EASYCALIB: ONLINE SPATIO-TEMPORAL RADAR-CAMERA CALIBRATION FROM OBJECT OBSERVATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Radar and cameras are widely paired in vehicles, roadside infrastructure, and robotics, yet estimating their relative pose and timing online from routine object observations remains difficult. Camera detection centers and radar returns may describe different physical points; measurement errors differ, and object motion couples pose with timing. We present EasyCalib, which estimates six spatial perturbations and an added radar timestamp offset from observations accumulated over time and a nominal sensor transform. Its XYT representation encodes planar position and reported time while retaining visual height as an attribute. A coarse pose estimate is refined through rigid-transform composition, then soft candidate association estimates the offset after spatial compensation. In one forward pass, EasyCalib outputs 3D translation, 3D rotation, and the offset without supplied point correspondences or a pre-estimated delay. We evaluate pose perturbations, offset ranges, training continuations, and perception frontends on an internal nuScenes split with disjoint recording logs. At the widest R4–D4 setting (± 10 m per translation axis, $\pm 30^\circ$ per rotation axis, and ± 1 s added offset), mean per-axis translation and rotation MAEs are 51.89 cm and 0.950°, and added-offset MAE is 238.45 ms. Spatial refinement reduces translation and rotation errors by 10.5% and 23.1%; offset error is 52.3% below a zero-offset control.

1 INTRODUCTION

Vehicles, roadside installations, and robots combine camera semantics with radar range and motion. Their fusion depends on a stable relative pose and timing relationship. Maintenance, sensor replacement, and changes to the timestamp pipeline can invalidate calibration obtained at installation. Detections collected during ordinary operation can then support calibration checks and candidate corrections without deploying dedicated targets.

Calibration remains difficult because visual detections and radar returns may observe different parts of an object. Camera depth depends on the upstream detector, while nuScenes radar supplies sparse planar returns without measured object height or orientation (Hwang et al., 2022; Luu et al., 2025). Height-related pose components therefore rely more heavily on visual geometry and the nominal mounting transform. Motion can make delay look like translation, particularly when targets move at similar velocities.

Dedicated targets supply controlled geometry (Cheng et al., 2023). Targetless methods exploit sensor motion (Wise et al., 2023), align matched trajectories (Durmaz & Cevikalp, 2025), or establish object correspondences before geometric estimation (He et al., 2026; Li et al., 2026). Our input is less structured: asynchronous visual detections and native radar returns accumulated over a window. The unknown pose determines which observations can be associated, while those associations provide evidence for pose and delay.

EasyCalib addresses this dependency in stages. Predicted prealignment narrows the association problem, a rigid residual refines pose, and temporal consensus aggregates candidate time differences after spatial compensation. The same observation window yields six spatial corrections and one added offset.

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.

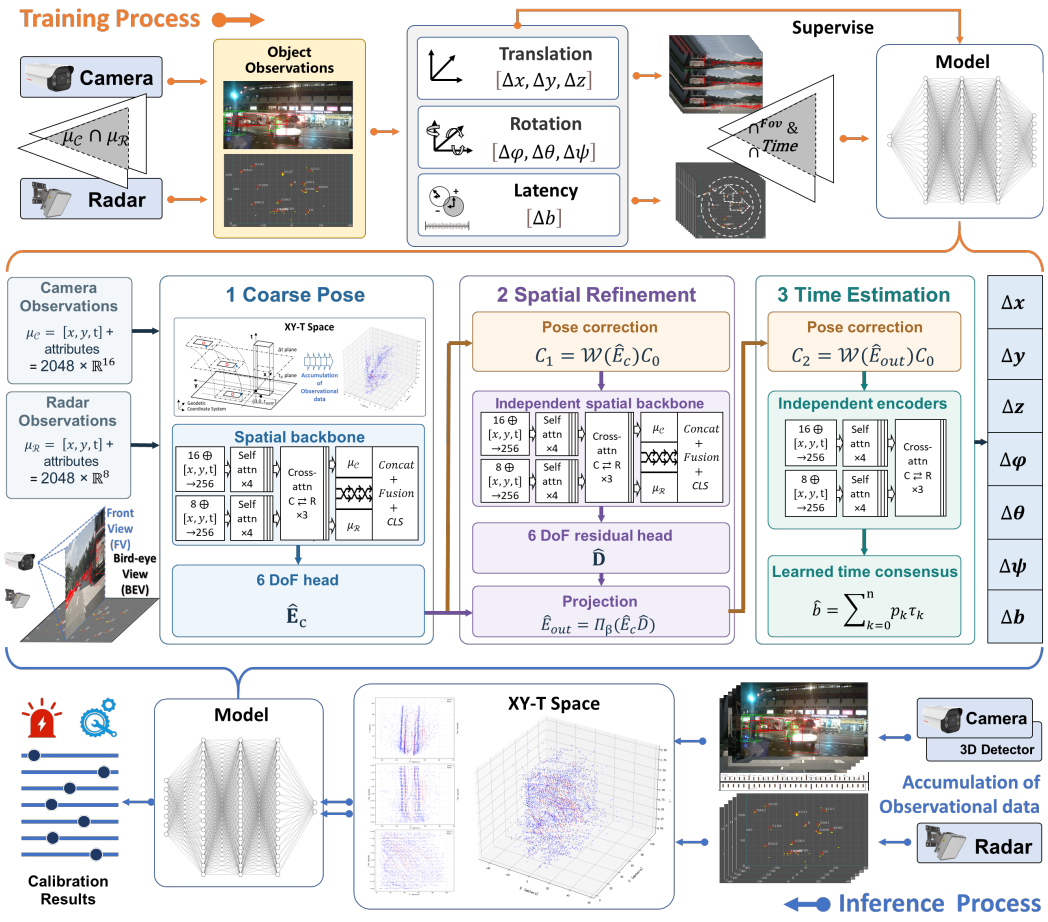

Figure 1: EasyCalib training (top) and window-level inference (bottom). The coarse and residual poses are composed before original visual observations are re-warped for offset estimation. Spatial networks are frozen during temporal training; labels and correspondences are not inference inputs.

Here online denotes forward inference over an accumulated observation window (Figure 1). Given camera detections, native radar returns, a nominal transform, and sensor-pair identity, EasyCalib predicts bounded pose and added-offset corrections without supplied correspondences or a pre-estimated offset.

Our contributions are:

1. We formulate object-based calibration as recovery from coupled pose, timing, and association residuals. Predicted prealignment makes soft association informative, while the associations supply evidence for spatial and temporal correction.
2. We develop a window-level radar–camera estimator that uses camera detections, native radar returns, reported timestamps, and a nominal transform to predict six pose parameters and one added offset in a single forward pass.
3. We replace height with time in XYT position encoding while retaining visual height as an observation attribute. Coarse estimation, rigid refinement, and temporal consensus turn this representation into a seven-parameter output without a pre-estimated offset.
4. We evaluate perturbation ranges, training continuations, perception frontends, and datasets on disjoint scene and recording-log splits, reporting error alongside admission coverage and timestamp-only controls.

2 RELATED WORK

Spatial and temporal calibration. Target-based radar-camera calibration uses controlled geometry (Cheng et al., 2023). Targetless approaches use track association (Liu et al., 2025), trajectory alignment (Durmaz & Cevikalp, 2025), or continuous-time motion (Wise et al., 2021). FusionCalib estimates the road-relative pose of an integrated roadside sensor under assumptions about its internal radar-camera alignment (Deng et al., 2023). Joint spatial and temporal calibration predates our work: Gaussian-process tracking supports multisensor calibration (Peršić et al., 2021), Wise et al. (2023) analyze radar-camera calibration with motion excitation and identifiability, and iKalibr integrates inertial and exteroceptive sensors through continuous-time optimization (Chen et al., 2025). He et al. (2026) match micro-Doppler and visual skeleton motion for timing, then associate radar and camera tracks for PnP-based spatial calibration; dynamic time warping associates mapped pedestrian tracks for spatio-temporal synchronization (Li et al., 2026). These methods draw on controlled targets, motion, matched tracks, or auxiliary sensing. EasyCalib predicts window-level corrections from object observations and a nominal sensor transform.

Learning to recover calibration. Early radar-camera learning methods estimated rotation with coarse and fine networks (Schöller et al., 2019). For lidar-camera calibration, CalibNet uses geometric supervision (Iyer et al., 2018), NetCalib uses paired depth representations (Wu et al., 2021), LCCNet uses cost volumes (Lv et al., 2021), and CalibDepth combines depth estimation with iterative refinement (Zhu et al., 2023). DeepI2P combines learned classification with pose optimization (Li & Lee, 2021). RC-AutoCalib learns online radar-camera spatial calibration from front-view and bird's-eye-view features with supervised matching (Luu et al., 2025); CLRNet covers cameras, lidar, and 4D radar (Kegl et al., 2026). Perturbation construction and preprocessing also shape the regression task (Huang et al., 2025); our protocol specifies the prediction coordinates, nominal inputs, output bounds, and data selection.

Cross-modal object evidence. CenterFusion associates radar returns with image detections (Nabati & Qi, 2021), and CramNet uses ray-constrained cross-attention to accommodate depth and elevation uncertainty (Hwang et al., 2022). Geometric Transformer demonstrates the value of contextual geometry for low-overlap point registration (Qin et al., 2022). EasyCalib uses soft interactions between object observations, with reported time differences routed into a learned temporal-consensus module. Timestamp-only controls quantify timing information carried by reported timestamps.

3 METHOD

3.1 CALIBRATION VARIABLES AND COORDINATE CONVENTIONS

For sensor pair g, let $\mathcal{C} = \{(\mathbf{c}_i, \mathbf{a}_i^C, t_i^C)\}$ and $\mathcal{R} = \{(\mathbf{r}_j, \mathbf{a}_j^R, t_j^R)\}$ denote model-input visual and native radar observations, with positions, attributes, and reported timestamps. Their point counts and sampling times can differ. The known nominal transform $N \in \mathrm{SE}(3)$ maps native radar coordinates to camera coordinates. A camera-side perturbation $E = (R, \mathbf{d})$ and an added radar timestamp offset b define the learning targets:

$$(\widehat{E}, \widehat{b}) = F_\Theta(\mathcal{C}, \mathcal{R}; N, g), \qquad \widehat{T}_f = \widehat{E}N. \tag{1}$$

The reported $\widehat{E}$ is the projected output defined below. We write E_{gt} and b_{gt} for the injected pose and offset labels. We parameterize E by $\mathbf{e} = (d_x, d_y, d_z, \phi, \theta, \psi)^\top$, with $R = R_z(\psi)R_y(\theta)R_x(\phi)$. In the optical camera frame (x right, y down, z forward), ϕ, θ, and ψ rotate about right, down, and forward, respectively; d_z is depth, not installation height. Homogeneous transforms act on positions; their rotation blocks act on heading and velocity.

Let $\mathbf{z}(t)$ be a target trajectory in the perturbed camera frame. For small pose errors, write $\widehat{E}E_{\mathrm{gt}}^{-1}$ as a translation error $\Delta\mathbf{d}$ and rotation error $\Delta\boldsymbol{\omega}$. Comparing the transformed point at t with the visual trajectory at $t - \Delta b$ gives the first-order residual

$$\boldsymbol{\delta}_{\mathrm{align}}(t) \simeq \Delta\mathbf{d} - [\mathbf{z}(t)]_\times \Delta\boldsymbol{\omega} + \dot{\mathbf{z}}(t)\Delta b, \qquad \Delta b = \widehat{b} - b_{\mathrm{gt}}. \tag{2}$$

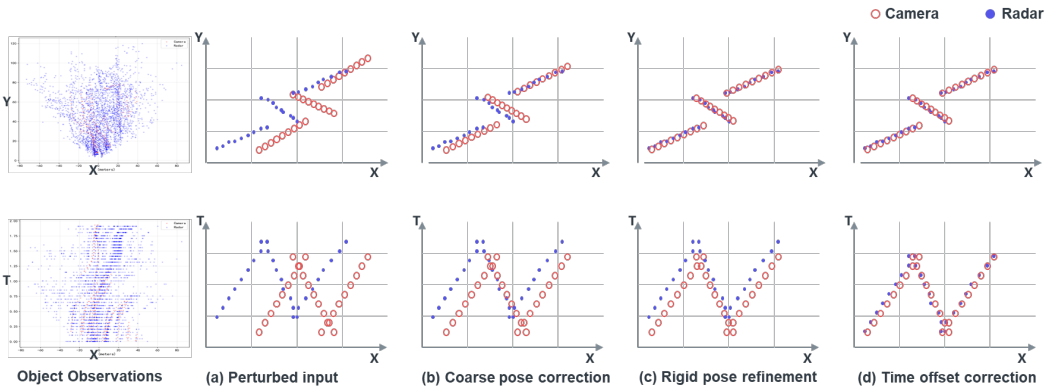

Figure 2: Schematic of the three correction stages. Red circles denote camera observations and blue dots radar observations; columns (a)–(d) show perturbed input, coarse pose correction, composed rigid refinement, and time-offset correction. The upper row depicts planar position and the lower row position versus reported time. The temporal branch re-warps the original visual observations using the complete spatial estimate.

The sign of the temporal term follows $t^R_{\text{corr}} = t^R - \widehat{b}$. Similar target velocities can make $\dot{\mathbf{z}}\Delta b$ difficult to distinguish from translation. Variation in geometry and motion strengthens these constraints. Height-related pose estimates rely heavily on visual 3D structure and the configuration prior because radar elevation is weak.

Figure 2 illustrates the correction sequence: coarse spatial estimation reduces the planar discrepancy, rigid refinement further aligns positions, and the time branch addresses the remaining displacement along moving trajectories.

3.2 Perturbations and Asynchronous Observation Windows

We perturb recorded visual observations while preserving radar returns. For a recorded homogeneous camera point $\overline{\mathbf{c}}^{\text{rec}}_i$, the injected input point $\overline{\mathbf{c}}'_i$ is

$$\overline{\mathbf{c}}'_i = E_{\text{gt}}\overline{\mathbf{c}}^{\text{rec}}_i, \qquad T_{f,\text{gt}} = E_{\text{gt}}N, \qquad E_{\text{gt}} = T_{f,\text{gt}}N^{-1}. \tag{3}$$

The injected observations form $\mathcal{C}$ during evaluation. Position, heading, and velocity are updated consistently; object dimensions are unchanged. This simulates calibration error on recorded observations without modeling new scattering or occlusion after physical reinstallation.

Let s denote acquisition time and ϵ the native timestamp error. An independently sampled, window-constant offset is added to radar timestamps:

$$t^C_i = s^C_i + \epsilon^C_i, \qquad t^R_j = s^R_j + \epsilon^R_j + b_{\text{gt}}, \qquad t^R_{j,\text{corr}} = t^R_j - \widehat{b}. \tag{4}$$

The native timing factors ϵ are not separately labeled. Consequently, b_{gt} measures recovery of an added offset, not the complete physical clock error. We crop both streams to the intersection of their reported-time intervals, mask each single-modality tail, and use the first retained camera timestamp as their shared origin. This preprocessing uses neither b_{gt} nor pre-injection radar times. Native asynchrony remains, and the retained reported-time boundaries can still carry offset information.

3.3 XYT Representation

A fixed homogeneous axis conversion A maps native radar axes (forward, left, up) to common axes (right, forward, up). With the injected camera observation $\overline{\mathbf{c}}'_i$ from Eq. (3), physical coordinates and the position-encoder input are

$$\begin{aligned} \overline{\mathbf{p}}^C_i &= AN^{-1}\overline{\mathbf{c}}'_i, & \overline{\mathbf{p}}^R_j &= A\overline{\mathbf{r}}_j, \\ \mathbf{u}^m_k &= \left(\frac{X^m_k}{50\,\text{m}}, \frac{Y^m_k}{50\,\text{m}}, \frac{t^m_k - t_0}{10\,\text{s}}\right)^\top, & m &\in \{C, R\}. \end{aligned} \tag{5}$$

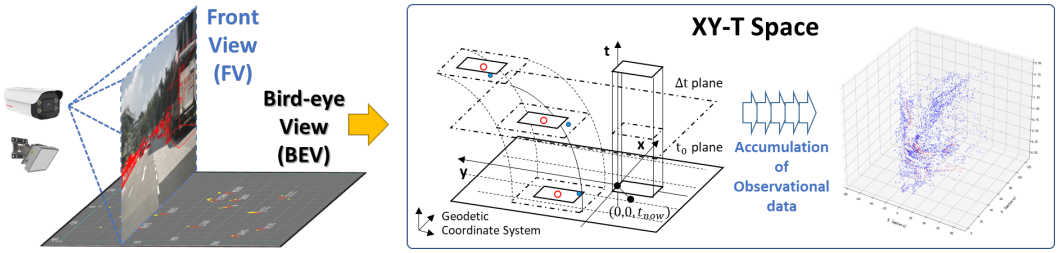

Figure 3: Conceptual construction of XYT evidence from camera detections and radar returns. The left panel shows front-view camera evidence and a planar radar view; the center depicts observations accumulated across time; the right shows an XYT point cloud. In the model, the position encoder uses common-frame planar X, Y and each modality's reported time t, while visual height and other native properties remain attributes. The center panel is schematic; Eq. (5) defines the model coordinates.

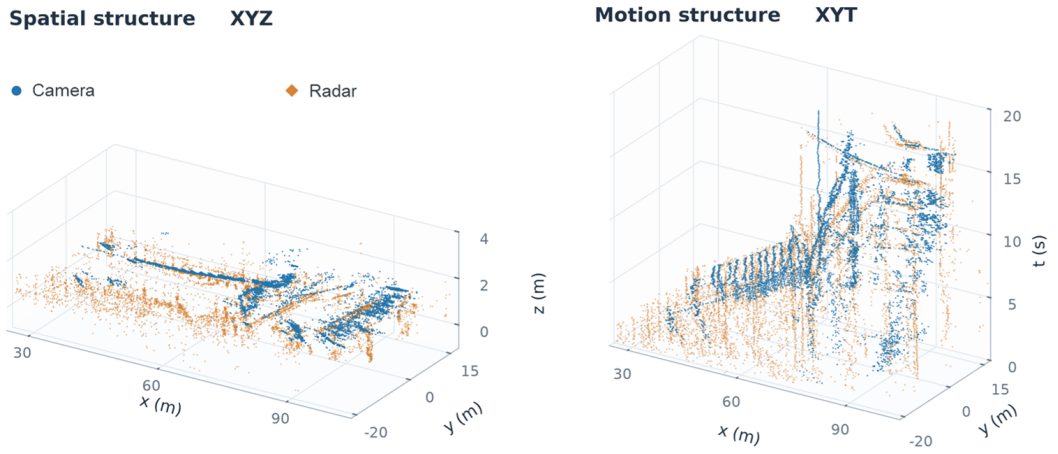

Figure 4: XYZ and XYT views of the same 20-second observation window. Blue points are visual detection centers and orange points are radar returns. Dataset calibration and ego poses place points in a common display frame for this illustration; this motion-compensated visualization differs from model input construction at native event times. Height is enlarged fivefold.

Each stream contains up to 2,048 masked observations. The 16-channel visual tokens use detector-produced 3D boxes, so center, height, heading, and motion attributes depend on the visual frontend. The 8-channel native radar tokens use planar returns and radial Doppler; radar supplies neither object height nor orientation, and Doppler is not a complete object velocity. Positions remain in their event-time sensor frames rather than being aligned by ego poses; ego poses are used only for the visualization in Figure 4. Reported time is never rotated as a physical coordinate. Figures 3 and 4 show the construction.

3.4 Network

Figure 5 specifies the coarse, residual, and temporal branches. The two spatial backbones are independent. Coarse-pose-compensated observations feed the residual branch; the temporal branch instead re-warps the original visual input using the complete spatial estimate.

The coarse estimate $\widehat{E}_c$ and residual $\widehat{D}$ compose as

$$\widehat{E}_{\text{raw}} = \widehat{E}_c\widehat{D}, \qquad \widehat{R}_{\text{raw}} = \widehat{R}_c\widehat{R}_d, \qquad \widehat{\mathbf{d}}_{\text{raw}} = \widehat{\mathbf{d}}_c + \widehat{R}_c\widehat{\mathbf{d}}_d. \tag{6}$$

Euler components are converted to matrices before composition. Evaluation uses $\widehat{E}_{\text{out}} = \Pi_{\mathcal{B}}(\widehat{E}_{\text{raw}})$, where $\Pi_{\mathcal{B}}$ clips the six translation–Euler coordinates to the declared task bounds. This known output prior is separate from the $SE(3)$ composition; Appendix B also reports raw outputs.

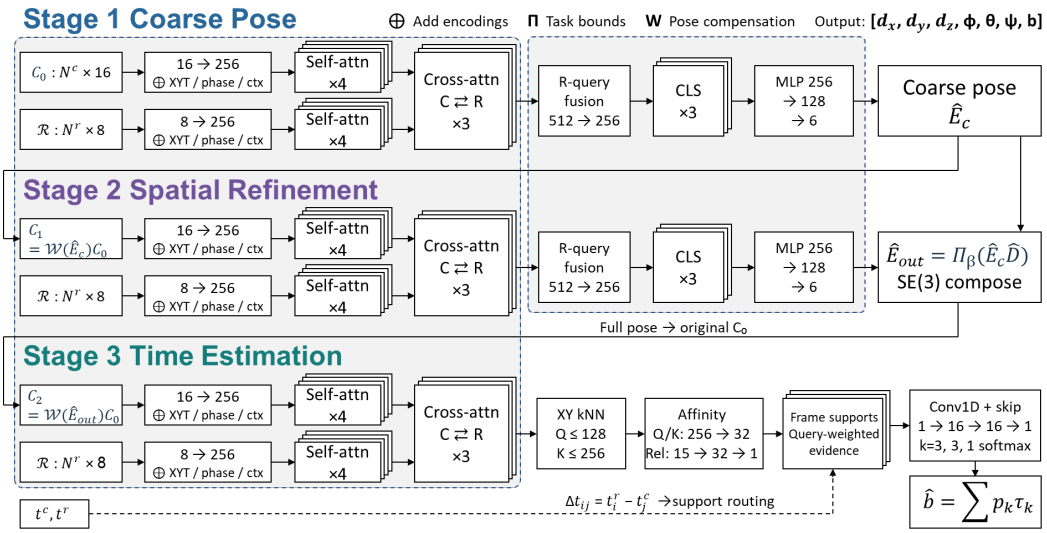

Figure 5: Three-stage network. Camera and radar tokens pass through modality-specific encoders, self- and cross-attention, then separate pose heads. The residual pose is composed in $SE(3)$ and projected to task bounds. For time prediction, reported timestamps bypass the content encoder and route candidate scores to 41 offset supports. W denotes spatial compensation and Π range projection.

3.5 Temporal Consensus and Supervision

Let $\mathcal{C}_0$ denote the original injected visual observations in the common physical coordinates of Eq. (5). The temporal branch compensates them using the complete spatial prediction. Its position operator is

$$W(\widehat{E}) = AN^{-1}\widehat{E}^{-1}NA^{-1}, \qquad \mathcal{C}_2 = W\Big(\mathrm{sg}(\widehat{E}_{\mathrm{out}})\Big)\,\mathcal{C}_0, \tag{7}$$

where sg stops gradients. The operator transforms positions, headings, and velocities according to their geometric roles and preserves timestamps and other attributes.

Reported time bypasses the temporal content encoder. For radar query j and visual candidate i, let $\widetilde{\mathbf{q}}_j^R$ and $\widetilde{\mathbf{k}}_i^C$ be normalized features and $\boldsymbol{\eta}_{ji}$ a 15-dimensional spatial, appearance, and motion descriptor. Candidate affinity is

$$s_{ji} = \mathrm{sigmoid}\Big(h_{\mathrm{rel}}(\boldsymbol{\eta}_{ji}) + (\widetilde{\mathbf{q}}_j^R)^\top \widetilde{\mathbf{k}}_i^C\Big), \qquad \Delta t_{ji} = t_j^R - t_i^C. \tag{8}$$

Scores vote on 41 offset supports $\tau_k = -B + 2Bk/40$, where B is the specified offset half-range. Normalized votes give probabilities p_k and the final estimate $\widehat{b} = \sum_{k=0}^{40} p_k \tau_k$. Queries use an interior reported-time margin of $B + 0.15$ s.

Spatial training minimizes a physically scaled Smooth-L1 loss:

$$\mathcal{L}_{\mathrm{sp}} = \sum_{k=1}^{3} \rho\left(\frac{\widehat{d}_{\mathrm{raw},k} - d_{\mathrm{gt},k}}{1\,\mathrm{m}}\right) + \sum_{k=1}^{3} \rho\left(\frac{\mathrm{wrap}(\widehat{\vartheta}_{\mathrm{raw},k} - \vartheta_{\mathrm{gt},k})}{3^\circ}\right), \tag{9}$$

where $\boldsymbol{\vartheta} = (\phi, \theta, \psi)$, wrap maps each angle difference to its principal interval, and ρ is Smooth-L1 with unit transition. Temporal training uses a normalized Gaussian target distribution q_k over the 41 supports and an absolute-error term:

$$\mathcal{L}_{\mathrm{time}} = -\sum_{k=0}^{40} q_k \log p_k + 0.25\,\frac{|\widehat{b} - b_{\mathrm{gt}}|}{0.12\,\mathrm{s}}. \tag{10}$$

Spatial refinement freezes the coarse network; temporal training freezes both pose networks. Evaluation uses predicted compensation, without labels or unperturbed radar times (Appendix A).

3.6 OBSERVATION ADMISSION

We admit windows with vehicle observations both within and beyond 40 m and in at least half of each modality's native keyframes. The offline radar selection uses annotated vehicle boxes for classification and deduplication; identities do not enter the model. Appendix A gives the thresholds and rules, and Appendix D reports all-window performance.

4 EXPERIMENTS

4.1 PROTOCOL AND METRICS

We use nuScenes (Caesar et al., 2020) with PGD detections (Wang et al., 2022), an adapted RepDETR3D frontend, and V2X-Radar (Yang et al., 2025). The admitted PGD split has 2,145/267/267 TRAIN/VAL/TEST scene–sensor pairs from disjoint recording logs. Because the TEST population has been used during development, we report it as internal. Appendix A details the other splits and evaluation protocol.

Spatial ranges R1–R4 have per-axis translation/angle bounds $(0.25\,\mathrm{m}, 10^\circ)$, $(1.5\,\mathrm{m}, 20^\circ)$, $(5\,\mathrm{m}, 25^\circ)$, and $(10\,\mathrm{m}, 30^\circ)$; temporal half-ranges D1–D4 are 120, 250, 500, and 1,000 ms. Checkpoints are selected on validation; each test row scores 32 fixed draws per pair.

We report per-axis translation MAE in cm, wrapped Euler-angle MAE in degrees, and added-offset MAE in ms, weighting every registered view equally. Three-run error bars are sample standard deviations across continuations from one common initialization.

4.2 SPATIAL AND TEMPORAL RANGES

Table 1 reports all seven parameter errors at R2–D2 and the widest R4–D4 setting. At R4–D4, refinement improves every spatial axis, while offset error falls 52.3% below zero prediction. Appendix B gives the remaining ranges and the spread across three R2–D2 continuations.

Table 1: EasyCalib (ours) coarse and final seven-parameter MAE on 8,544 admitted nuScenes/PGD test views per row. R2–D2 denotes ± 1.5 m, $\pm 20^\circ$, ± 250 ms; R4–D4 denotes ± 10 m, $\pm 30^\circ$, ± 1 s. Each range uses one selected checkpoint. Blue marks the lower error within each range.

EasyCalib stage	Translation MAE (cm)			Rotation MAE (°)			Offset MAE (ms)
	d_x	d_y	d_z	ϕ	θ	ψ	b
R2–D2 coarse	59.20	25.26	49.47	0.553	1.778	0.920	93.93
R2–D2 final	**52.86**	**17.81**	**45.38**	**0.268**	**1.578**	**0.505**	**93.47**
Reduction vs. coarse (%)	10.7%	29.5%	8.3%	51.5%	11.3%	45.1%	0.5%
R4–D4 coarse	73.91	35.69	64.32	0.679	1.842	1.185	242.42
R4–D4 final	**66.12**	**29.45**	**60.10**	**0.397**	**1.695**	**0.758**	**238.45**
Reduction vs. coarse (%)	10.5%	17.5%	6.6%	41.5%	8.0%	36.0%	1.6%

Table 2 compares matched-view controls, including a planar coherent point drift (CPD) adapter (Myronenko & Song, 2010) and timestamp-only estimates. Table 3 places our R2–D2 result beside published R2 spatial values. Different inputs and protocols preclude a controlled ranking.

4.3 TIMING CONTROLS AND CROSS-DOMAIN RESULTS

At R2–D2, learned timing reduces the zero-offset MAE by 25.3%, while the endpoint statistic reaches a lower error. Window boundaries thus carry useful timing information in the injected-offset task. Additional timing controls and failure counts are in Appendix C.

Table 2: EasyCalib (ours) and matched controls on 8,544 R2–D2 views. Translation and rotation means average three axis MAEs. CPD is coherent point drift and fits 3DoF. The last two rows reuse our refined pose; dashes mark unchanged pose values. Blue marks the lowest error in each comparable column.

Method	Translation MAE (cm)				Rotation MAE (°)				Offset MAE (ms)
	Mean	d_x	d_y	d_z	Mean	ϕ	ψ	θ	b
Planar CPD (3DoF)	115.18	131.78	76.38	137.38	9.788	10.109	10.173	9.081	–
EasyCalib (ours): coarse	44.64	59.20	25.26	49.47	1.084	0.553	0.920	1.778	93.93
EasyCalib (ours): refined	**38.68**	**52.86**	**17.81**	**45.38**	**0.784**	**0.268**	**0.505**	**1.578**	93.47
Ours + zero-offset control	–	–	–	–	–	–	–	–	125.04
Ours + endpoint statistic	–	–	–	–	–	–	–	–	**68.48**

Table 3: Published spatial-only R2 baselines from RC-AutoCalib Table 1 (Luu et al., 2025) and EasyCalib (ours) at R2–D2, averaged over three training continuations. Published axes retain source labels; our axes follow the camera-optical convention used in Table 2. Blue marks column minima among published baselines and identifies our row. Input and evaluation protocols differ between groups.

Method	Translation MAE (cm)				Rotation MAE (°)				Offset MAE (ms)
	Mean	X	Y	Z	Mean	roll	pitch	yaw	b
LCCNet-3	89.672	**71.660**	89.605	107.751	2.156	1.526	2.364	2.579	–
NetCalib2	71.037	76.001	57.204	79.906	2.778	1.465	4.688	2.180	–
CalibDepth	55.380	77.146	12.918	76.078	1.686	1.149	0.808	3.102	–
RC-AutoCalib	**47.537**	74.777	**5.415**	**62.420**	**0.852**	**0.3597**	**0.4423**	**1.7544**	–
	Mean	d_x	d_y	d_z	Mean	ϕ	ψ	θ	b
EasyCalib (ours, R2–D2)	**38.90**	**53.35**	**18.01**	**45.34**	**0.788**	**0.267**	**0.515**	**1.583**	**94.29**

On the 57 common PGD/RepDETR3D test pairs, the adapted RepDETR3D system lowers rotation error but raises translation and offset error. Each system is trained on its own frontend's observations. On V2X-Radar, offset error remains close to zero prediction (Tables 8 and 9).

4.4 Ablations and operational scope

Removing predicted prealignment raises rotation MAE from $0.804°$ to $1.008°$ and also increases translation error. With pose frozen, removing learned association raises offset MAE from 114.91 to 125.04 ms. Table 10 reports the other component controls and their distinct training budgets.

Across all 539 PGD test pairs, including admission-rejected windows, the model yields 46.18 cm, $1.478°$, and 103.81 ms. Batch-one model inference takes 28.40 ms on an RTX 5090 after features are cached; Appendix D reports coverage and latency boundaries.

5 Limitations and conclusion

EasyCalib recovers six pose corrections and an added radar timestamp offset from accumulated camera detections and native radar returns. XYT encoding retains reported time in the position representation, and rigid refinement gives the temporal branch a compensated window for soft candidate association. On admitted nuScenes/PGD windows, refinement improves every spatial axis with injected perturbations of up to ± 10 m per translation axis, $\pm 30°$ per rotation axis, and ± 1 s of added radar timestamp offset.

The internal test panel was inspected during development, and radar admission uses annotated vehicle boxes. With perturbations of up to ± 1.5 m per translation axis, $\pm 20°$ per rotation axis, and ± 250 ms of added offset, a timestamp endpoint statistic outperforms the learned temporal branch. This shows that window boundaries carry useful timing evidence. Sealed-panel evaluation and tests with physical clock changes and continuous operation are needed to assess deployment.

AI USE STATEMENT

Generative AI tools assisted with literature organization, review of mathematical formulations and experimental design, interpretation of recorded results, figure and document code, translation, and prose editing. We checked numerical claims against recorded experiment artifacts. The authors are responsible for the correctness, originality, references, and final contents of the submission.

ETHICS STATEMENT

The experiments use existing driving datasets and offline perception outputs, with no new collection of human-subject data. Dataset access and redistribution remain subject to the original terms. A predicted calibration correction can degrade an otherwise accurate system; deployment requires application-specific checks of geometry, timing, and downstream perception.

REPRODUCIBILITY STATEMENT

Appendices specify the coordinate convention, observation interface, admission criteria, splits, perturbation ranges, training budgets, checkpoint selection, and metric aggregation. The source package contains a machine-readable numerical snapshot and deterministic table-generation code. Each reported row is associated with an experiment family, evaluation population, and source digest. Independent reaggregation checked numerical consistency across 42 completed test panels comprising 329,344 stored prediction rows.

REFERENCES

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Caesar_nuScenes_A_Multimodal_Dataset_for_Autonomous_Driving_CVPR_2020_paper.html.

Shuolong Chen, Xingxing Li, Shengyu Li, Yuxuan Zhou, and Xiaoteng Yang. iKalibr: Unified targetless spatiotemporal calibration for resilient integrated inertial systems. *IEEE Transactions on Robotics*, 41:1618–1638, 2025. doi: 10.1109/TRO.2025.3532506. URL https://doi.org/10.1109/TRO.2025.3532506.

Lei Cheng, Arindam Sengupta, and Siyang Cao. 3d radar and camera co-calibration: A flexible and accurate method for target-based extrinsic calibration. In *IEEE Radar Conference*, pp. 1–6, 2023. doi: 10.1109/RadarConf2351548.2023.10149669. URL https://radar.ece.arizona.edu/publications/3d-radar-camera-calibration-2023/.

Jiayin Deng, Zhiqun Hu, Zhaoming Lu, and Xiangming Wen. FusionCalib: Automatic extrinsic parameters calibration based on road plane reconstruction for roadside integrated radar camera fusion sensors. *Pattern Recognition Letters*, 176:7–13, 2023. doi: 10.1016/j.patrec.2023.10.015. URL https://doi.org/10.1016/j.patrec.2023.10.015.

Ozan Durmaz and Hakan Cevikalp. Targetless radar–camera calibration via trajectory alignment. *Sensors*, 25(24):7574, 2025. doi: 10.3390/s25247574. URL https://www.mdpi.com/1424-8220/25/24/7574.

Wei He, Gui Zhang, Xinyu Liu, Zhenmiao Deng, Xinglong Li, and Yishan Ye. A low-cost temporal-spatial calibration method for radar and camera systems. *Expert Systems with Applications*, 306: 130879, 2026. doi: 10.1016/j.eswa.2025.130879.

Shujuan Huang, Chunyu Lin, and Yao Zhao. What really matters for learning-based LiDAR-camera calibration. In *Proceedings of the 7th ACM International Conference on Multimedia in Asia*, pp. 1–8, 2025. doi: 10.1145/3743093.3771027. URL https://doi.org/10.1145/3743093.3771027.

Jyh-Jing Hwang, Henrik Kretzschmar, Joshua Manela, Sean Rafferty, Nicholas Armstrong-Crews, Tiffany Chen, and Dragomir Anguelov. CramNet: Camera-radar fusion with ray-constrained cross-attention for robust 3d object detection. In *European Conference on Computer Vision*, pp. 388–405, 2022. doi: 10.1007/978-3-031-19839-7_23. URL https://arxiv.org/abs/2210.09267.

Ganesh Iyer, R. Karnik Ram, J. Krishna Murthy, and K. Madhava Krishna. CalibNet: Geometrically supervised extrinsic calibration using 3d spatial transformer networks. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1110–1117, 2018. doi: 10.1109/IROS.2018.8593693. URL https://arxiv.org/abs/1803.08181.

Marcell Kegl, Andras Palffy, Csaba Benedek, and Dariu M. Gavrila. CLRNet: Targetless extrinsic calibration for camera, lidar and 4d radar using deep learning. *arXiv preprint arXiv:2603.15767*, 2026. URL https://arxiv.org/abs/2603.15767.

Jiaxin Li and Gim Hee Lee. DeepI2P: Image-to-point cloud registration via deep classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15960–15969, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Li_DeepI2P_Image-to-Point_Cloud_Registration_via_Deep_Classification_CVPR_2021_paper.html.

Jingjing Li, Juan Liu, Xiuping Li, Chengliang Zhong, and Xiyan Sun. A dtw-based spatio-temporal synchronization method for radar and camera fusion. *Sensors*, 26(7):2108, 2026. doi: 10.3390/s26072108.

Xinyu Liu, Zhenmiao Deng, and Gui Zhang. Targetless radar–camera extrinsic parameter calibration using track-to-track association. *Sensors*, 25(3):949, 2025. doi: 10.3390/s25030949. URL https://www.mdpi.com/1424-8220/25/3/949.

Van-Tin Luu, Yon-Lin Cai, Vu-Hoang Tran, Wei-Chen Chiu, Yi-Ting Chen, and Ching-Chun Huang. RC-AutoCalib: An end-to-end radar-camera automatic calibration network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Luu_RC-AutoCalib_An_End-to-End_Radar-Camera_Automatic_Calibration_Network_CVPR_2025_paper.html.

Xudong Lv, Boya Wang, Ziwen Dou, Dong Ye, and Shuo Wang. LCCNet: LiDAR and camera self-calibration using cost volume network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2894–2901, 2021. URL https://openaccess.thecvf.com/content/CVPR2021W/WAD/html/Lv_LCCNet_LiDAR_and_Camera_Self-Calibration_Using_Cost_Volume_Network_CVPRW_2021_paper.html.

Andriy Myronenko and Xubo Song. Point set registration: Coherent point drift. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(12):2262–2275, 2010. doi: 10.1109/TPAMI.2010.46.

Ramin Nabati and Hairong Qi. CenterFusion: Center-based radar and camera fusion for 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1527–1536, 2021. URL https://openaccess.thecvf.com/content/WACV2021/html/Nabati_CenterFusion_Center-Based_Radar_and_Camera_Fusion_for_3D_Object_Detection_WACV_2021_paper.html.

Juraj Peršić, Luka Petrović, Ivan Marković, and Ivan Petrović. Spatiotemporal multisensor calibration via gaussian processes moving target tracking. *IEEE Transactions on Robotics*, 37(5):1401–1415, 2021. doi: 10.1109/TRO.2021.3061364. URL https://lamor.fer.hr/_news/26917/2020-persic-temporal-tro.pdf.

Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11143–11152, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Qin_Geometric_Transformer_for_Fast_and_Robust_Point_Cloud_Registration_CVPR_2022_paper.html.

Christoph Schöller, Maximilian Schnettler, Annkathrin Krämmer, Gereon Hinz, Maida Bakovic, Müge Güzet, and Alois Knoll. Targetless rotational auto-calibration of radar and camera for intelligent transportation systems. In *IEEE Intelligent Transportation Systems Conference*, pp. 3934–3941, 2019. doi: 10.1109/ITSC.2019.8917135. URL https://arxiv.org/abs/1904.08743.

Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 1475–1485, 2022. URL https://proceedings.mlr.press/v164/wang22i.html.

Emmett Wise, Juraj Peršić, Christopher Grebe, Ivan Petrović, and Jonathan Kelly. A continuous-time approach for 3d radar-to-camera extrinsic calibration. In *IEEE International Conference on Robotics and Automation*, pp. 13164–13170, 2021. doi: 10.1109/ICRA48506.2021.9561938. URL https://arxiv.org/abs/2103.07505.

Emmett Wise, Qilong Cheng, and Jonathan Kelly. Spatiotemporal calibration of 3-d millimetre-wavelength radar-camera pairs. *IEEE Transactions on Robotics*, 39(6):4552–4566, 2023. doi: 10.1109/TRO.2023.3311680. URL https://arxiv.org/abs/2211.01871.

Shan Wu, Amnir Hadachi, Damien Vivet, and Yadu Prabhakar. NetCalib: A novel approach for LiDAR-camera auto-calibration based on deep learning. In *2020 25th International Conference on Pattern Recognition*, pp. 6648–6655, 2021. doi: 10.1109/ICPR48806.2021.9412653. URL https://doi.org/10.1109/ICPR48806.2021.9412653.

Lei Yang, Xinyu Zhang, Jun Li, Chen Wang, Jiaqi Ma, Zhiying Song, Tong Zhao, Ziying Song, Li Wang, Mo Zhou, Yang Shen, Kai Wu, and Chen Lv. V2X-Radar: A multi-modal dataset with 4D radar for cooperative perception. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. doi: 10.52202/085713-3789. URL https://papers.neurips.cc/paper_files/paper/2025/hash/a501f3238029713afdad57ce7924667a-Abstract-Datasets_and_Benchmarks_Track.html.

Jiangtong Zhu, Jianru Xue, and Pu Zhang. CalibDepth: Unifying depth map representation for iterative LiDAR-camera online calibration. In *IEEE International Conference on Robotics and Automation*, pp. 726–733, 2023. doi: 10.1109/ICRA48891.2023.10161575. URL https://doi.org/10.1109/ICRA48891.2023.10161575.

A Implementation and experimental protocol

A.1 Network and inference interface

Figures 1 and 5 show the architecture and staged flow. Validity masks exclude padded tokens from attention and aggregation. Inference receives the nominal transform, sensor-pair identity, native observations, reported timestamps, and masks; it uses no perturbation labels, unperturbed radar timestamps, event identifiers, or ground-truth associations. Temporal training freezes the spatial networks. Training-only empirical residual compensation assists fitting, while evaluation uses predicted compensation.

A.2 Sampling, split, and optimization

The primary nuScenes/PGD split has 2,145, 267, and 267 admitted scene–sensor pairs for training, validation, and test, drawn from 55, 6, and 7 disjoint recording logs, respectively. This internal test population was used during development and underwent earlier split migrations. RepDETR3D has 2,033/284/58 admitted pairs. V2X-Radar has 65/7/15. The frontend comparison evaluates separately adapted systems on the 57 PGD/RepDETR3D test pairs common to both.

Each admitted training pair has 512 pre-generated joint perturbation draws. Validation uses eight fixed draw slots per pair; selected-checkpoint reporting uses 32 fixed slots. Thus each principal test row has 8,544 views. Metrics weight all registered views equally. A complete checkpoint is selected on validation, and the initial checkpoint is eligible. Reported three-run uncertainty is the sample standard deviation across continuations from one shared R4 initialization.

Training uses two 2,048-token streams, BF16, batch size 80, and an RTX 5090. Spatial refinement uses 4,096 optimizer updates per range. Its learning rate warms from 5×10^{-6} to 2×10^{-5} over 128 updates, then follows a cosine decay; gradient norm is clipped at 10. Temporal adaptation uses its own registered range-specific schedule. A new training cell resets AdamW; an explicit continuation restores model, optimizer, random-number, and committed-sample-cursor state. The exact per-cell configs, parent checkpoint digests, data manifests, and source hashes are recorded in the evidence snapshot and summarized in `data/results.json`.

A.3 Data admission and evaluation failures

Admission is evaluated before virtual perturbation. In each modality, at least four vehicle sightings must occur at distances at most 40 m and at least four beyond 40 m; sightings must be present in at least half of native keyframes. Empty keyframes count in the denominator. Repeated sightings across time are allowed, while same-vehicle radar returns within a packet are deduplicated. Camera counts use post-NMS detections. Offline radar admission uses annotated vehicle boxes for classification and deduplication. Deployment would require a label-free selection rule.

The all-window evaluation scores every registered view, including identity/no-update fallbacks when an input modality or usable overlap is absent. We report results for all, admitted, and rejected windows. Empty-modality timing placeholders occur only in fully masked views that do not enter the model.

B Additional spatial, temporal, and frontend results

B.1 Perturbation-range campaigns

Tables 4 and 5 report the range campaigns summarized in the main text. Spatial-range rows use D4; their smallest-range row is an earlier fixed reference. Temporal-range rows use R2. Changing the offset half-range also changes the zero-prediction reference.

Table 4: EasyCalib (ours) across spatial perturbation ranges with added offset up to ± 1 s. The time branch is fixed; the smallest-range row uses an earlier checkpoint. Ranges differ, so rows are not ranked.

Pose perturbation	Translation MAE (cm)			Rotation MAE (°)			Offset MAE (ms)
	d_x	d_y	d_z	ϕ	θ	ψ	b
± 0.25 m, $\pm 10°$	20.55	9.69	19.56	0.243	1.538	0.411	227.88
± 1.5 m, $\pm 20°$	52.92	17.98	46.05	0.267	1.604	0.515	232.34
± 5 m, $\pm 25°$	64.19	22.53	54.10	0.303	1.608	0.590	236.28
± 10 m, $\pm 30°$	66.12	29.45	60.10	0.397	1.695	0.758	238.45

Table 5: EasyCalib (ours) across added-offset ranges at the fixed spatial range of ± 1.5 m and $\pm 20°$. Each row uses one selected checkpoint; rows have different offset difficulty and are not ranked.

Added-offset range	Translation MAE (cm)			Rotation MAE (°)			Offset MAE (ms)
	d_x	d_y	d_z	ϕ	θ	ψ	b
± 120 ms	52.80	17.85	45.36	0.270	1.576	0.505	58.34
± 250 ms	52.86	17.81	45.38	0.268	1.578	0.505	93.47
± 500 ms	52.80	17.85	45.64	0.268	1.589	0.509	148.80
$\pm 1{,}000$ ms	52.92	17.98	46.05	0.267	1.604	0.515	222.79

B.2 Per-continuation results and output projection

Table 6 reports the three R2–D2 continuations and their shared-initialization mean and sample standard deviation. Each row comes from one selected checkpoint. Range projection is an explicit prior, so Table 7 contrasts raw predictions with bounded outputs.

Table 6: EasyCalib (ours): three R2–D2 training continuations from one shared initialization on the same 8,544 test views. Blue marks the lowest run in each column; SD is across continuations.

Training run	Translation MAE (cm)	Rotation MAE (°)	Offset MAE (ms)
Training continuation 1	**38.68**	**0.784**	93.47
Training continuation 2	39.18	0.785	**93.37**
Training continuation 3	38.83	0.796	96.02
Mean $\pm$ SD	38.90 $\pm$ 0.26	0.788 $\pm$ 0.007	94.29 $\pm$ 1.50

B.3 Dataset and frontend results

Table 8 reports adapted-system results on the internal test populations and the matched 57-pair subset. V2X-Radar evaluates offsets derived from frame timing; recovery of physical clock offsets remains future work (Section 5).

B.4 Planar registration control

The same-panel CPD control in Table 2 uses a third-party coherent-point-drift solver with a fixed-scale variance correction and our planar radar–camera adapter. It estimates three pose degrees of freedom and supplies no time offset. The solver reported nonconvergence on 8,504 of 8,544 views; those outputs remain in the spatial means.

C Temporal controls and ablations

C.1 Timestamp-only and registration controls

Table 9 compares matched-view timing controls. The endpoint statistic is best on nuScenes; trajectory registration retains 130 fallback-scored failures.

Table 7: EasyCalib (ours): R2–D2 mean per-axis error before and after projection to the declared pose range, on identical test views. Blue marks the lowest training continuation within each column.

Training run	Translation MAE (cm)		Rotation MAE (°)	
	Raw	Range-projected	Raw	Range-projected
Training continuation 1	48.73	**38.68**	**0.811**	**0.784**
Training continuation 2	49.22	39.18	0.812	0.785
Training continuation 3	**48.70**	38.83	0.824	0.796

Table 8: EasyCalib (ours) trained separately with each perception frontend at R2–D2. The bottom two rows use the same 57 scene–sensor pairs (1,824 views); blue marks their column minima. Full-population rows have different pair counts and are not ranked.

Perception frontend and test set	Pairs	Translation MAE (cm)	Rotation MAE (°)	Offset MAE (ms)
PGD (all admitted)	267	38.68	0.784	93.47
RepDETR3D (all admitted)	58	34.38	0.585	105.45
V2X-Radar (all admitted)	15	64.04	1.969	124.08
PGD (common pairs)	57	**28.91**	0.645	**86.74**
RepDETR3D (common pairs)	57	34.68	**0.584**	105.57

Table 9: EasyCalib (ours) and added-offset controls on matched test views. P95 is the 95th percentile of absolute error. Registration failures remain scored; dashes mean no iterative solver. Blue marks the lowest error within each dataset block.

Offset estimator	Offset MAE (ms)	Offset P95 (ms)	Solver failures
nuScenes / PGD (8,544 views)			
EasyCalib (ours)	93.47	231.84	–
Planar CPD time adapter	147.54	398.46	0
Window endpoint statistic	**68.48**	**143.30**	–
Frame-phase statistic	126.72	248.50	–
Frame phase + endpoint	71.99	159.48	–
Trajectory registration	200.44	449.04	130
Zero added offset	125.04	237.48	–
V2X-Radar (480 views)			
EasyCalib (ours)	124.08	231.50	–
Window endpoint statistic	65.04	119.06	–
Frame-phase statistic	121.46	200.00	–
Frame phase + endpoint	**60.00**	**100.00**	–
Zero added offset	125.86	238.89	–

C.2 MATCHED COMPONENT CONTROLS

Table 10 reports 15 component controls. Temporal rows share a frozen pose model and training budget, using a checkpoint distinct from the main comparison. Spatial and fresh-training blocks differ in initialization or budget, so component effects are read within each block. Learned association, query weighting, and consensus correction reduce offset error in the temporal block.

Table 10: EasyCalib (ours) component controls on the internal test population. Temporal rows share the fixed pose errors shown in the first row; dashes in other blocks mean an output was not estimated. Blocks differ in initialization or budget. Blue marks column minima within each block.

Model variant	Translation MAE (cm)	Rotation MAE (°)	Offset MAE (ms)
Temporal module			
Full temporal module (ours)	38.68	0.784	114.91
Uniform query weighting	–	–	121.96
No Doppler relation	–	–	114.99
No consensus correction	–	–	118.72
No training residual compensation	–	–	114.99
Factorized residual compensation	–	–	**114.90**
Time encoder only	–	–	125.01
No learned association	–	–	125.04
Spatial module			
Full spatial module (ours)	39.33	0.804	–
No predicted prealignment	43.15	1.008	–
No velocity in refinement	**38.53**	**0.795**	–
Fresh-training controls			
Full EasyCalib (ours), fresh training	**53.86**	**1.184**	–
Calibration prior only	75.39	10.082	–
Camera observations only	75.39	10.082	–
Radar observations only	75.34	10.082	–

D ADMISSION COVERAGE AND RUNTIME

D.1 ADMISSION COVERAGE

Table 11 reports the admitted, rejected, and combined PGD test populations. Rejected windows have larger translation and rotation errors.

Table 11: EasyCalib (ours) R2–D2 admission coverage on the internal PGD test population. Each scene–sensor pair yields 32 scored views; rejected pairs remain in the all-pairs result, with identity/no-update fallback where needed.

Test population	Pairs	Translation MAE (cm)	Rotation MAE (°)	Offset MAE (ms)
All registered pairs	539	46.18	1.478	103.81
Admitted pairs	267	38.68	0.784	93.47
Admission-rejected pairs	272	53.53	2.161	113.96

D.2 CACHED-FEATURE MODEL RUNTIME

Batch-one timing on an RTX 5090 uses BF16, 2,048 tokens per stream, ten warm-ups, and two repeats over 32 pairs. The 28.40 ms PGD model time excludes detection, window collection, cached-feature loading, transfer, and physical acquisition.

E RADAR-SIDE R1 SIMULATION STUDY

Radar-side virtual-pose tests appear in prior work (Luu et al., 2025), so we include an R1 pose-only diagnostic. A rigid warp preserves which radar returns were detected; we therefore do not treat it as a physical remounting simulation. We fix camera observations, perturb radar points by ± 0.25 m and $\pm 10^\circ$ per axis, and add no time offset. The model receives nominal geometry and radar radial directions. Table 12 places our validation-selected result and frozen parent on 8,544 previously inspected internal test views below the published R1 baselines.

Table 12: Published R1 baselines from Table 1 of Luu et al. (2025) and EasyCalib (ours) radar-side pose-only tests. Source and local axes and protocols differ. SO(3) P95 is the 95th-percentile rotation error; dashes mark unavailable values. Blue marks column minima within each group separately.

Method / stage	Translation MAE (cm)				Rotation MAE (°)				SO(3) P95 (°)
	Mean	X	Y	Z	Mean	roll	pitch	yaw	
RC-AutoCalib (published R1)	**9.498**	12.563	**3.295**	12.635	**0.427**	0.130	**0.198**	0.953	–
LCCNet-1	16.531	22.992	17.648	**8.954**	1.603	**0.123**	3.130	1.556	–
NetCalib2	12.297	**12.532**	12.076	12.284	1.205	0.387	2.289	**0.941**	–
CalibDepth	12.608	12.860	12.250	12.715	0.807	0.390	0.345	1.686	–
Schöller et al. (coarse)	–	–	–	–	2.035	0.581	1.519	4.004	–
Schöller et al. (fine)	–	–	–	–	1.692	0.442	0.939	3.695	–
	Mean	d_x	d_y	d_z	Mean	ϕ	θ	ψ	
Ours: parent checkpoint	1.044	1.642	0.201	1.290	0.4424	0.0109	1.2083	0.1082	4.577
Ours: validation-selected checkpoint	**1.005**	**1.586**	**0.199**	**1.230**	**0.4252**	0.0109	**1.1564**	0.1082	**4.294**