_{\mathrm{det}} = \mathcal{L}_{\mathrm{focal}} + \lambda_{\mathrm{box}}\mathcal{L}_{\text{smooth-}\ell_1} + \lambda_{\mathrm{dir}}\mathcal{L}_{\mathrm{dir}}, \tag{6a}$$

$$\mathcal{L}_{\mathrm{pred}} = \mathcal{L}_{\mathrm{reg}}(\hat{\mathcal{Y}}_{m^*}, \mathcal{Y}) + \lambda_{\mathrm{score}}\mathcal{L}_{\mathrm{score}}(\boldsymbol{\pi}, \mathbf{q}), \tag{6b}$$

$$\mathcal{L}_{\mathrm{S1}} = \mathcal{L}_{\mathrm{det}} + \lambda_{\mathrm{pred}}\mathcal{L}_{\mathrm{pred}}, \tag{6c}$$

$$\mathcal{L}_{\mathrm{S2}} = \mathcal{L}_{\mathrm{pred}}. \tag{6d}$$

Stage 1 optimizes the encoder, WSA, temporal aggregator, and both task heads end to end under Eq. (6c). The zero-initialised WSA output projection makes optimization begin from the base fusion path. Stage 2 loads this checkpoint and optimizes only WSD and the trajectory head under Eq. (6d); all other parameters and batch-normalization statistics are frozen. Consequently, the detected boxes, matched set, and recall window remain those of Stage 1, making the trajectory comparison between rows (g) and (h) of Table 3 paired. Because the zero-initialised WSD output projection initially blocks gradients to the preceding projection, $\mathbf{V}$ is excluded from weight decay.

5 Experiments

5.1 Implementation Details

Datasets. We evaluate on three collaborative driving datasets spanning real-world and simulated data. They cover different cities and environments, providing diversity in road layouts and traffic conditions. **V2XPnP-Seq** (Zhou et al., 2025a) is real-world V2X data with up to two vehicles and two roadside units per scenario. Release 1.0 publishes 68 of the 96 scenarios used in the original paper, split $48/6/14$ into train, validation, and test; we use the published test split and report the vehicle-centric (VC) and vehicle-to-vehicle (V2V) collaboration modes. **V2X-Seq-SPD** (Yu et al., 2023b) provides real-world vehicle-to-infrastructure data, while **OPV2V** (Xu et al., 2022c) provides simulated vehicle-to-vehicle data. All three datasets are recorded at 10 Hz.

Settings. Each sample uses 2 s of history and a 3 s prediction horizon at 2 Hz, corresponding to five past frames and six future steps. Point clouds are voxelised at 0.4 m over the training range $x \in [-102.4, 102.4]$ and $y \in [-51.2, 51.2]$ m. We train for 20 epochs with Adam, a batch size of 2, an initial learning rate of 10^{-3}, weight decay of 5×10^{-4}, and a cosine schedule with one

Table 1: V2XPnP-Seq results under vehicle-centric (VC) and vehicle-to-vehicle (V2V) collaboration. AP_{50} measures detection, ADE/FDE/MR top-1 forecasting, and EPA the joint task. E2E/Map denote end-to-end training/HD-map input; dark/light cells mark the best/second-best results.

			V2XPnP-Seq-VC					V2XPnP-Seq-V2V				
Method	E2E	Map	AP_{50} ↑	ADE↓	FDE↓	MR↓	EPA↑	AP_{50} ↑	ADE↓	FDE↓	MR↓	EPA↑
No Fusion	✓		43.9	1.87	3.24	33.8	24.3	40.8	1.99	3.38	34.0	19.8
No Fusion-FaF	✓		53.4	1.55	2.81	34.3	31.6	51.9	1.67	3.12	39.3	27.5
Late Fusion		✓	58.1	1.59	2.82	32.4	33.0	55.3	1.75	3.07	34.0	30.5
Early Fusion	✓	✓	60.3	1.37	2.49	33.8	36.7	53.0	1.64	3.11	40.2	26.9
CoBEVFlow	✓	✓	63.3	1.36	2.49	33.0	41.9	58.7	1.72	3.15	40.3	33.6
FFNet	✓	✓	64.6	1.36	2.47	34.7	42.3	56.5	1.68	3.12	39.8	31.2
V2X-ViT	✓	✓	69.6	1.39	2.56	35.2	44.7	64.6	1.68	3.13	39.8	36.7
V2XPnP	✓	✓	71.6	1.35	2.36	31.7	48.2	70.5	1.78	3.28	39.9	40.6
CoReST (Ours)	✓	✓	**82.0**	**1.05**	**1.95**	**28.0**	**52.8**	**81.6**	**1.07**	**2.00**	**28.7**	**52.2**

warm-up epoch and a minimum learning rate of 2×10^{-6}. The checkpoint with the lowest validation loss is used for evaluation. The communication radii are 50, 200, and 70 m, and the maximum numbers of agents are 4, 2, and 5 for V2XPnP, V2X-Seq, and OPV2V, respectively. The evaluation ranges follow the dataset protocols: $x \in [-70, 70]$ for V2XPnP, $x \in [-100, 100]$ for V2X-Seq, and $x \in [-102.4, 102.4]$ for OPV2V, with $y \in [-40, 40]$ m for all three. We retain the vehicle classes Car, Van, Truck, and Bus for training and evaluation.

Metrics. Detection is evaluated by frame-wise average precision at IoU thresholds 0.5 and 0.7. AP summarizes current-frame localization, whereas the trajectory metrics evaluate future object-center motion after detection matching. For matched objects, ADE is the mean Euclidean displacement over the six predicted steps, FDE is the displacement at the final step, and MR is the percentage of predictions whose FDE exceeds 2 m. Following Zhou et al. (2025a), we also report end-to-end perception and prediction accuracy. EPA counts an object as a hit only when detection matching and the trajectory displacement criterion are both satisfied, and penalizes false positives:

$$\mathrm{EPA} = \frac{|\hat{S}_{\mathrm{match,hit}}| - \alpha N_{\mathrm{FP}}}{N_{\mathrm{GT}}},$$

with $\tau_{\mathrm{EPA}} = 2$ m and $\alpha = 0.5$, where $\hat{S}_{\mathrm{match,hit}}$ contains matched detections whose predictions satisfy the displacement threshold, and N_{FP} and N_{GT} are the numbers of false-positive and ground-truth objects. For multimodal predictions, top-1 is the mode with the highest predicted score; single-mode baselines contribute their only forecast. In every table, the best value in each column is bold on a grey cell and the runner-up is shown on a lighter cell.

Baselines. On V2XPnP-Seq, we use the results reported by Zhou et al. (2025a) in Table 1, including No Fusion-FaF, FFNet, V2X-ViT, and V2XPnP. For V2X-Seq and OPV2V, we reimplement all baselines and train them under the settings above. We compare no fusion, box-level late fusion, point-level early fusion, F-Cooper (Chen et al., 2019), CoAlign (Lu et al., 2023), and CoBEVFlow (Wei et al., 2023). No Fusion uses only ego observations; Early Fusion combines point clouds before encoding, Late Fusion combines decoded boxes, and F-Cooper, CoAlign, and CoBEVFlow operate on intermediate features. CoAlign targets pose error, while CoBEVFlow compensates temporal misalignment. In Table 1, E2E denotes end-to-end training and Map denotes HD-map use.

5.2 Experimental Results

Main results. Across all four evaluation settings, CoReST achieves the best ADE, FDE, MR, and EPA. On V2XPnP-Seq, it also leads detection, improving AP_{50} over V2XPnP by 10.4 points in VC mode and 11.1 points in V2V mode, while reducing ADE by 22% and 40%. On V2X-Seq-SPD and OPV2V, it reduces ADE by 26% and 45% despite CoBEVFlow's 0.7- and 0.2-point AP_{50} leads. Its EPA gains over the best baseline are 4.6, 11.6, 12.5, and 11.0 points on V2XPnP-Seq-VC, V2XPnP-Seq-V2V, V2X-Seq-SPD, and OPV2V, respectively.

Table 2: Cross-dataset results on real-world V2X-Seq-SPD and simulated OPV2V, covering vehicle-to-infrastructure and vehicle-to-vehicle collaboration. CoReST leads trajectory and joint-task metrics on both while remaining close in AP.

	V2X-Seq-SPD						OPV2V					
Method	AP_{50} ↑	AP_{70} ↑	ADE↓	FDE↓	MR↓	EPA↑	AP_{50} ↑	AP_{70} ↑	ADE↓	FDE↓	MR↓	EPA↑
No Fusion	32.9	18.1	2.88	4.86	54.2	5.9	68.2	40.7	5.33	9.46	85.0	−2.5
Late Fusion	21.8	10.6	4.56	7.64	65.6	−38.7	71.5	59.6	4.42	7.92	82.9	−6.5
Early Fusion	32.7	13.6	3.47	5.88	58.5	−1.0	85.9	73.2	4.26	7.66	81.4	4.7
F-Cooper	39.5	17.2	2.58	4.35	49.4	14.4	86.0	69.1	3.83	6.88	79.9	8.0
CoAlign	43.9	23.5	2.63	4.44	51.8	19.4	88.1	77.6	3.90	7.01	81.3	9.3
CoBEVFlow	**48.9**	**27.8**	3.38	5.66	54.6	20.3	**90.1**	**82.3**	4.12	7.33	82.7	9.2
CoReST (Ours)	48.2	26.7	**1.91**	**3.24**	**42.2**	**32.8**	89.9	81.6	**2.11**	**4.28**	**68.5**	**20.3**

Table 3: Component ablation on V2XPnP-Seq-VC. Rows (a)–(b) test co-registration, (c)–(g) build WSA, and (h) adds WSD to the frozen detector. Δpar. is relative to (b); ADE* is oracle over modes.

		Δpar.	AP_{50} ↑	AP_{70} ↑	ADE↓	FDE↓	MR↓	ADE* ↓	EPA↑
(a)	No shared reference across time	0	78.4	55.8	2.504	4.349	42.1	2.504	42.8
(b)	**Co-registered fusion**	0	80.3	59.1	1.248	2.316	31.3	0.924	**52.9**
World-State Attention									
(c)	(b) + visibility gate on the write	0	80.0	62.3	1.136	2.137	30.5	0.901	52.4
(d)	(c) + hub write and read, K=4	216k	81.5	63.9	1.145	2.153	31.5	0.913	51.0
(e)	(d) + 6.4 m read window	216k	**82.4**	62.2	1.122	2.086	29.5	0.877	**52.9**
(f)	(d) + writer signature	216k	81.7	63.3	1.122	2.101	30.9	0.869	52.7
(g)	**WSA** = (d) + window + signature	216k	82.0	**65.1**	**1.038**	**1.953**	28.5	0.807	52.5
World-State Difference, over the frozen detector of (g)									
(h)	**(g) + WSD** (full model)	234k	82.0	**65.1**	1.047	**1.953**	**28.0**	**0.795**	52.8

Ablation results. Co-registration gives the largest parameter-free gain (ADE 2.504→1.248; EPA 42.8→52.9). WSA's window and signature reach 1.038 ADE and 65.1 AP_{70}; WSD preserves AP and improves MR and ADE* with about 18k parameters.

Robustness to pose error and latency. We perturb the V2X-Seq-SPD validation set and evaluate the checkpoints of Table 2 without retraining. Pose error adds Gaussian noise of standard deviation σ to the infrastructure position (m) and yaw (°), while the ground-truth ego pose stays exact. A latency of t ms replaces the infrastructure point cloud and calibration with those recorded k frames earlier, k uniform in $\{1, \ldots, t/100\}$ at 10 Hz. In Fig. 3, axes denote perturbation levels, dashed circles mark each method's clean EPA, and the solid circle marks EPA = 0. The radius is linear except that two ranges containing no data are compressed, and Late Fusion, below the range at every level, is omitted. Our model degrades gradually under both perturbations: pose error up to $\sigma = 0.4$ costs it 1.7 EPA points, and each additional 100 ms of latency costs at most 1.1. At the strongest levels it keeps 78% ($\sigma = 0.8$) and 88% (500 ms) of its clean EPA and still exceeds the clean EPA of every baseline. These trends are consistent with WSA's spatial tolerance and show that the co-registered state sequence remains useful as collaborator observations become noisy or stale.

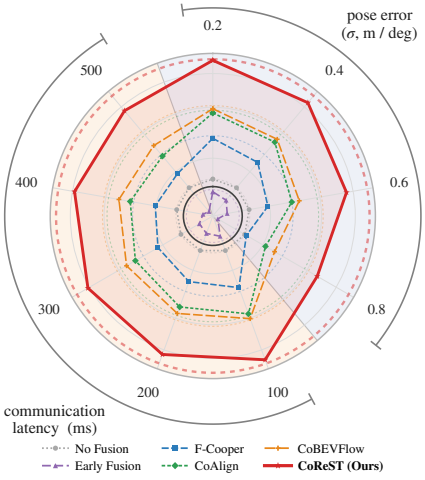

Figure 3: V2X-Seq-SPD EPA under pose error (upper right) and latency (lower left).

5.3 VISUALIZATION

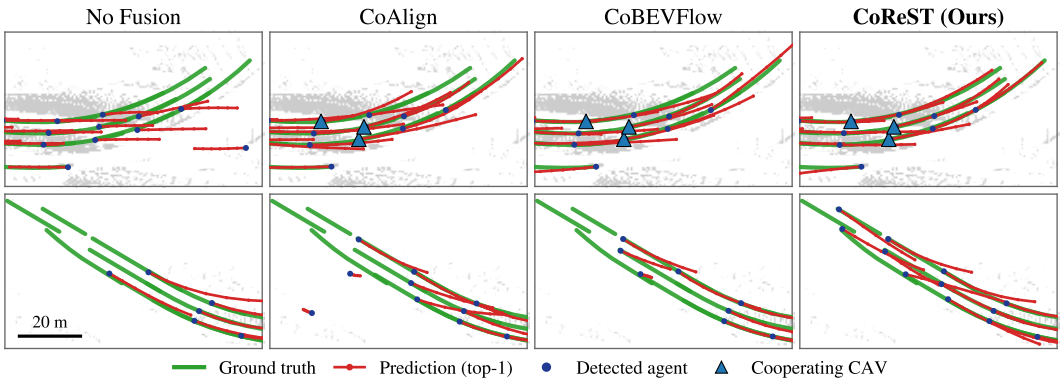

Figure 4: Top-1 forecasts for two OPV2V scenes (rows), using a common crop. Columns compare No Fusion, CoAlign, CoBEVFlow, and CoReST. Green/red lines denote ground truth/prediction; blue dots mark detections and triangles cooperating vehicles.

Figure 4 shows two OPV2V test scenes with the checkpoints of Table 2, each cropped to the same 80×48 m window for every method. Only the top-1 forecast is drawn, since the baselines predict a single mode. In the first scene, where the road bends ahead of the ego, every method detects all ten vehicles in the window. No Fusion extrapolates them in straight lines through the bend, and CoAlign and CoBEVFlow cut across it, while our forecasts follow the curve, with a top-1 ADE of 1.5 m over these vehicles against 3.2 and 3.6 m. In the second scene, behind the ego, No Fusion misses 7 of the 13 vehicles in the window, CoAlign and CoBEVFlow miss 5 each and our model misses 1. This detection gap is specific to the scene: over the whole test set our AP@0.5 is within 0.2 points of CoBEVFlow (Table 2).

6 CONCLUSION

We studied how multi-agent observations should be organized across space and time for collaborative perception and prediction. CoReST aligns and fuses all agents' observations at each timestamp into a shared world state before aggregating the resulting sequence over time. WSA constructs each state while accounting for visibility, information source, and pose error, and WSD exposes state changes directly to the trajectory head. Experiments across three real-world and simulated datasets show consistent improvements in trajectory prediction and end-to-end accuracy, with co-registration providing the largest gain without adding parameters. These results support co-registered world-state sequences as an effective representation for collaborative prediction.

Limitations and Future Work. CoReST is not implemented as a streaming model and repeatedly processes the history at each prediction step, resulting in redundant communication and computation. Future work will explore streaming world modeling with incremental state updates to better support real-world deployment.

AI USE STATEMENT

In this work, we used generative AI tools for writing polish and grammar correction. We did not use generative AI tools to generate research ideas, develop methods, write or execute code, conduct experiments, analyze results, or formulate scientific claims. The authors reviewed the AI-assisted text for technical accuracy and consistency with the intended meaning. We take responsibility for the final content of this work, including material produced with the aid of generative AI.

REFERENCES

Qi Chen, Xu Ma, Sihai Tang, Jingda Guo, Qing Yang, and Song Fu. F-Cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3D point clouds. In *ACM/IEEE Symposium on Edge Computing (SEC)*, 2019.

Minh-Quan Dao, Julie Stephany Berrio, Vincent Frémont, Mao Shan, Elwan Héry, and Stewart Worrall. Practical collaborative perception: A framework for asynchronous and multi-agent 3D object detection. *IEEE Transactions on Intelligent Transportation Systems*, 2024.

Peng Gao, Williard Jose, and Hao Zhang. Bandwidth-adaptive spatiotemporal correspondence identification for collaborative perception. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

Shixin Hong, Yu Liu, Zhi Li, Shaohui Li, and You He. Multi-agent collaborative perception via motion-aware robust communication network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Senkang Hu, Zhengru Fang, Yihang Tao, Zihan Fang, Sam Tak Wu Kwong, and Yuguang Fang. Multi-agent embodied autonomous driving: From V2X information exchange to shared world models. *arXiv preprint arXiv:2606.13840*, 2026.

Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Lingzhao Kong, Jiacheng Lin, Siyu Li, Kai Luo, Zhiyong Li, and Kailun Yang. CoBEVMoE: Heterogeneity-aware feature fusion with dynamic mixture-of-experts for collaborative perception. *arXiv preprint arXiv:2509.17107*, 2025.

Mingyue Lei, Zewei Zhou, Hongchen Li, Jia Hu, and Jiaqi Ma. CooperRisk: A driving risk quantification pipeline with multi-agent cooperative perception and prediction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

Zixing Lei, Shunli Ren, Yue Hu, Wenjun Zhang, and Siheng Chen. Latency-aware collaborative perception. In *European Conference on Computer Vision (ECCV)*, 2022.

Chi Li, Rui Lin, Aobo Ji, and Dongzhu Xu. CoAnchor: Robust collaborative perception under spatio-temporal misalignment via object-level anchors. In *ACM International Conference on Multimedia (MM)*, 2026.

Lantao Li, Kang Yang, Wenqi Zhang, Xiaoxue Wang, and Chengyong Sun. RG-Attn: Radian glue attention for multi-modal multi-agent cooperative perception. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2025.

Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, and Wenjun Zhang. Learning distilled collaboration graph for multi-agent perception. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Fangfu Liu, Kai He, Tianchang Shen, Tianshi Cao, Sanja Fidler, Yueqi Duan, Jun Gao, Igor Gilitschenski, Zian Wang, and Xuanchi Ren. Gamma-world: Generative multi-agent world modeling beyond two players. *arXiv preprint arXiv:2605.28816*, 2026.

Yuntao Liu, Qian Huang, Rongpeng Li, Xianfu Chen, Zhifeng Zhao, Shuyuan Zhao, Yongdong Zhu, and Honggang Zhang. Select2Col: Leveraging spatial-temporal importance of semantic information for efficient collaborative perception. *IEEE Transactions on Vehicular Technology*, 2024.

Yifan Lu, Quanhao Li, Baoan Liu, Mehrdad Dianati, Chen Feng, Siheng Chen, and Yanfeng Wang. Robust collaborative 3D object detection in presence of pose errors. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

Yifan Lu, Yue Hu, Yiqi Zhong, Dequan Wang, Siheng Chen, and Yanfeng Wang. An extensible framework for open heterogeneous collaborative perception. In *International Conference on Learning Representations (ICLR)*, 2024.

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.

Zongheng Tang, Yi Liu, Yifan Sun, Yulu Gao, Jinyu Chen, Runsheng Xu, and Si Liu. CoST: Efficient collaborative perception from unified spatiotemporal perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1120–1129, 2025.

Chengchang Tian, Jianwei Ma, Yan Huang, Zhanye Chen, Honghao Wei, Hui Zhang, and Wei Hong. DATA: Domain-and-time alignment for high-quality feature fusion in collaborative perception. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Lei Wan, Hannan Ejaz Keen, and Alexey Vinel. Towards collaborative joint perception and prediction: Framework, baseline evaluation, and deployment perspectives. *arXiv preprint arXiv:2608.09541*, 2026.

Jinlong Wang, Yuang Jia, Junhong Lin, Nannan Li, and Wei Gao. CoDS: Robust collaborative perception via expert-driven detection and BEV segmentation. *arXiv preprint arXiv:2608.14085*, 2026.

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 *European Conference on Computer Vision (ECCV)*, 2020.

Sizhe Wei, Yuxi Wei, Yue Hu, Yifan Lu, Yiqi Zhong, Siheng Chen, and Ya Zhang. Asynchrony-robust collaborative perception via bird's eye view flow. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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

Kun Yang, Dingkang Yang, Jingyu Zhang, Mingcheng Li, Yang Liu, Jing Liu, Hanqi Wang, Peng Sun, and Liang Song. Spatio-temporal domain awareness for multi-agent collaborative perception. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Kun Yang, Zhi Xu, Dingkang Yang, Qiang Fu, Rui Tang, Liang Song, and Lihua Zhang. Robust multi-agent collaborative perception via spatio-temporal awareness. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Huilin Yin, Yangwenhui Xu, Jiaxiang Li, Hao Zhang, and Gerhard Rigoll. Knowledge-informed multi-agent trajectory prediction at signalized intersections for infrastructure-to-everything. *IEEE Transactions on Intelligent Transportation Systems*, 27(7):8480–8494, 2026. doi: 10.1109/TITS.2026.3673925.

Haibao Yu, Yingjuan Tang, Enze Xie, Jilei Mao, Ping Luo, and Zaiqing Nie. Flow-based feature fusion for vehicle-infrastructure cooperative 3D object detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Haibao Yu, Wenxian Yang, Hongzhi Ruan, Zhenwei Yang, Yingjuan Tang, Xu Gao, Xin Hao, Yifeng Shi, Yifeng Pan, Ning Sun, Juan Song, Jirui Yuan, Ping Luo, and Zaiqing Nie. V2X-Seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023b.

Haibao Yu, Wenxian Yang, Jiaru Zhong, Zhenwei Yang, Siqi Fan, Ping Luo, and Zaiqing Nie. End-to-end autonomous driving through V2X cooperation. In *AAAI Conference on Artificial Intelligence*, 2025.

Yunshuang Yuan and Monika Sester. StreamLTS: Query-based temporal-spatial LiDAR fusion for cooperative object detection. In *European Conference on Computer Vision Workshops (ECCVW)*, 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)*, 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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025a.

Zewei Zhou, Seth Z. Zhao, Tianhui Cai, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. TurboTrain: Towards efficient and balanced multi-task learning for multi-agent perception and prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025b.