HYPERSYNC: HYPERBOLIC FEATURE MATCHING FOR MULTI-AGENT SPATIAL SYNCHRONIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

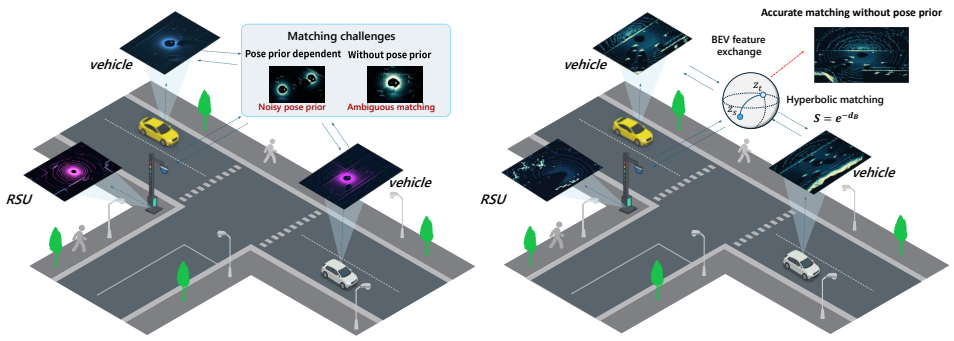
Multi-agent spatial synchronization is a critical prerequisite for cooperative perception and information fusion in autonomous driving. Existing methods typically establish cross-agent feature correspondences using metrics such as inner products and bilinear similarities in euclidean feature spaces, or perform association using object trajectories and motion-consistency constraints. When substantial viewpoint differences, incomplete observations, or only a limited number of co-visible objects exist among different agents, existing methods have limited ability to distinguish true correspondences, false correspondences, and outliers, making them prone to erroneous associations and consequently degrading the stability of relative pose estimation and spatial alignment. To address these challenges, we propose HyperSync, a multi-agent spatial synchronization method based on hyperbolic-space representations. By exploiting the negative curvature and nonlinear distance properties of hyperbolic space, HyperSync maps candidate matching features extracted from different agents into the Poincaré ball model and constructs a cross-agent feature similarity metric based on hyperbolic distance, thereby enhancing the discriminability of feature correspondences under weak-overlap and sparse-co-visibility conditions. On this basis, spatial alignment among multi-agent observations is achieved through correspondence filtering and relative pose estimation. The proposed method is evaluated on the OPV2V, V2X-Sim and DAIR-V2X datasets. Experimental results demonstrate that, compared with existing methods, the proposed method achieves more accurate feature matching and relative pose estimation and exhibits better robustness under substantial viewpoint differences and conditions with only a limited number of co-visible objects.

1 INTRODUCTION

Multi-agent cooperative perception extends the sensing range of individual agents and improves the perception of occluded and distant objects by exchanging observations among vehicles, roadside units, and other sensing nodes Liu et al. (2023); Wang et al. (2020); Hu et al. (2022); Xu et al. (2022a). Since each observation is represented in the agent's local coordinate system, accurate relative rotation and translation estimation is essential for cross-agent feature alignment and fusion.

Existing systems typically obtain agent poses from global navigation satellite systems (GNSS), inertial measurement units (IMUs), high-definition maps, or pre-calibrated extrinsics. However, signal blockage and multipath propagation in urban canyons, tunnels, elevated roads, and densely built environments can substantially degrade localization accuracy He et al. (2023); Lei et al. (2024). Such errors propagate through coordinate transformations, leading to feature displacement, duplicated responses, and unreliable fusion. As illustrated in Figure 1, existing cross-agent matching schemes face two typical challenges. Pose-prior-dependent methods can align observations when the prior is accurate, but noisy poses directly introduce spatial misalignment and imprecise correspondences. In contrast, pose-prior-free methods recover relative relationships directly from cross-agent observations, yet may suffer from ambiguous matching under large viewpoint differences, incomplete observations, and sparse co-visibility. Therefore, establishing accurate and stable cross-agent correspondences without relying on reliable localization priors is a key challenge for spatial synchronization.

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.



(a) Matching challenges

(b) HyperSync:Accurate matching without pose prior

Figure 1: Perception data are exchanged among multiple agents, and hyperbolic matching enables spatial synchronization without localization priors.

Existing spatial synchronization methods mainly rely on point-cloud registration, object-level association, or feature-level matching. Point-cloud registration requires sufficient geometric overlap and often incurs considerable communication and computation costs Rusinkiewicz & Levoy (2001); Yang et al. (2020); Huang et al. (2021). Object-level methods estimate relative poses from bounding boxes, trajectories, or structural relationships among co-visible objects Shi et al. (2022); Song et al. (2024); Lu et al. (2022); Lei et al. (2024), but their performance depends strongly on detection quality and the number and distribution of shared objects. Intermediate bird's-eye-view (BEV) features provide denser matching cues by encoding object structure, local geometry, and scene context Hao et al. (2024); Xu et al. (2026); Ni et al. (2024).

Most feature-level methods establish correspondences in Euclidean space using inner products, cosine similarity, or bilinear transformations. Under large viewpoint variations, occlusion, sparse overlap, and repetitive road structures, however, true correspondences and hard negatives may exhibit similar Euclidean similarities. The negative curvature and nonlinear geodesic distance of the Poincaré ball provide a more flexible metric geometry for separating such descriptors Nickel & Kiela (2017); Ganea et al. (2018). Nevertheless, reliable synchronization also requires accurate correspondence locations and robust pose estimation, since discrete BEV matching introduces quantization errors and a few high-confidence outliers may cause severe pose failures.

To address these challenges, we propose HyperSync, a hyperbolic feature matching framework for multi-agent spatial synchronization. HyperSync constructs pair-conditioned BEV descriptors and embeds them into a Poincaré ball, where hyperbolic geodesic distance is used to establish cross-agent correspondences. Local matching responses are then exploited to refine discrete candidate locations into continuous coordinates. Finally, geometric consistency filtering, robust special Euclidean group in two dimensions ($\mathrm{SE}(2)$) estimation, and pose self-consistency evaluation are combined to suppress outliers and reject unreliable synchronization results.

The main contributions are summarized as follows:

- We introduce hyperbolic matching for multi-agent BEV features by embedding pair-conditioned descriptors into a Poincaré ball and replacing Euclidean bilinear similarity with hyperbolic geodesic distance.
- We propose a sub-pixel refinement strategy that recovers continuous correspondence coordinates from local matching responses, reducing BEV-grid quantization errors.
- We develop a robust synchronization procedure combining geometric consistency filtering, $\mathrm{SE}(2)$ pose estimation, and pose self-consistency evaluation to suppress false correspondences and long-tailed failures.
- Experiments on OPV2V, V2X-Sim, and DAIR-V2X demonstrate accurate and robust relative pose estimation in simulated vehicle-to-vehicle and real-world vehicle-to-infrastructure scenarios.

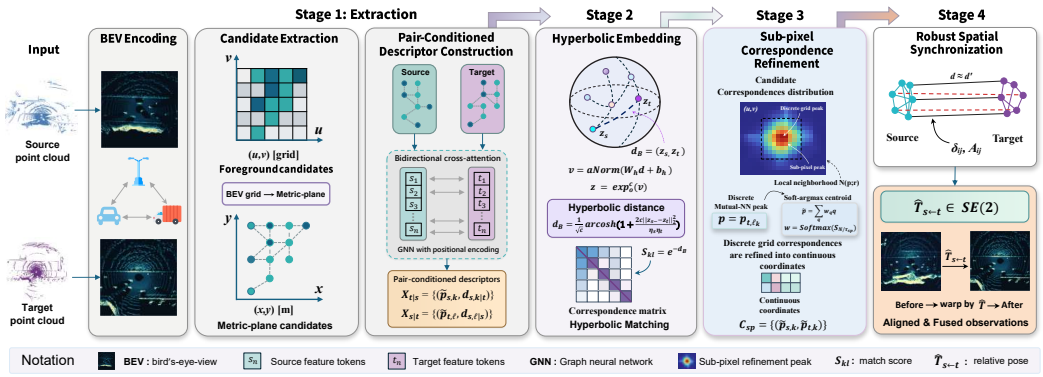


Figure 2: Overall framework of HyperSync. Shared BEV encoding produces source and target features, followed by foreground candidate extraction and pair-conditioned descriptor construction. The descriptors are matched in a Poincaré ball, refined to continuous correspondences, and filtered geometrically for robust $\mathrm{SE}(2)$ estimation and cross-agent alignment.

2 Related Work

2.1 Multi-Agent Spatial Synchronization and Relative Pose Estimation

When global poses are unavailable or unreliable, relative transformations can be estimated directly from multi-agent observations. Existing methods can be broadly categorized into point-cloud registration, object-level association, and feature-level matching.

Point-cloud registration estimates rigid transformations from geometric correspondences. Classical iterative closest point (ICP) methods iteratively update nearest-neighbor associations Rusinkiewicz & Levoy (2001), while learning-based approaches such as DCP, Predator, and PEAL improve correspondence estimation under challenging conditions Wang & Solomon (2019); Huang et al. (2021); Yu et al. (2023). Robust estimators including TEASER++ and Quatro further suppress outliers Yang et al. (2020); Lim et al. (2022). However, these methods generally require sufficient geometric overlap and incur substantial communication and computation costs when raw point clouds are exchanged.

Object-level methods estimate relative poses from bounding boxes, trajectories, or spatial relationships among co-visible objects. Representative approaches use object trajectories for roadside calibration Ren et al. (2023), associate commonly observed objects across vehicle and infrastructure sensors Qu et al. (2024), or exploit object-level geometric consistency for pose recovery and feature alignment Shi et al. (2022); Song et al. (2024); Lu et al. (2022); Lei et al. (2024). Although communication-efficient, these methods depend strongly on detection quality and the number and spatial distribution of co-visible objects, making them vulnerable to sparse traffic, severe occlusion, and limited overlap.

Intermediate bird's-eye-view (BEV) features provide denser matching cues by encoding object structure, local geometry, and scene context. Recent methods construct BEV keypoint or foreground descriptors and combine learned similarity with geometric consistency Xu et al. (2026); Hao et al. (2024); Ni et al. (2024). Nevertheless, viewpoint variations, uneven point density, repetitive structures, and partial occlusion can still produce ambiguous correspondences. Discrete BEV grids also introduce coordinate quantization errors, while a small number of high-confidence outliers may cause substantial pose deviations. These limitations motivate more discriminative matching metrics and more robust correspondence-to-pose estimation.

2.2 Hyperbolic Representation Learning and Feature Matching

Most visual representations are learned in Euclidean space and compared using inner products, cosine similarity, or Euclidean distance. Euclidean geometry is computationally convenient, but its

spatial capacity grows polynomially with radius. In contrast, hyperbolic space has constant negative curvature and exponentially expanding volume, offering an alternative geometry for modeling non-uniform and complex neighborhood structures.

Nickel and Kiela introduced Poincaré embeddings for hierarchical representation learning Nickel & Kiela (2017). Ganea et al. subsequently established Möbius operations and exponential and logarithmic maps for hyperbolic neural networks Ganea et al. (2018). Hyperbolic geometry has since been extended to graph representation learning Chami et al. (2019) and visual embedding Khrulkov et al. (2020). Related studies show that nonlinear geodesic metrics can reshape local distance distributions and improve the separation of hard negatives and closely distributed samples Yue et al. (2024); Khrulkov et al. (2020).

The Poincaré ball is a conformal model in which features are constrained within an open ball. Its geometry is approximately Euclidean near the origin, whereas geodesic distance becomes increasingly sensitive near the boundary. This nonlinear distance structure has been exploited in visual outlier detection, cross-modal learning, and contrastive representation learning Gonzalez-Jimenez et al. (2024); Hu et al. (2024); Ge et al. (2023).

Multi-agent BEV descriptors are affected by viewpoint, occlusion, sensing distance, and local point density. Consequently, true correspondences and hard negatives may occupy nearby Euclidean neighborhoods. Hyperbolic distance provides an alternative metric for reorganizing such candidate relationships. However, existing hyperbolic visual learning methods mainly address classification, retrieval, outlier detection, or semantic representation. Their objectives do not directly consider one-to-one spatial correspondence or subsequent relative pose estimation. HyperSync therefore introduces Poincaré-ball embedding into cross-agent BEV matching and combines hyperbolic correspondence learning with geometric pose recovery.

3 METHOD

3.1 PROBLEM FORMULATION AND OVERALL FRAMEWORK

Consider a collaborative perception scenario containing N agents,

$$\mathcal{A} = \{A_i\}_{i=1}^{N}, \tag{1}$$

where A_i denotes the i-th vehicle or roadside unit (RSU). Each agent acquires a LiDAR point cloud $\mathbf{P}_i$ in its local coordinate system and extracts an intermediate bird's-eye-view (BEV) feature map through a parameter-shared feature extractor:

$$\mathbf{F}_i = \mathcal{E}_{\text{bev}}(\mathbf{P}_i), \qquad \mathbf{F}_i \in \mathbb{R}^{C\times H\times W}, \tag{2}$$

where C, H, and W denote the number of feature channels, feature-map height, and feature-map width, respectively.

For an agent pair (A_s, A_t), referred to as the source and the target, the objective is to estimate the relative pose that maps observations from A_t into the coordinate system of A_s. The estimate uses only the two BEV features when reliable global localization is unavailable or the given pose is badly corrupted:

$$\widehat{\mathbf{T}}_{s\leftarrow t} = \mathcal{H}(\mathbf{F}_s, \mathbf{F}_t), \qquad \widehat{\mathbf{T}}_{s\leftarrow t} \in \mathrm{SE}(2). \tag{3}$$

The subscript $s \leftarrow t$ means that observations of A_t are transformed into the frame of A_s. The planar rigid transform contains the relative yaw and the planar translation, and is the alignment used by subsequent collaborative fusion.

As illustrated in Figure 4, HyperSync has four stages. Foreground keypoints are taken from the two BEV maps and encoded as pair-conditioned descriptors. These descriptors are embedded in a Poincaré ball and matched by hyperbolic geodesic distance. Discrete grid matches are then refined to continuous coordinates using the local score field. Finally, geometric consistency filtering and a robust $\mathrm{SE}(2)$ fit reject outliers, and a pose self-consistency gate drops pairs that lack sufficient geometric support.

3.2 BEV Candidate Extraction and Pair-Conditioned Descriptor Construction

A pillar-based BEV encoder Lang et al. (2019) converts each local LiDAR point cloud into a BEV feature map. Because the BEV observation is sparse, HyperSync matches occupied cells instead of every location on the feature map. Let $O_i(u, v, z) \in \{0, 1\}$ denote the voxel occupancy of agent A_i, with u, v, and z indexing the lateral, longitudinal, and vertical axes. Collapsing occupancy along the vertical axis yields

$$M_i(u, v) = \mathbb{I}\left[\sum_{z=1}^{Z} O_i(u, v, z) > 0\right], \tag{4}$$

where Z is the number of vertical voxels and $\mathbb{I}[\cdot]$ is the indicator. Thus $M_i(u, v) = 1$ only when the column (u, v) contains points. The candidate set of agent A_i is

$$\mathcal{K}_i = \left\{\mathbf{p}_{i,k} = [u_k, v_k]^\top \;\middle|\; M_i(u_k, v_k) = 1\right\}, \qquad N_i = |\mathcal{K}_i|, \tag{5}$$

and each discrete cell is converted to a metric planar coordinate $\widetilde{\mathbf{p}}_{i,k} \in \mathbb{R}^2$ by the sensing range, the voxel size, and the encoder stride. This conversion ties every descriptor to the physical plane on which the pose is estimated.

Descriptors of a pair (A_s, A_t) are built jointly. Features sampled at the candidates are augmented with their metric positional encodings. Bidirectional cross-attention exchanges context between the two candidate sets, responses are aggregated by feature similarity, and a linear projection fuses the result with the local representation. The pair-conditioned sets are

$$\mathcal{X}_{s|t} = \left\{\left(\widetilde{\mathbf{p}}_{s,k}, \mathbf{d}_{s,k|t}\right)\right\}_{k=1}^{N_s}, \qquad \mathcal{X}_{t|s} = \left\{\left(\widetilde{\mathbf{p}}_{t,\ell}, \mathbf{d}_{t,\ell|s}\right)\right\}_{\ell=1}^{N_t}, \tag{6}$$

where $\mathbf{d}_{s,k|t}$ is the descriptor of source candidate k conditioned on the target observation, and $\mathbf{d}_{t,\ell|s}$ is defined analogously. The coordinates enter pose estimation; the descriptors enter matching.

3.3 Hyperbolic Feature Embedding and Matching Metric

Large viewpoint changes, occlusion, and repeated structure can make a true match and a hard negative look alike under an inner product or a cosine score. HyperSync embeds the pair-conditioned descriptors in a Poincaré ball and compares them by geodesic distance.

Let the curvature be $-c$ with $c > 0$. The d-dimensional Poincaré ball is Nickel & Kiela (2017); Ganea et al. (2018)

$$\mathbb{B}_c^d = \left\{\mathbf{z} \in \mathbb{R}^d \;\middle|\; c\,\|\mathbf{z}\|_2^2 < 1\right\}. \tag{7}$$

A Euclidean descriptor is first mapped to a tangent vector at the origin. For the source descriptor $\mathbf{d}_{s,k|t}$,

$$\widetilde{\mathbf{d}}_{s,k} = \mathbf{W}_h \mathbf{d}_{s,k|t} + \mathbf{b}_h, \qquad \mathbf{v}_{s,k} = \alpha \frac{\widetilde{\mathbf{d}}_{s,k}}{\left\|\widetilde{\mathbf{d}}_{s,k}\right\|_2 + \varepsilon}, \tag{8}$$

where $\mathbf{W}_h$ and $\mathbf{b}_h$ are the projection parameters, α is a shared scale, and ε stabilizes the normalization. The exponential map then sends the tangent vector into the ball Ganea et al. (2018),

$$\mathbf{z}_{s,k} = \exp_{\mathbf{0}}^c(\mathbf{v}_{s,k}) = \tanh\left(\sqrt{c}\,\|\mathbf{v}_{s,k}\|_2\right) \frac{\mathbf{v}_{s,k}}{\sqrt{c}\,\|\mathbf{v}_{s,k}\|_2 + \varepsilon}, \tag{9}$$

and the target descriptor $\mathbf{d}_{t,\ell|s}$ is mapped in the same way to $\mathbf{z}_{t,\ell}$.

Because every tangent vector is ℓ_2-normalized and scaled by the same α, all embeddings share one Poincaré radius, $\|\mathbf{z}\|_2 = \tanh(\sqrt{c}\,\alpha)/\sqrt{c}$. On this sphere the geodesic is a strictly increasing function of the angle between descriptors, so mutual nearest neighbors coincide with cosine matching. The role of the hyperbolic distance is the training metric, not a different neighbor ranking at test time. With a positive temperature τ, the training logit and the inference score of a candidate pair are

$$\ell_{k\ell} = -\,d_{\mathbb{B}}(\mathbf{z}_{s,k}, \mathbf{z}_{t,\ell})/\tau, \qquad S_{k\ell} = \exp\!\big(-d_{\mathbb{B}}(\mathbf{z}_{s,k}, \mathbf{z}_{t,\ell})\big). \tag{10}$$

The geodesic is the Poincaré distance

$$d_{\mathbb{B}}(\mathbf{z}_{s,k}, \mathbf{z}_{t,\ell}) = \frac{1}{\sqrt{c}} \operatorname{arcosh}\left(1 + \frac{2c\|\mathbf{z}_{s,k} - \mathbf{z}_{t,\ell}\|_2^2}{\eta_{s,k}\eta_{t,\ell}}\right), \tag{11}$$

where

$$\eta_{s,k} = 1 - c\|\mathbf{z}_{s,k}\|_2^2, \qquad \eta_{t,\ell} = 1 - c\|\mathbf{z}_{t,\ell}\|_2^2. \tag{12}$$

A smaller distance yields a larger logit and a higher inference score.

3.4 Training Objective

Correspondence labels are built in the metric plane after the ground-truth relative pose has aligned the target candidates with the source. For source candidate k, let ℓ_k be its nearest target. The pair is a positive when this distance is smaller than r_p. Other targets that fall inside radius r_n, except ℓ_k itself, are ignored; the remaining targets are negatives. InfoNCE on the hyperbolic logits is

$$\mathcal{L}_{\text{match}} = -\ell_{k\ell_k} + \log \sum_{\ell \notin \mathcal{I}_k} \exp(\ell_{k\ell}), \tag{13}$$

where $\mathcal{I}_k$ is the ignored set, and the loss is averaged over source candidates that have a positive. The shared BEV backbone is also trained with the standard detection objective, so the full loss is

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{dir}} + \mathcal{L}_{\text{match}}. \tag{14}$$

Here $\mathcal{L}_{\text{cls}}$ is a focal classification loss, $\mathcal{L}_{\text{reg}}$ is a smooth-ℓ_1 box regression loss, and $\mathcal{L}_{\text{dir}}$ is a direction classification loss. During training, intermediate feature fusion still uses the ground-truth relative pose so that detection supervision stays stable. At test time the reported pose is produced only by $\mathcal{H}(\mathbf{F}_s, \mathbf{F}_t)$, without a localization prior.

3.5 Sub-Pixel Correspondence Refinement

Mutual nearest neighbors selected by $S_{k\ell}$ lie on discrete BEV cells. Voxelization snaps a continuous location to a nearby cell, and that quantization enters both translation and rotation. HyperSync therefore moves each matched coordinate according to the local score field, without changing which candidates are paired.

Let $\mathcal{C}_0 = \{(k, \ell_k)\}$ be the mutual-nearest-neighbor set, and let $\mathcal{N}_r(\ell_k)$ be a neighborhood of radius r around the matched target cell. Normalized weights are computed from the inference scores,

$$w_{k,j} = \frac{\exp\left(S_{kj}/\tau_{\text{sp}}\right)}{\sum_{j' \in \mathcal{N}_r(\ell_k)} \exp\left(S_{kj'}/\tau_{\text{sp}}\right)}, \qquad j \in \mathcal{N}_r(\ell_k), \tag{15}$$

where $\tau_{\text{sp}} > 0$ is the refinement temperature. The refined target coordinate is the weighted centroid

$$\widetilde{\mathbf{p}}_{t,k} = \sum_{j \in \mathcal{N}_r(\ell_k)} w_{k,j} \mathbf{p}_{t,j}. \tag{16}$$

Applying the same operation on the source side gives the continuous correspondence set

$$\mathcal{C}_{\text{sp}} = \left\{(\widetilde{\mathbf{p}}_{s,k}, \widetilde{\mathbf{p}}_{t,k})\right\}. \tag{17}$$

The refinement uses only local matching scores and does not introduce a pose prior.

3.6 Robust Spatial Synchronization

Continuous matches can still be outliers because of repeated structure, occlusion, or small overlap. HyperSync removes them by geometric consistency, then estimates a rigid pose from the remaining inliers.

A planar rigid motion preserves distances. For two correspondences i and j, the residual and the compatibility weight are

$$\delta_{ij} = \left| \|\widetilde{\mathbf{p}}_{s,i} - \widetilde{\mathbf{p}}_{s,j}\|_2 - \|\widetilde{\mathbf{p}}_{t,i} - \widetilde{\mathbf{p}}_{t,j}\|_2 \right|, \qquad A_{ij} = \max\left(0,\, 1 - \frac{\delta_{ij}^2}{\sigma^2}\right), \tag{18}$$

where σ controls the geometric tolerance. Correspondences with sufficient support are passed to MAGSAC Barath et al. (2019), which returns an inlier set consistent with the dominant transform. Those inliers are refit by a planar rigid motion with no extra scale,

$$\widehat{\mathbf{T}}_{s\leftarrow t} = \begin{bmatrix} \cos\theta & -\sin\theta & t_x \\ \sin\theta & \cos\theta & t_y \\ 0 & 0 & 1 \end{bmatrix} \in \mathrm{SE}(2), \tag{19}$$

where θ is the relative yaw and (t_x, t_y) is the planar translation. Dropping the scale degree of freedom matches the rigid layout of vehicles and roadside sensors, and prevents a free scale from absorbing matching noise into the rotation. Finally, correspondences are reprojected by $\widehat{\mathbf{T}}_{s\leftarrow t}$ and the inlier ratio is measured. If the ratio is below a threshold, the pair is rejected; otherwise the pose is kept for cross-agent feature alignment.

4 Experiments

To comprehensively evaluate the performance of HyperSync, we conduct cross-agent spatial synchronization experiments on OPV2V Xu et al. (2022b), V2X-Sim Li et al. (2022), and DAIR-V2X Yu et al. (2022). HyperSync is compared with representative feature-level matching methods and object-level association methods.

4.1 Datasets

The experiments cover two representative collaborative perception settings: vehicle-to-vehicle (V2V) collaboration on OPV2V and the V2V subset of V2X-Sim, and vehicle-to-infrastructure (V2I) collaboration on the V2I subset of V2X-Sim and DAIR-V2X.

OPV2V is a simulated dataset designed for multi-vehicle collaborative perception. We use its official test split to evaluate relative pose estimation between different vehicles.

V2X-Sim is a multi-agent collaborative perception dataset constructed using CARLA. We evaluate cross-agent spatial synchronization on its test split.

DAIR-V2X is a real-world V2I collaborative perception dataset containing heterogeneous observations acquired by vehicle-mounted and roadside sensors. We use its validation split to evaluate relative pose estimation between vehicle-side and infrastructure-side agents.

4.2 Evaluation Metrics

We evaluate relative pose estimation using the relative translation error (RTE) and relative rotation error (RRE), together with their median values. SR@1 and SR@2 denote the percentages of agent pairs with RTE below $1\,\mathrm{m}$ and $2\,\mathrm{m}$, respectively. An estimate is considered successful if

$$\mathrm{RTE} \le 3\,\mathrm{m}, \qquad \mathrm{RRE} \le 3^\circ. \tag{20}$$

To account for methods that may abstain from producing a valid pose, we additionally report

$$\begin{aligned} \mathrm{Cover} &= \frac{N_{\mathrm{eval}}}{N_{\mathrm{cand}}} \times 100\%, \\ \mathrm{CSR} &= \frac{N_{\mathrm{success}}}{N_{\mathrm{eval}}} \times 100\%, \\ \mathrm{Native} &= \frac{N_{\mathrm{success}}}{N_{\mathrm{cand}}} \times 100\%. \end{aligned} \tag{21}$$

Here, Cover measures valid-output coverage, CSR denotes the conditional success rate among valid estimates, and Native reflects the overall success rate over all candidate agent pairs. Detailed metric definitions are provided in the appendix.

4.3 Relative Pose Estimation Results

As shown in Table 1, HyperSync consistently outperforms existing methods across OPV2V, V2X-Sim, and DAIR-V2X. On OPV2V, where feature-level methods already achieve near-saturated

Table 1: Relative pose estimation results on OPV2V, V2X-Sim, and DAIR-V2X.

Dataset	Method	RTE(m)↓	RRE(°)↓	med RTE(m)↓	med RRE(°)↓	SR@1/2(%)↑	CSR(%)↑
OPV2V	CBM+SVD	0.653	0.793	0.348	0.345	93.1/97.9	97.9
	VIPS	1.676	2.664	0.590	0.535	70.7/86.5	86.4
	V2X-Reg++	3.227	4.318	0.386	0.294	85.3/90.8	91.3
	CPCoCo	0.098	0.220	0.075	0.157	99.7/99.9	99.7
	HyperSync	**0.085**	**0.158**	**0.064**	**0.100**	**99.9/100**	**99.9**
V2X-Sim	CBM+SVD	4.043	14.955	0.524	1.028	76.3/84.8	80.2
	VIPS	2.461	8.040	0.632	0.862	69.5/86.2	79.6
	V2X-Reg++	12.635	31.401	0.735	1.023	59.3/65.6	64.2
	CPCoCo	0.333	1.290	**0.134**	0.647	98.4/99.5	98.3
	HyperSync	**0.300**	**0.785**	0.151	**0.563**	**99.0/99.6**	**98.5**
DAIR-V2X	CBM+SVD	8.321	13.082	1.416	1.417	36.3/66.3	76.3
	VIPS	4.514	4.682	1.831	1.779	26.5/53.9	65.2
	V2X-Reg++	8.406	10.587	1.506	1.558	33.4/63.9	75.5
	CPCoCo	1.325	3.688	0.709	1.613	65.2/91.7	88.8
	HyperSync	**0.694**	**0.930**	**0.499**	**0.603**	**82.8/99.0**	**96.4**

Table 2: Difficulty-stratified relative pose estimation results on DAIR-V2X.

Level	Method	Cover (%)↑	RTE(m)↓	RRE(°)↓	SR@1/2/3 (%)↑	CSR(%)↑	Native (%)↑
Easy ($N = 509$)	CBM+SVD	**100**	2.146	2.792	48.7/78.6/91.9	85.5	85.5
	VIPS	60.5	3.581	4.568	34.7/60.7/78.6	67.5	40.9
	V2X-Reg++	**100**	7.785	11.324	42.6/72.1/85.5	78.2	78.2
	CPCoCo	36.1	0.911	1.667	69.6/86.4/98.9	93.3	32.8
	HyperSync	92.5	**0.678**	**0.931**	**77.1/99.8/100**	**96.6**	**89.4**
Medium ($N = 777$)	CBM+SVD	**99.1**	6.476	9.015	38.3/71.2/83.6	80.1	79.4
	VIPS	61.0	4.329	3.765	28.1/53.6/69.6	65.0	39.6
	V2X-Reg++	92.4	6.403	8.814	35.2/67.7/81.6	78.1	72.2
	CPCoCo	25.9	2.030	6.521	57.2/92.2/98.0	89.7	23.6
	HyperSync	85.6	**0.642**	**1.009**	**89.0/99.4/99.8**	**96.5**	**82.8**
Hard ($N = 503$)	CBM+SVD	**92.4**	18.135	31.080	19.4/44.7/63.4	59.8	55.3
	VIPS	37.8	6.486	7.156	9.5/43.7/66.3	62.1	23.5
	V2X-Reg++	72.4	13.225	13.055	16.8/45.1/72.5	66.8	48.3
	CPCoCo	24.5	0.792	2.082	71.50/**99.2/99.2**	79.7	19.5
	HyperSync	90.1	**0.788**	**0.811**	**78.6**/97.1/99.1	**96.0**	**86.5**

performance, HyperSync further reduces the mean RTE/RRE from $0.098\,\mathrm{m}/0.220^\circ$ (CPCoCo) to $0.085\,\mathrm{m}/0.158^\circ$, with a 99.9% CSR (conditional success rate). The advantage becomes more evident under larger cross-agent observation differences on V2X-Sim, where HyperSync reduces RTE/RRE by 9.9% and 39.1%, respectively, compared with CPCoCo. On the real-world DAIR-V2X benchmark, it further reduces RTE/RRE by 47.6% and 74.8%, while improving CSR from 88.8% to 96.4%. These results demonstrate robust pose estimation across varying viewpoints, sensor configurations, spatial baselines, and observation overlaps.

CBM, VIPS, and V2X-Reg++ all produce mean translation errors at the meter level, which are insufficient for accurate BEV feature alignment. Notably, although CBM and VIPS are provided with coarse relative-pose initialization, they still perform substantially worse than HyperSync, which does not rely on localization priors. These results further highlight the importance of stable feature-level correspondences for spatial synchronization in real-world vehicle-to-infrastructure scenarios.

4.4 Difficulty-Stratified Evaluation

To evaluate the effectiveness of hyperbolic feature representations under different levels of scene difficulty, we conduct a stratified evaluation on all 1,789 candidate agent pairs in the DAIR-V2X validation set. Based on the inter-agent distance and the number of co-visible objects, the samples are divided into Easy, Medium, and Hard subsets containing 509, 777, and 503 pairs, respectively.

Table 3: Ablation of Robust Sync and sub-pixel refinement on DAIR-V2X.

Robust Sync	Sub-pixel	RTE(m)↓	RRE(°)↓	med RTE(m)↓	med RRE(°)↓	SR@1/2(%)↑	CSR(%)↑
×	×	4.451	8.863	0.561	0.659	76.6/93.6	91.7
✓	×	0.758	1.101	0.510	0.577	82.1/99.0	96.2
✓	✓	**0.694**	**0.930**	**0.499**	**0.603**	**82.8/99.0**	**96.4**

To reduce the influence of differences in valid-pose determination across methods, we jointly report pose estimation accuracy, valid-output coverage, and full-set effectiveness.

HyperSync maintains high coverage and full-set effectiveness across all three difficulty levels. In contrast, CBM+SVD, VIPS, and V2X-Reg++, which establish correspondences using bounding boxes and object-level spatial relationships, exhibit substantial performance degradation as the inter-agent distance increases and co-visibility decreases. In the Hard setting, the SR@1 values of these object-level methods fall to approximately 10%-19%, whereas HyperSync still achieves an SR@1 of 78.9% and a CSR of 95.4%. These results indicate that hyperbolic feature matching can maintain relatively stable correspondence estimation even when shared observations are sparse.

Although CPCoCo achieves a relatively low conditional RTE in the Hard setting, its Cover and Native values are only 24.5% and 19.5%, respectively. CBM+SVD benefits from an initial pose estimate, whereas V2X-Reg++ relies on dataset-provided bounding-box annotations to establish object correspondences and still degrades substantially as the number of co-visible objects decreases. Overall, the stratified results demonstrate that HyperSync offers stronger adaptability and effectiveness under larger cross-agent observation discrepancies and more limited co-visibility.

4.5 Ablation

We ablate the two post-matching components applied to the hyperbolic correspondences: sub-pixel refinement and Robust Sync, which combines geometric consistency filtering with pose gating. The underlying robust pose estimator is kept fixed across all variants, allowing the effects of correspondence refinement and consistency filtering to be evaluated independently of the pose solver. We report the mean and median RTE/RRE, SR@1/2, and CSR on the DAIR-V2X validation set.

As shown in Table 3, Robust Sync contributes the largest performance gain. Without it, the mean RTE and RRE remain high at $4.451\,\mathrm{m}$ and $8.863°$, respectively, despite relatively moderate median errors, indicating that residual outlier correspondences substantially affect the overall estimation accuracy. Enabling Robust Sync reduces the mean RTE/RRE to $0.758\,\mathrm{m}$ and $1.101°$, while increasing SR@1 and Success to 82.1% and 96.2%, respectively. Adding sub-pixel refinement further improves the full model to a mean RTE/RRE of $0.694\,\mathrm{m}$ and $0.930°$, with median errors of $0.499\,\mathrm{m}$ and $0.603°$. The corresponding SR@1/2 reaches $82.8\%/99.0\%$, and CSR increases to 96.4%. These results indicate that Robust Sync primarily removes geometrically inconsistent correspondences, whereas sub-pixel refinement provides an additional accuracy improvement after reliable matches have been established.

5 Conclusion

This paper presents HyperSync, a hyperbolic feature matching framework for multi-agent spatial synchronization, designed to address large viewpoint differences, incomplete observations, and sparse co-visibility. HyperSync embeds pair-conditioned bird's-eye-view features into a Poincaré ball and integrates hyperbolic matching, correspondence refinement, and geometric consistency filtering for robust relative pose estimation. Experiments on OPV2V, V2X-Sim, and DAIR-V2X demonstrate consistent improvements over representative object-level and feature-level baselines, as well as stronger robustness and adaptability in challenging scenarios characterized by large inter-agent distances and few co-visible objects.

AI USE STATEMENT

We used generative AI tools in this work solely to assist with manuscript editing, including language refinement, formatting of mathematical expressions, and improvement of figure captions. The core research process including methodology development, theoretical and mathematical formulation, experimental design and implementation, and analysis and interpretation of experimental results was conducted by the authors without generative AI assistance. We also used generative AI tools to support preliminary literature exploration. All AI-assisted materials were carefully reviewed by the authors, and the corresponding text and technical descriptions were checked against the original manuscript and experimental records. Experimental results and related claims were independently verified by the authors. The authors retain full responsibility for the accuracy, originality, and integrity of the final manuscript and all associated research content.

ETHICS STATEMENT

This work uses publicly available benchmark datasets and synthetic data instances. It does not involve human subjects, personal data, or the collection or processing of sensitive information.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, the implementation of the proposed method, data-processing procedures, learning-rate settings, and training and inference configurations are provided in the accompanying code package. The appendix further provides detailed algorithmic descriptions, experimental settings, pseudocode, and visualizations of intermediate outputs from different stages of the proposed method.

REFERENCES

Daniel Barath, Jiri Matas, and Jana Noskova. Magsac: Marginalizing sample consensus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10197–10205, 2019.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. *Advances in neural information processing systems*, 32, 2019.

Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. *Advances in neural information processing systems*, 31, 2018.

Songwei Ge, Shlok Mishra, Simon Kornblith, Chun-Liang Li, and David Jacobs. Hyperbolic contrastive learning for visual representations beyond objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6840–6849, 2023.

Alvaro Gonzalez-Jimenez, Simone Lionetti, Dena Bazazian, Philippe Gottfrois, Fabian Gröger, Alexander Navarini, and Marc Pouly. Hyperbolic metric learning for visual outlier detection. In *European Conference on Computer Vision*, pp. 327–344. Springer, 2024.

Jiuwu Hao, Liguo Sun, Ti Xiang, Yuting Wan, Haolin Song, and Pin Lv. Feakm: Robust collaborative perception under noisy pose conditions. In *Proceedings of the 2024 4th International Joint Conference on Robotics and Artificial Intelligence*, pp. 98–102, 2024.

Yuan He, Jichuan Li, and Junjie Liu. Research on gnss ins & gnss/ins integrated navigation method for autonomous vehicles: A survey. *IEEE Access*, 11:79033–79055, 2023.

Naiwen Hu, Haozhe Cheng, Yifan Xie, Pengcheng Shi, and Jihua Zhu. Hyperbolic image-and-pointcloud contrastive learning for 3d classification. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4973–4979. IEEE, 2024.

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

Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, and Konrad Schindler. Predator: Registration of 3d point clouds with low overlap. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 4267–4276, 2021.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6418–6428, 2020.

Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12697–12705, 2019.

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

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

Hyungtae Lim, Suyong Yeon, Soohyun Ryu, Yonghan Lee, Youngji Kim, Jaeseong Yun, Euigon Jung, Donghwan Lee, and Hyun Myung. A single correspondence is enough: Robust global registration to avoid degeneracy in urban environments. In *2022 international conference on robotics and automation (ICRA)*, pp. 8010–8017. IEEE, 2022.

Si Liu, Chen Gao, Yuan Chen, Xingyu Peng, Xianghao Kong, Kun Wang, Runsheng Xu, Wentao Jiang, Hao Xiang, Jiaqi Ma, et al. Towards vehicle-to-everything autonomous driving: A survey on collaborative perception. *arXiv preprint arXiv:2308.16714*, 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. *arXiv preprint arXiv:2211.07214*, 2022.

Zhenyang Ni, Zixing Lei, Yifan Lu, Dingju Wang, Chen Feng, Yanfeng Wang, and Siheng Chen. Self-localized collaborative perception. *arXiv preprint arXiv:2406.12712*, 2024.

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017.

Qianxin Qu, Yijin Xiong, Guipeng Zhang, Xin Wu, Xiaohan Gao, Xin Gao, Hanyu Li, Shichun Guo, and Guoying Zhang. V2i-calib: A novel calibration approach for collaborative vehicle and infrastructure lidar systems. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 892–897. IEEE, 2024.

Haojie Ren, Sha Zhang, Sugang Li, Yao Li, Xinchen Li, Jianmin Ji, Yu Zhang, and Yanyong Zhang. Trajmatch: Toward automatic spatio-temporal calibration for roadside lidars through trajectory matching. *IEEE Transactions on Intelligent Transportation Systems*, 24(11):12549–12559, 2023.

Szymon Rusinkiewicz and Marc Levoy. Efficient variants of the icp algorithm. In *Proceedings third international conference on 3-D digital imaging and modeling*, pp. 145–152. IEEE, 2001.

Shuyao Shi, Jiahe Cui, Zhehao Jiang, Zhenyu Yan, Guoliang Xing, Jianwei Niu, and Zhenchao Ouyang. Vips: Real-time perception fusion for infrastructure-assisted autonomous driving. In *Proceedings of the 28th annual international conference on mobile computing and networking*, pp. 133–146, 2022.

Zhiying Song, Tenghui Xie, Hailiang Zhang, Jiaxin Liu, Fuxi Wen, and Jun Li. A spatial calibration method for robust cooperative perception. *IEEE Robotics and Automation Letters*, 9(5):4011–4018, 2024.

Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *European conference on computer vision*, pp. 605–621. Springer, 2020.

Yue Wang and Justin M Solomon. Deep closest point: Learning representations for point cloud registration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3523–3532, 2019.

Langkun Xu, Chenyang Li, Zhongxiang Zhao, and Lin Wang. Cpcoco: Localization-free collaborative perception via attentive contrastive consensus. *IEEE Transactions on Intelligent Transportation Systems*, 2026.

Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *European conference on computer vision*, pp. 107–124. Springer, 2022a.

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

Heng Yang, Jingnan Shi, and Luca Carlone. Teaser: Fast and certifiable point cloud registration. *IEEE Transactions on Robotics*, 37(2):314–333, 2020.

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

Junle Yu, Luwei Ren, Yu Zhang, Wenhui Zhou, Lili Lin, and Guojun Dai. Peal: Prior-embedded explicit attention learning for low-overlap point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17702–17711, 2023.

Yun Yue, Fangzhou Lin, Guanyi Mou, and Ziming Zhang. Understanding hyperbolic metric learning through hard negative sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1891–1903, 2024.

APPENDIX A: MATHEMATICAL DETAILS

This appendix collects the closed-form expressions used in HyperSync for hyperbolic matching, sub-pixel correspondence refinement, geometric consensus filtering, and localization-free $\mathrm{SE}(2)$ synchronization.

POINCARÉ BALL EMBEDDING

Let $c > 0$ denote the curvature magnitude. The d-dimensional Poincaré ball is

$$\mathbb{B}_c^d = \big\{\mathbf{z} \in \mathbb{R}^d \;\big|\; c\|\mathbf{z}\|_2^2 < 1\big\}. \tag{22}$$

Given a Euclidean pair-conditioned descriptor $\mathbf{d}$, we first form a tangent vector at the origin,

$$\mathbf{v} = \alpha\,\mathrm{Norm}\big(W_h\mathbf{d} + \mathbf{b}_h\big), \tag{23}$$

where W_h and $\mathbf{b}_h$ are a linear projection, $\mathrm{Norm}(\cdot)$ denotes ℓ_2 normalization, and $\alpha > 0$ scales the tangent radius. The exponential map at the origin embeds $\mathbf{v}$ into $\mathbb{B}_c^d$:

$$\mathbf{z} = \exp_0^c(\mathbf{v}) = \tanh\big(\sqrt{c}\,\|\mathbf{v}\|_2\big)\,\frac{\mathbf{v}}{\sqrt{c}\,\|\mathbf{v}\|_2}. \tag{24}$$

Source and target descriptors $\mathbf{d}_{s,k|t}$ and $\mathbf{d}_{t,\ell|s}$ are mapped independently to $\mathbf{z}_{s,k}$ and $\mathbf{z}_{t,\ell}$.

GEODESIC DISTANCE AND MATCHING SCORES

Define the conformal factors

$$\eta(\mathbf{z}) = 1 - c\|\mathbf{z}\|_2^2. \tag{25}$$

The geodesic distance on $\mathbb{B}_c^d$ is

$$d_{\mathbb{B}}(\mathbf{z}_s, \mathbf{z}_t) = \frac{1}{\sqrt{c}} \operatorname{arcosh}\left(1 + \frac{2c\,\|\mathbf{z}_s - \mathbf{z}_t\|_2^2}{\eta(\mathbf{z}_s)\,\eta(\mathbf{z}_t)}\right). \tag{26}$$

At training time, matching logits are the negative scaled distance

$$\ell_{k\ell} = -\frac{d_{\mathbb{B}}(\mathbf{z}_{s,k}, \mathbf{z}_{t,\ell})}{\tau}, \tag{27}$$

and are supervised with an InfoNCE-style contrastive objective. At inference, we convert distances into similarity scores

$$S_{k\ell} = \exp\big(-d_{\mathbb{B}}(\mathbf{z}_{s,k}, \mathbf{z}_{t,\ell})\big) \tag{28}$$

(or, equivalently, reuse $\ell_{k\ell}$ when the training/inference score maps are aligned). Mutual nearest-neighbor search on S (or ℓ) yields the discrete correspondence set $\mathcal{C}_0 = \{(k, \ell_k)\}$. Because $\exp(-d_{\mathbb{B}})$ and $-d_{\mathbb{B}}/\tau$ are strictly monotone in $d_{\mathbb{B}}$, they induce the same candidate ranking.

Sub-pixel Correspondence Refinement

Integer BEV indices introduce quantization bias. For each mutual match (k, ℓ_k) with discrete coordinates $\mathbf{p}_{s,k}$ and $\mathbf{p}_{t,\ell_k}$, let $S_{k,:}$ be the source-to-target score row. Soft weights in a neighborhood of radius r are

$$w_{k,j} = \frac{\exp\big(S_{kj}/\tau_{\mathrm{sp}}\big)\,\mathbb{I}\big(\|\mathbf{p}_{t,j} - \mathbf{p}_{t,\ell_k}\|_2 \le r\big)}{\sum_{j'} \exp\big(S_{kj'}/\tau_{\mathrm{sp}}\big)\,\mathbb{I}\big(\|\mathbf{p}_{t,j'} - \mathbf{p}_{t,\ell_k}\|_2 \le r\big)}, \tag{29}$$

where $\tau_{\mathrm{sp}} > 0$ is the refinement temperature. The refined target coordinate is the local soft-argmax centroid

$$\tilde{\mathbf{p}}_{t,k} = \sum_j w_{k,j}\,\mathbf{p}_{t,j}. \tag{30}$$

The symmetric target-to-source pass produces $\tilde{\mathbf{p}}_{s,k}$. Mapping continuous grid coordinates to metric BEV yields

$$\mathcal{C}_{\mathrm{sp}} = \big\{(\tilde{\mathbf{p}}_{s,k}, \tilde{\mathbf{p}}_{t,k})\big\}. \tag{31}$$

Distance-Preservation Consensus

Planar rigid motions approximately preserve pairwise distances. For any two refined correspondences $i, j \in \mathcal{C}_{\mathrm{sp}}$, define the distance residual

$$\delta_{ij} = \big\|\tilde{\mathbf{p}}_{s,i} - \tilde{\mathbf{p}}_{s,j}\big\|_2 - \big\|\tilde{\mathbf{p}}_{t,i} - \tilde{\mathbf{p}}_{t,j}\big\|_2 \tag{32}$$

and the pairwise compatibility

$$A_{ij} = \max\Big(0,\, 1 - \delta_{ij}^2/\sigma^2\Big), \tag{33}$$

with tolerance $\sigma > 0$. The support of correspondence i is the number of partners j with $A_{ij} > \eta$. Correspondences whose support falls below a fraction of $|\mathcal{C}_{\mathrm{sp}}|$ are discarded, producing an inlier candidate set $\mathcal{C}$. This filter uses only predicted geometry and does not rely on ground-truth poses.

SE(2) Estimation and Self-Consistency Gate

On $\mathcal{C}$, a robust estimator (MAGSAC) identifies an inlier subset $\mathcal{I}$. We then refit a planar rigid transform with fixed scale,

$$\mathbf{T}_{s\leftarrow t} = \begin{bmatrix} \cos\theta & -\sin\theta & t_x \\ \sin\theta & \cos\theta & t_y \end{bmatrix} \in \mathrm{SE}(2), \tag{34}$$

estimating only yaw θ and translation (t_x, t_y). Constraining the scale removes a free degree of freedom that would otherwise absorb matching noise.

Finally, let r_i denote the metric reprojection residual of correspondence i under $\mathbf{T}_{s\leftarrow t}$, and let

$$\rho = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \mathbb{I}(r_i < \epsilon) \tag{35}$$

be the inlier rate under threshold ϵ. If ρ is below a preset minimum, the pair is rejected; otherwise HyperSync outputs $\hat{\mathbf{T}}_{s\leftarrow t} = \mathbf{T}_{s\leftarrow t}$.

APPENDIX B: EVALUATION PROTOCOL AND METRICS

This appendix details the relative-pose evaluation protocol used throughout the paper. Unless otherwise stated, all main-table RTE/RRE statistics are computed over the set of pairs that return a finite pose estimate (N_{eval}).

COORDINATE SYSTEM AND RELATIVE POSE

For each agent pair (i, j), HyperSync estimates a planar rigid transform

$$\hat{\mathbf{T}}_{i\leftarrow j} = \begin{bmatrix} \cos\hat{\theta} & -\sin\hat{\theta} & \hat{t}_x \\ \sin\hat{\theta} & \cos\hat{\theta} & \hat{t}_y \end{bmatrix} \in \mathrm{SE}(2), \tag{36}$$

i.e., yaw $\hat{\theta}$ and planar translation $(\hat{t}_x, \hat{t}_y)$ in meters. Ground-truth pairwise transforms are likewise reduced to the BEV xy plane (ignoring the vertical component). All errors below are therefore **2D**: translation in the ground plane and rotation about the vertical axis.

RTE AND RRE

Let $\mathbf{T}^{\text{gt}}$ and $\hat{\mathbf{T}}$ denote the ground-truth and estimated $\mathrm{SE}(2)$ transforms. Write $\mathbf{R}^{\text{gt}}, \hat{\mathbf{R}} \in \mathrm{SO}(2)$ and $\mathbf{t}^{\text{gt}}, \hat{\mathbf{t}} \in \mathbb{R}^2$ for their rotation and translation parts. The relative translation / rotation errors are

$$\mathrm{RTE} = \left\|\hat{\mathbf{t}} - \mathbf{t}^{\text{gt}}\right\|_2 \quad (\mathrm{m}), \tag{37}$$

$$\mathrm{RRE} = \frac{180}{\pi} \left\|\log\!\left(\hat{\mathbf{R}}^{\top}\mathbf{R}^{\text{gt}}\right)\right\|_{\mathrm{F}} \quad (^\circ), \tag{38}$$

where $\log(\cdot)$ is the matrix logarithm on $\mathrm{SO}(2)$. For a set of evaluated pairs $\mathcal{E}$ with $|\mathcal{E}| = N_{\text{eval}}$, we report the mean and median:

$$\overline{\mathrm{RTE}} = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \mathrm{RTE}_e, \qquad \mathrm{med\,RTE} = \mathrm{median}\{\mathrm{RTE}_e\}_{e\in\mathcal{E}}, \tag{39}$$

$$\overline{\mathrm{RRE}} = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \mathrm{RRE}_e, \qquad \mathrm{med\,RRE} = \mathrm{median}\{\mathrm{RRE}_e\}_{e\in\mathcal{E}}. \tag{40}$$

Unless noted, "RTE / RRE" in tables refer to these means.

SUCCESS RATES

Translation-only success rates at thresholds $\lambda \in \{1, 2, 3\}\,\mathrm{m}$ are

$$\mathrm{SR@}\lambda = \frac{1}{N_{\text{eval}}}\sum_{e\in\mathcal{E}} \mathbb{I}\big(\mathrm{RTE}_e < \lambda\big). \tag{41}$$

The joint success criterion used in the main tables is

$$\mathrm{Success} = \mathbb{I}\big(\mathrm{RTE} \le 3\,\mathrm{m} \ \wedge\ \mathrm{RRE} \le 3^\circ\big), \tag{42}$$

and CRS is the fraction of evaluated pairs that satisfy Success:

$$\mathrm{CSR} = \frac{N_{\text{success}}}{N_{\text{eval}}}, \qquad N_{\text{success}} = \sum_{e\in\mathcal{E}} \mathrm{Success}_e. \tag{43}$$

COVERAGE, CONDITIONAL SUCCESS, AND NATIVE SUCCESS

Methods may abstain (e.g., insufficient correspondences or failed pose gating). Let

- N_{cand}: number of candidate agent pairs in the evaluation split;
- N_{eval}: number of pairs that return a finite pose ($N_{\text{eval}} = |\mathcal{E}|$);
- N_{success}: number of pairs among $\mathcal{E}$ that satisfy Success.

We define

$$\text{Cover} = \frac{N_{\text{eval}}}{N_{\text{cand}}}, \quad \text{CSR} = \frac{N_{\text{success}}}{N_{\text{eval}}}, \quad \text{Native} = \frac{N_{\text{success}}}{N_{\text{cand}}}. \tag{44}$$

Thus $\text{Native} = \text{Cover} \times \text{CSR}$. Cover% measures how often a method produces a usable estimate; CSR measures accuracy conditional on producing an estimate; Native% reflects end-to-end usefulness on the full candidate set and is preferred when comparing methods with different abstention rates.

LOCALIZATION PRIORS AND FAIR COMPARISON

Compared methods fall into two groups:

Table 4: Pose-prior settings of compared methods.

Method	Pose prior	Notes
HyperSync (Ours)	None (localization-free)	Hyperbolic BEV matching + robust $\mathrm{SE}(2)$ sync
CPCoCo	None (localization-free)	Euclidean bilinear matching
V2X-Reg++	None (localization-free)	Detection-box association
CBM+SVD	Noisy prior: GT $+\mathcal{N}(5\,\mathrm{m}, 3^\circ)$	Object-graph matching; SVD pose (no ICP)
VIPS	Noisy prior: GT $+\mathcal{N}(5\,\mathrm{m}, 3^\circ)$	Category-consistent association; same SVD setting

For the object-association baselines on OPV2V and V2X-Sim, we use per-agent ground-truth bounding boxes corrupted by independent Gaussian noise with standard deviations of $0.5\,\mathrm{m}$ in translation and 2° in yaw. On DAIR-V2X, we use the vehicle-side and infrastructure-side annotations provided by the dataset. Following their original designs, CBM and VIPS are additionally given a noisy pose prior obtained by perturbing the ground-truth relative pose with $\mathcal{N}(5\,\mathrm{m}, 3^\circ)$. Baseline implementations follow the official source code and/or released checkpoints when available. Accordingly, the main-table comparisons evaluate whether localization-free feature matching can match or surpass object association with pose priors.

DATASET LIDARS EQUIPMENT

Table 5: Specifications of the LiDAR sensors in the evaluated datasets.

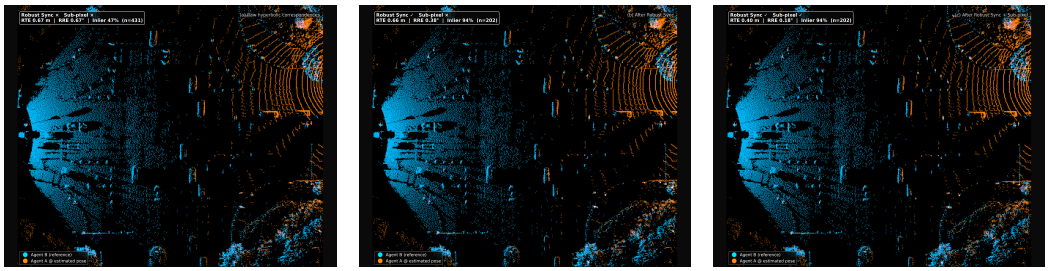
Category	**Dataset**	**Channels**	**Range (m)**
Vehicle-side LiDARs	OPV2V	64	120
	V2X-Sim	32	70
	DAIR-V2X	40	200
Infrastructure-side LiDARs	V2X-Sim	32	70
	DAIR-V2X	300	280

EXPERIMENTAL HARDWARE

All model training, inference, and evaluation were conducted on the same Linux workstation. The system was equipped with two NVIDIA RTX A6000 GPUs, each with 48 GB of memory, and two AMD EPYC 9754 processors, providing a total of 256 physical cores and 512 logical threads.

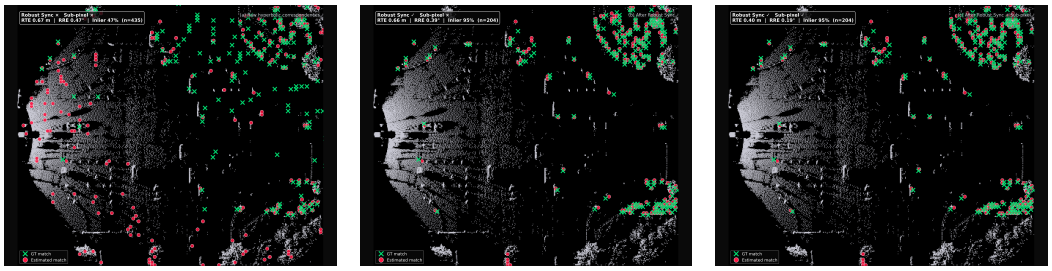
APPENDIX C: VISUALIZATION OF THE ABLATION STUDY

Figure 3 presents the visual ablation results on a representative frame from DAIR-V2X. Without Robust Sync and Sub-pixel refinement, the inlier ratio is only 47%. After enabling Robust Sync, the inlier ratio increases to 94-95%, the RRE decreases from approximately 0.47° to 0.39°, and the RTE decreases from 0.67 m to 0.66 m. With the further introduction of Sub-pixel refinement, the RTE and RRE are reduced to 0.40 m and 0.19°, respectively. The corresponding point-cloud registration results are shown in Figure 4.



(a) Raw hyperbolic correspondences (b) After Robust Sync (c) Robust Sync + Sub-pixel

Figure 3: Matching-point visualization ablation on a representative frame from the DAIR-V2X validation set: (a) original hyperbolic correspondences; (b) with Robust Sync enabled; and (c) with Sub-pixel refinement additionally enabled.



(a) Raw hyperbolic correspondences (b) After Robust Sync (c) Robust Sync + Sub-pixel

Figure 4: Point-cloud registration visualization for an ablation study on a representative frame from the DAIR-V2X validation set: (a) original hyperbolic correspondences; (b) with Robust Sync enabled; and (c) with Sub-pixel refinement additionally enabled. Cyan denotes the reference point cloud from Agent B, while orange denotes the point cloud from Agent A transformed using the estimated pose.

Robust Sync improves correspondence accuracy through geometric-consensus filtering, whereas Sub-pixel refinement reduces BEV quantization errors by refining discrete grid-center coordinates. The average refinement displacement is 0.27 m, thereby providing more accurate geometric constraints for relative pose estimation.

APPENDIX D: MODEL PIPELINE VISUALIZATION

Figure 5 presents the spatial synchronization pipeline on DAIR-V2X. The model takes vehicle-side and roadside-unit (RSU) LiDAR point clouds as input, as shown in (a) and (b), respectively. PointPillars scatter encoding produces the pillar feature maps in (c) and (d), which are then fed into the cross-agent matching branch to extract the matcher BEV features in (e) and (f). These features are further refined using a LoFTR-style module, yielding the refined matcher features in (g) and (h). Under the hyperbolic OverlapHead, cross-agent correspondences are first established in the feature-level BEV space in (i), where the left and right panels represent the vehicle-side and RSU-side feature maps and the connecting lines denote matched pairs. The same correspondences are then visualized in the point-cloud BEV space in (j) and used for relative pose estimation. Finally, the RSU point cloud is transformed into the vehicle coordinate frame using the estimated relative pose, producing the pose-aligned fused point cloud in (k), without relying on additional global localization priors. All subfigures visualize the actual forward activations and matching results for the same frame.

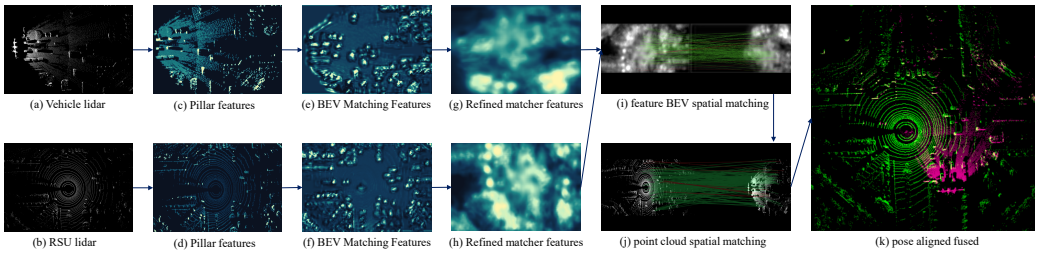


Figure 5: Model-pipeline visualization on a representative DAIR-V2X sample: (a-b) vehicle-side and RSU-side LiDAR inputs; (c-d) pillar-encoded features; (e-f) BEV matching features; (g-h) refined matcher features; (i) feature-level BEV spatial matching; (j) point-cloud spatial matching; and (k) the pose-aligned fused result.

APPENDIX E: ALGORITHMIC WORKFLOW AND PSEUDOCODE

Algorithm 1 Multi-agent cooperative perception with learned cross-view alignment

Input: point clouds $\{\mathbf{P}_i\}_{i=1}^{N}$; labels $\mathcal{Y}$ (training only)
Output: ego detections $\mathcal{D}$; relative poses $\{\widehat{\mathbf{T}}_{0j}\}_{j\neq 0}$ (inference)
# Feature encoding
for $i = 1$ to N **do**
    $\mathbf{F}_i \leftarrow \Phi_{\text{enc}}(\mathbf{P}_i) \in \mathbb{R}^{H\times W\times D}$ ▷ Lift sensors to a shared BEV grid
end for
if training **then**
    # Cross-agent representation alignment
    for each ordered pair (i, j) within communication range **do**
        Extract foreground tokens $\mathbf{u}_i, \mathbf{v}_j$ from $(\mathbf{F}_i, \mathbf{F}_j)$
        $\mathbf{S}_{ij} \leftarrow \mathcal{G}(\mathbf{u}_i, \mathbf{v}_j; \theta)$ ▷ Joint similarity in mixed metric spaces
        $\mathcal{L}_{\text{align}} \mathrel{+}= \ell\big(\mathbf{S}_{ij}, \mathbf{M}_{ij}^{\text{gt}}\big)$
    end for
    # Supervised cooperative detection
    Warp $\{\mathbf{F}_j\}$ to ego frame using $\{\mathbf{T}_{0j}^{\text{gt}}\}$
    $\mathbf{F}_0^{\text{fus}} \leftarrow f_{\text{fus}}(\{\mathbf{F}_i\})$
    $\widehat{\mathcal{Y}} \leftarrow \Phi_{\text{dec}}(\mathbf{F}_0^{\text{fus}})$
    Minimize $\mathcal{L}_{\text{det}}(\widehat{\mathcal{Y}}, \mathcal{Y}) + \lambda\, \mathcal{L}_{\text{align}}$
    return
end if
# Inference: pose from correspondences
for each pair $(0, j)$ within communication range **do**
    Score all token pairs; keep mutual-consistent matches $\mathcal{C}_{0j}$
    Apply geometric plausibility mask on $\mathcal{C}_{0j}$
    if $|\mathcal{C}_{0j}| < \tau_{\min}$ **then**
        Mark pair $(0, j)$ as invalid; **continue**
    end if
    $\widehat{\mathbf{T}}_{0j} \leftarrow \text{RobustFit}\big(\mathcal{C}_{0j}\big)$ ▷ SE(2) estimate with outlier rejection
end for
# Cooperative detection
Warp received maps with $\{\widehat{\mathbf{T}}_{j0}\}$; fuse to $\mathbf{F}_0^{\text{fus}}$
$\mathcal{D} \leftarrow \Phi_{\text{dec}}(\mathbf{F}_0^{\text{fus}})$
return $\mathcal{D}$, $\{\widehat{\mathbf{T}}_{0j}\}$