RELITRACK: CROSS-SENSOR 3D TRACKING VIA OBSERVATION–TRANSITION FACTORIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Autonomous-driving fleets do not keep one sensing configuration: LiDAR layout, return policy, sampling density, coordinate processing, and annotation cadence evolve as platforms and data pipelines change. Recollecting tracking archives is expensive, yet a model trained on one configuration may bind appearance, motion scale, and confidence to it, making reuse on a new platform difficult. We study this source-only transfer problem for LiDAR-based 3D single-object tracking, where each localization error changes the next crop and can accumulate over time. ReLiTrack separates a history-only, continuous-time transition prior from acquisition-sensitive point observations. It encodes sparse crops with masks rather than repeated points, matches the current crop to a permanent initialization anchor and recent valid memories, and predicts an appearance innovation with uncertainty. Paired range–density degradation discourages configuration-specific evidence shortcuts, while covariance intersection fuses motion and appearance without assuming independent errors. A frozen protocol evaluates all six directed transfers among KITTI, Waymo, and nuScenes, applying each source checkpoint unchanged to the other two datasets. ReLiTrack achieves mean Success/Precision of 53.72/65.13 for Vehicle/Car and 44.42/62.93 for Pedestrian, exceeding the strongest baseline means by 5.48/6.18 and 4.48/5.20, respectively.

1 INTRODUCTION

Autonomous-driving fleets evolve continuously: LiDAR layouts are upgraded, return policies and calibration pipelines change, and datasets are annotated at different temporal rates. Recollecting and relabelling every tracking sequence after each change is expensive. A practical tracker should reuse supervision from an existing platform when deployed with a new sensing configuration. This requirement is especially important for LiDAR-based three-dimensional single-object tracking (3D SOT), which estimates the location and orientation of a designated target throughout a point-cloud sequence given only its first-frame bounding box. Each prediction determines the next search region, so a transfer error at one frame can alter every observation that follows.

Most 3D SOT methods localize the target by comparing its current point-cloud appearance with an initialized template or a recent prediction. Shape completion, point-to-box voting, box-aware features, and attention-based template–search interaction have progressively strengthened this paradigm (Giancola et al., 2019; Qi et al., 2020; Zheng et al., 2021; Zhou et al., 2022; Hui et al., 2022; Xu et al., 2023a). Its effectiveness depends on the target template and search crop exposing comparable geometric evidence. LiDAR observations provide sparse samples of visible surfaces instead of dense texture: range, viewpoint, occlusion, and sensor layout determine which parts of an object are observed. A sensor change can therefore transform the same target from a well-supported shape into a handful of returns. Once an appearance mismatch shifts the estimated box, the following crop and stored memory also move, turning a local error into recursive drift.

Motion-centric and memory-based trackers supply a complementary source of information. M2-Track estimates inter-frame target motion, P2P-Voxel models part-level displacement, and MBP-Track aggregates historical appearance with box priors (Zheng et al., 2022; Nie et al., 2025; Xu et al., 2023b). These methods show that tracking benefits from temporal continuity in addition to appearance matching. Cross-sensor reuse, however, changes two inputs at once. The observation changes because the point support and visible surface change; the transition changes because a

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.

frame step can represent a different elapsed time and a different displacement. A motion module trained around one annotation cadence can internalize its step size, while a point-conditioned update can entangle displacement with source-specific density. Memory then carries either error into later frames.

The three datasets studied here make this problem concrete. KITTI, Waymo, and nuScenes differ in sensor coverage, target support, object morphology, and annotated frame interval. Figure 1 measures these differences directly. Whole-frame point count is a poor proxy for target evidence: a dense scan can still contain very few returns inside a distant box, and the zero-return rate rises sharply with range. The sensor-relative displacement presented at each annotated step also shifts across datasets because sampling cadence, target motion, and ego motion jointly determine the change between observations. These measurements identify two distinct transfer pressures—observation reliability and transition scale—that act at different points in the recursive update.

This leads to the central question of our work: *which information should remain stable when the sensing configuration changes?* Target dynamics are expressed by the history of estimated poses and their physical timestamps. Observation quality is expressed by the real support in the current crop. Treating them as separate interfaces allows the temporal model to extrapolate across frame rates and allows the observation model to decide how strongly the current cloud should correct that prediction. It also provides an explicit motion-only update when the cloud supplies little or no usable target evidence.

Our source-only setting makes this distinction consequential. Target scans, pseudo-labels, and target-tuned model selection are absent during training; the same source checkpoint carries its dynamics and confidence estimates into both unseen targets. Transfer must therefore come from the recursive update structure itself.

We instantiate this principle in ReLiTrack, an observation–transition factorized tracker. A continuous-time transition branch consumes past poses, timestamps, validity, and initialization morphology. Four dynamics experts represent static, velocity, acceleration, and turning regimes; moment matching produces a single motion prior whose uncertainty grows when the experts disagree. An observation branch then crops around this prior and estimates an appearance correction. Masked point encoding preserves the number of physical returns instead of manufacturing density by repeating points. A permanent initialization anchor maintains target identity, while recent valid memories capture changes in visible surface. Their frame-specific correspondences are aggregated only after anchor and recent evidence have been matched separately.

Two additional mechanisms make this factorization effective under acquisition shift. First, paired range–density degradation creates full and thinned views of each source example. Pose, feature, and variance-ordering objectives train the observation branch to preserve correspondence while lowering confidence as point support decreases. Second, covariance intersection fuses the motion and appearance estimates because both inherit error from the same track history. An availability gate retains the motion prior for empty or extremely sparse crops. The resulting recursion has a clear division of labor: the transition carries state through time, the observation corrects it with current geometry, and the fusion stage controls how evidence enters the next history.

We evaluate source-only transfer in all six directed pairs among KITTI, Waymo, and nuScenes for Vehicle/Car and Pedestrian. One checkpoint is trained on each source and applied unchanged to the other two targets. ReLiTrack achieves the highest six-direction mean Success and Precision in both categories. Relative to the strongest baseline mean, it improves Vehicle/Car by 5.48/6.18 Success/Precision points and Pedestrian by 4.48/5.20 points. It also leads the target-conditioned averages on Waymo and nuScenes. A representative component study in the main paper and focused density, transition, memory, and fusion studies in the supplement test the corresponding design choices.

Our main contributions are:

- We formulate cross-sensor 3D SOT as coupled shifts in observation reliability and transition scale, and quantify both across KITTI, Waymo, and nuScenes.
- We introduce an observation–transition factorized tracker with continuous-time dynamics, masked anchor–recent correction, paired degradation, and correlation-aware fusion.
- We establish source-only performance through six directed transfers in two categories, with a representative component ablation and focused supplementary studies.

2 Related Work

Appearance matching and geometric correspondence. Early 3D SOT methods complete a target shape or vote from template–search features (Giancola et al., 2019; Qi et al., 2020; Zheng et al., 2021); later designs use transformers, contextual cues, object-preserving representations, and point-to-box or spatio-temporal augmentation to strengthen sparse correspondence (Hui et al., 2022; Zhou et al., 2022; Xu et al., 2023a; Zhao et al., 2024; 2025b;a). Recent work further combines multiple fields and tasks, distills matching from 2D foundation trackers, or compresses point tokens (Liu et al., 2024; Wu et al., 2024a; Zhou et al., 2026). These trackers improve the observation model, whereas ReLiTrack additionally isolates its point-conditioned correction from a history-only transition and retains the physical support mask.

Motion and temporal memory. Motion-centric trackers regress object displacement from consecutive frames or corresponding parts (Zheng et al., 2022; Nie et al., 2025). Memory models instead aggregate earlier appearance and box evidence: MBPTrack couples memory with box priors, StreamTrack treats a tracklet as a continuous feature stream, and HVTrack expands memory and search context for large temporal gaps (Xu et al., 2023b; Luo et al., 2024; Wu et al., 2024b). ChronoTrack extends this direction with compact, temporally consistent long-term tokens (Yoo et al., 2026). In contrast, our transition consumes only poses, timestamps, validity, and initialization morphology; current point features enter through a separate innovation, preventing sensor-specific appearance from defining the motion prior.

Generalization across categories, conditions, and sensors. Unified trackers address category variation through adaptive shape priors, pretrained 3D representations, relation parsing, or structure-aware keypoints (Nie et al., 2024; Wang et al., 2025; Liu et al., 2026; Xiao et al., 2026); GSOT3D broadens evaluation to in-the-wild categories, while UAWTrack studies weather-corrupted observations (Jiao et al., 2025; Yang et al., 2025). Related cross-dataset 3D detection methods use feature alignment, point generation, pseudo-labels, or stationary-object aggregation (Zhang et al., 2021; Yang et al., 2021; Xu et al., 2021; Hu et al., 2023; Huang et al., 2024), and recent source-only studies expose architecture and density as central LiDAR generalization factors (Eskandar, 2024; Kim et al., 2024). Our protocol permits neither target scans nor target statistics during training and evaluates the recursive interaction between cadence-dependent transition and sensor-dependent evidence over all six directed dataset pairs.

3 Sensing and Temporal Distribution Shifts

Changing LiDAR configuration alters both the number and geometry of returns, while annotation cadence changes the displacement between successive tracking steps. Figure 1 measures these effects in the Vehicle/Car inputs. Whole-frame density differs from the evidence available inside a target box (a–b). Target support decreases with range and can vanish (c–d), while sensor-relative displacement per annotated step and sampled box dimensions differ across datasets (e–f). The displacement statistic captures the tracker-facing combination of target and ego motion.

These shifts affect separate stages of recursive tracking. Sparse target support weakens correspondence; larger step displacements challenge a fixed-step motion model; and an incorrect pose changes the next crop. The measurements motivate preserving valid point support and conditioning prediction on elapsed time. Dataset and preprocessing details are in Supplementary Sections S1–S2.

The target-conditioned plots also distinguish two forms of missing evidence. A low count permits a noisy appearance estimate, whereas a zero-return box provides no target geometry for a point matcher. Both cases occur even when the whole frame contains many returns. The displacement plot captures the motion presented between annotated observations, including ego motion and sampling cadence. Effective transfer therefore requires crop-level evidence modeling and elapsed-time conditioning rather than a fixed interpretation of scan density or frame index.

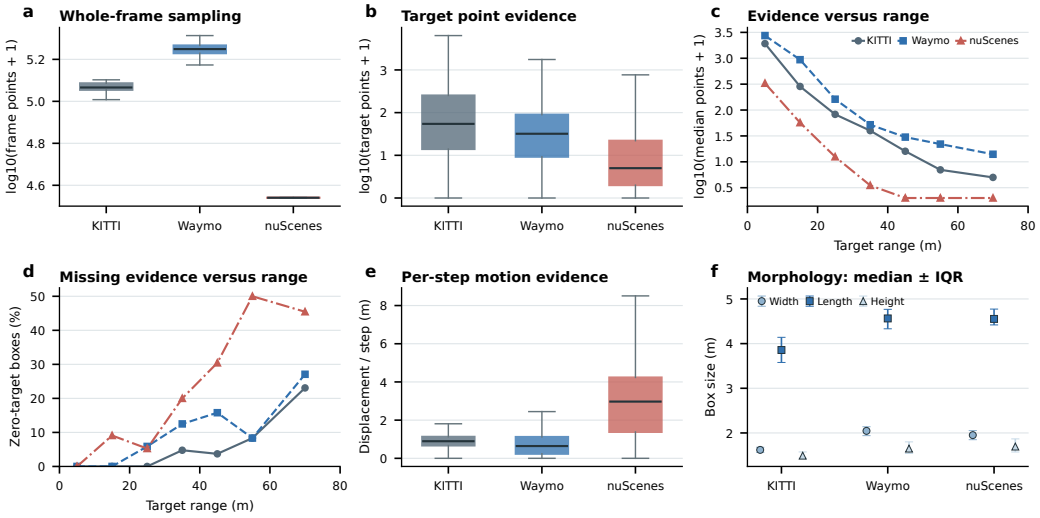

Figure 1: Measured Vehicle/Car acquisition shifts. (a) Whole-frame point count; (b) target-box point support; (c–d) target support and empty-box rate versus range; (e) sensor-relative displacement per annotated step; and (f) sampled box dimensions. Boxes show medians and interquartile ranges. Sampling counts and preprocessing are given in Supplementary Section S2.

4 METHOD

Given an initial target box, ReLiTrack recursively predicts one box per LiDAR frame. Its update has two parts (Figure 2): a transition from past states to a motion prior, followed by an appearance correction from the current crop. The transition uses past poses and timestamps, while the correction uses the point support in the current crop.

4.1 STATE AND OBSERVATION

Let $\mathbf{g}_t = (x_t, y_t, z_t, \theta_t) \in \mathrm{SE}(2) \times \mathbb{R}$ denote planar pose with vertical translation, and let $\mathbf{s}_1$ be the initial box size, held fixed during tracking. The first ground-truth box initializes $\mathbf{g}_1$ and the permanent appearance anchor. At time τ_t, the model has access to past posterior states, diagonal uncertainty approximations, and validity flags,

$$\mathcal{H}_{t-1} = \{(\hat{\mathbf{g}}_j^+, \tau_j, \mathbf{P}_j^+, v_j) : j < t\}. \tag{1}$$

Horizontal translation and yaw compose on $\mathrm{SE}(2)$; height is additive. We express residuals in a four-dimensional local tangent space, normalizing horizontal displacement by $\sqrt{l_1 w_1}$ and vertical displacement by h_1. Yaw is wrapped and represented in radians. This convention is used for the transition, appearance residual, and fusion.

A constant-velocity extrapolation from valid past states provides a crop reference. The learned prior subsequently centers a search region whose extent increases with predicted uncertainty. Historical crops use their associated predicted boxes; the initialization crop remains an anchor alongside recent valid crops. At most M real points are sampled from each crop. If fewer are present, unused slots are masked rather than filled by repeating measurements. A shared point Transformer encodes point descriptors, acquisition evidence, and first-box morphology:

$$\mathbf{F}_t = \mathrm{Tr}_{\mathrm{point}}\big(\phi_p(\mathbf{q}_t) + \phi_e(\mathbf{e}_t) + \phi_g(\mathbf{g}_1^{\mathrm{morph}}); \mathbf{m}_t\big). \tag{2}$$

Here $\mathbf{m}_t$ excludes padding from keys and values. The evidence descriptor records support and range; the morphology descriptor is fixed at initialization. Their definitions and tensor sizes are in Supplementary Section S4.

Point descriptors and evidence descriptors serve complementary roles. Point coordinates are normalized by the initial dimensions so that matching operates in a target-relative frame. The evidence vector records the number of real returns and target range for the correction and uncertainty heads,

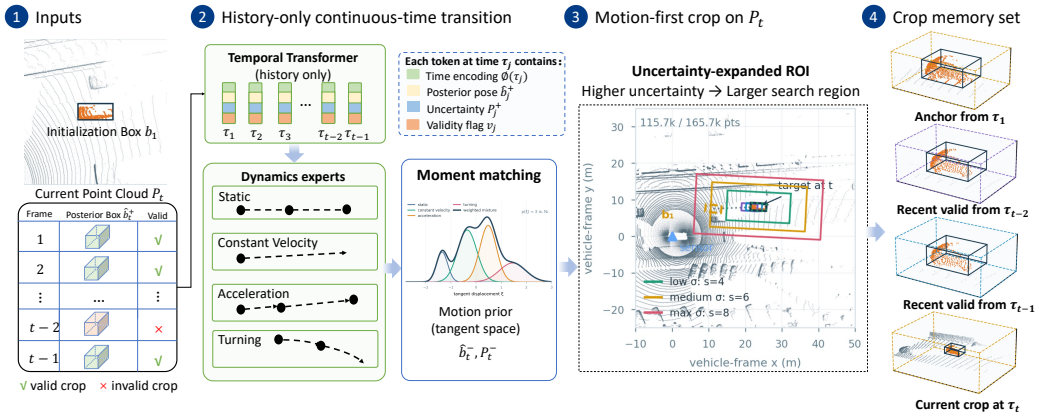

Figure 2: Recursive inference in ReLiTrack. Past states and timestamps define a motion prior and search crop; masked current points and stored crops produce an appearance correction; covariance intersection yields the next state.

while masked slots remain excluded from measurement evidence. The first-box morphology supplies a persistent scale reference. Each sparse return is encoded once, preserving the distinction between two physical returns and a dense observation. An empty crop carries one masked dummy point and routes the update through the motion prior.

4.2 Elapsed-time motion prior

The transition Transformer receives relative past poses, their timestamps and validity, anchor identity, and initialization morphology. It predicts probabilities $\pi_{k,t}$ over four dynamics experts: static, constant velocity, acceleration, and turning. Each expert outputs a four-dimensional tangent velocity $\mathbf{v}_{k,t}$, acceleration $\mathbf{a}_{k,t}$, and diagonal process variance $\mathbf{Q}_{k,t}$. For the observed interval $\Delta\tau_t = \tau_t - \tau_{t-1}$,

$$\boldsymbol{\xi}_{k,t} = \mathbf{v}_{k,t}\Delta\tau_t + \tfrac{1}{2}\mathbf{a}_{k,t}\Delta\tau_t^2, \qquad \sum_k \pi_{k,t} = 1. \tag{3}$$

The static expert has zero displacement, and the constant-velocity expert has zero acceleration. Moment matching yields one displacement and an uncertainty that includes disagreement among experts:

$$\bar{\boldsymbol{\xi}}_t = \sum_k \pi_{k,t}\boldsymbol{\xi}_{k,t}, \tag{4}$$

$$\mathbf{Q}_t = \sum_k \pi_{k,t}\left[\mathbf{Q}_{k,t} + \mathrm{diag}\big((\boldsymbol{\xi}_{k,t} - \bar{\boldsymbol{\xi}}_t)^{\odot 2}\big)\right]. \tag{5}$$

Yaw differences are wrapped before the second moment is computed. The predicted pose is $\hat{\mathbf{g}}_t^- = \hat{\mathbf{g}}_{t-1}^+ \circ \mathrm{Exp}(\bar{\boldsymbol{\xi}}_t)$. We use $\mathbf{P}_t^- \approx \mathbf{P}_{t-1}^+ + \mathbf{Q}_t$ as a diagonal propagation approximation in the normalized local tangent space. The explicit dependence on $\Delta\tau_t$ adapts the predicted displacement to the interval between observations.

The mixture represents uncertainty over displacement regimes before the current cloud is encoded and passes a single moment-matched prior to the observation branch. The second term inside $\mathbf{Q}_t$ increases variance when experts disagree even if each expert reports a small process variance. A model trained on short steps can represent a longer query interval through the same velocity and acceleration outputs, while disagreement among regimes enlarges the region searched by the observation branch. The current crop then supplies the appearance correction when point support is available.

4.3 Appearance correction from anchor and recent memory

For each valid memory, current tokens attend to historical tokens. Matched features, their differences and products, and matched coordinate displacements form a change representation. A change

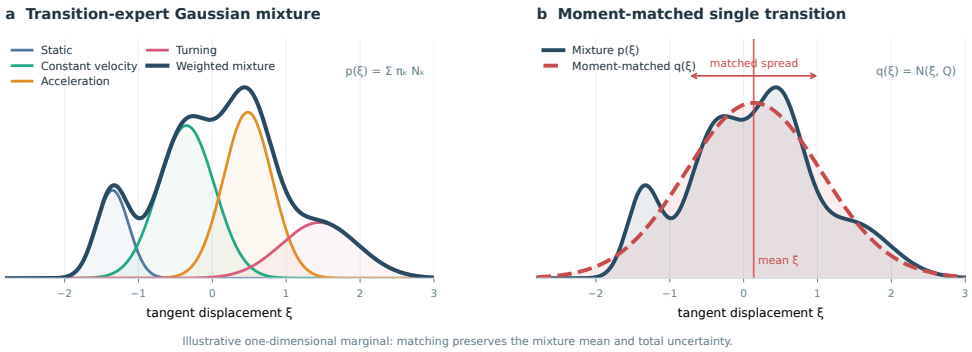

Figure 3: Illustrative one-dimensional marginal of the motion-expert mixture. Moment matching preserves the mixture mean and total variance, including spread between experts, while passing a single prior to the observation branch.

Transformer pools this information separately for the initialization anchor and each recent memory; a temporal Transformer then aggregates their summaries using age, match entropy, and point evidence. Thus the anchor can preserve the initial target reference while recent memories capture newer visible surfaces (Figure 4).

For a current token i and a historical frame h, cross attention returns a matched feature $\bar{\mathbf{f}}_{hi}$ and coordinate $\bar{\mathbf{x}}_{hi}$. The local descriptor includes $\mathbf{f}_i^t - \bar{\mathbf{f}}_{hi}$, $\mathbf{f}_i^t \odot \bar{\mathbf{f}}_{hi}$, and $\mathbf{x}_i^t - \bar{\mathbf{x}}_{hi}$ together with their input features and coordinates. Feature disagreement and geometric displacement can therefore influence the same pooled correspondence. Relevance pooling uses shared weights for learned and geometric terms; otherwise a strong match in one part of the crop could be paired with displacement from unrelated points. The anchor and recent summaries are aggregated only after this frame-specific matching, allowing the temporal model to weight identity evidence and recent appearance differently.

The appearance head predicts a correction $\boldsymbol{\delta}_t^a$ relative to the motion-centered crop and a diagonal variance $\mathbf{R}_t$. Its absolute pose estimate is $\hat{\mathbf{g}}_t^a = \hat{\mathbf{g}}_t^- \circ \mathrm{Exp}(\boldsymbol{\delta}_t^a)$. Sparse support and range increase predicted appearance variance. A dense crop can still contain a distractor, so point count is used as reliability evidence rather than an identity decision. Detailed change descriptors appear in Supplementary Section S4.

4.4 Fusion under unknown correlation

The prior and appearance estimate share historical tracking errors. Before fusion, both diagonal variances are expressed in the normalized tangent at $\hat{\mathbf{g}}_t^-$. With fixed weight $\omega = 0.5$, we use a diagonal covariance-intersection update (Julier & Uhlmann, 1997):

$$
\begin{aligned}
(\mathbf{P}_t^+)^{-1} &= \omega(\mathbf{P}_t^-)^{-1} + (1-\omega)\mathbf{R}_t^{-1}, \\
\mathbf{K}_t &= \mathbf{P}_t^+(1-\omega)\mathbf{R}_t^{-1}, \qquad \mathbf{r}_t = \mathrm{Log}\big((\hat{\mathbf{g}}_t^-)^{-1} \circ \hat{\mathbf{g}}_t^a\big), \\
\hat{\mathbf{g}}_t^+ &= \hat{\mathbf{g}}_t^- \circ \mathrm{Exp}(\mathbf{K}_t\mathbf{r}_t).
\end{aligned}
\tag{6}
$$

This weighting combines the two correlated estimates without estimating their cross-covariance. When the current crop contains fewer than three real points or all appearance memories are invalid, we set $(\hat{\mathbf{g}}_t^+, \mathbf{P}_t^+) = (\hat{\mathbf{g}}_t^-, \mathbf{P}_t^-)$ exactly. Otherwise, the posterior and accepted crop enter the next history. The fallback prevents padding from becoming a confident observation.

For one diagonal component with prior variance p^- and appearance variance r, the update has the equivalent form

$$
p^+ = \frac{p^- r}{\omega r + (1-\omega)p^-}, \qquad k = \frac{(1-\omega)p^-}{\omega r + (1-\omega)p^-}. \tag{7}
$$

As appearance variance increases relative to prior variance, its contribution to the pose correction decreases. When both variances are equal, CI retains their common variance instead of counting the two correlated estimates as independent evidence. The fixed ω is shared across transfer directions, and all computations use the same normalized local tangent representation.

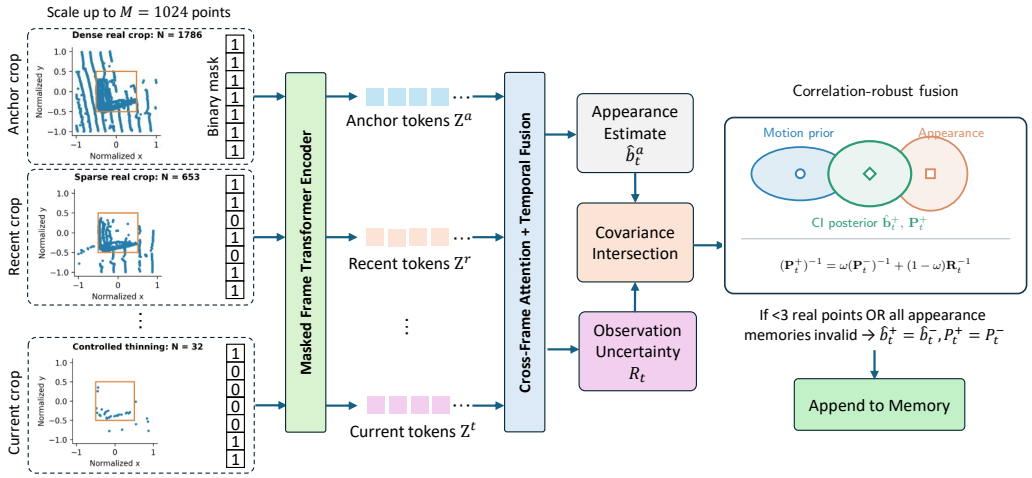

Figure 4: Appearance correction and state fusion. Anchor, recent, and current crops are encoded with validity masks. The covariance-intersection (CI) inset is schematic; with zero valid appearance evidence, the motion prior is retained.

4.5 Source-only paired degradation

To expose the observation branch to varying point support, training constructs full and range-dependent thinned views of the same source window. The full view supplies stop-gradient targets, and the degraded view recomputes masks, counts, and evidence. Both views share the target pose and timestamps. Let f, d indicate full and degraded views, $\bar{\mathbf{z}}$ the pooled change feature, d_{pose} a discrepancy in the normalized pose tangent, and sg stop-gradient. The paired objective is

$$\mathcal{L}_{\text{pair}} = \lambda_p d_{\text{pose}}\left(\hat{\mathbf{g}}_t^{a,d}, \text{sg}(\hat{\mathbf{g}}_t^{a,f})\right) + \lambda_f \left\|\bar{\mathbf{z}}_t^d - \text{sg}(\bar{\mathbf{z}}_t^f)\right\|_1 + \frac{\lambda_u}{4}\sum_{j=1}^{4}\left[\log R_{t,j}^f + m - \log R_{t,j}^d\right]_+ . \tag{8}$$

The last term penalizes degraded views that report lower variance than their full counterparts, component by component. Supervised source losses additionally train posterior, transition, and appearance poses, velocity and acceleration, and dynamics-mode classification. The full view supplies stop-gradient targets from the same network. Training variants and implementation details are in Supplementary Section S4.

The three paired terms constrain complementary aspects of transfer. Pose agreement preserves localization under point thinning; feature agreement preserves the change representation used by memory matching; and variance ordering lowers confidence when fewer returns remain. Because both views share the same target state, these terms combine directly with source supervision. At inference, the ordinary recursive tracker runs once per frame using the source-trained model.

5 Experiments

5.1 Evaluation protocol

We train each method on a source dataset and evaluate the resulting checkpoint on the other two datasets, without target-specific retraining. We report all six directed transfers among KITTI (K), Waymo (W), and nuScenes (N) for Vehicle/Car and Pedestrian. Tracks are initialized with the first ground-truth box; later states and crops depend on past predictions. We use the stated CD3DSOT-v0.1 splits and category mappings (Supplementary Sections S1 and S3). Success is the area under the oriented 3D IoU threshold curve; Precision is the area under the 3D center-error threshold curve. Scores are percentages, and the six-direction mean weights each transfer equally. The compared methods are M2-Track (Zheng et al., 2022), MBPTrack (Xu et al., 2023b), and P2P-Voxel (Nie et al., 2025).

5.2 CROSS-DATASET COMPARISON

Table 1(a) gives the complete transfer matrices. ReLiTrack achieves the highest mean Success/Precision for Vehicle/Car (53.72/65.13) and Pedestrian (44.42/62.93). Relative to MBPTrack, the strongest baseline by mean in both categories, the gains are 5.48/6.18 and 4.48/5.20 percentage points. ReLiTrack leads both metrics in five of six vehicle directions and leads four Success and three Precision directions for pedestrians. The vehicle gains cover both incoming sources to Waymo and nuScenes; for pedestrians, the strongest two-source consistency appears on Waymo.

Table 1: Cross-dataset transfer under CD3DSOT-v0.1. (a) Six directed transfers and their mean; (b) target-conditioned averages over two incoming sources. Entries are Success/Precision (%); best values are bold.

(a) Six-direction transfer

Method	K→W	K→N	W→K	W→N	N→K	N→W	Mean
Vehicle/Car							
M2-Track	50.1/62.3	23.2/40.6	68.8/79.4	42.5/43.4	51.2/59.8	46.2/55.8	47.00/56.88
MBPTrack	55.8/65.3	21.5/38.8	70.4/85.5	37.6/40.1	53.3/60.8	50.8/63.2	48.23/58.95
P2P-Voxel	35.6/47.7	20.6/35.4	**73.5/90.1**	26.6/33.7	56.7/62.2	40.8/49.2	42.30/53.05
ReLiTrack	**60.1/68.8**	**29.5/47.6**	71.2/84.3	**50.2/61.3**	**58.5/62.3**	**52.8/66.5**	**53.72/65.13**
Pedestrian							
M2-Track	32.0/49.7	**25.6/40.5**	63.2/90.9	30.8/45.3	40.5/60.7	38.9/52.8	38.50/56.65
MBPTrack	33.7/52.7	22.0/38.8	66.8/91.3	35.8/51.2	36.4/52.3	44.9/60.1	39.93/57.73
P2P-Voxel	27.8/43.2	16.8/21.2	**69.8/92.5**	31.3/44.0	45.5/**67.1**	43.8/60.0	39.17/54.67
ReLiTrack	**40.8/59.1**	24.5/39.2	64.5/90.6	**40.7/57.5**	**46.1**/66.9	**49.9/64.3**	**44.42/62.93**

(b) Target-conditioned averages

Method	Target KITTI	Target Waymo	Target nuScenes
Vehicle/Car			
M2-Track	60.00/69.60	48.15/59.05	32.85/42.00
MBPTrack	61.85/73.15	53.30/64.25	29.55/39.45
P2P-Voxel	65.10/76.15	38.20/48.45	23.60/34.55
ReLiTrack	64.85/73.30	56.45/67.65	39.85/54.45
Pedestrian			
M2-Track	51.85/75.80	35.45/51.25	28.20/42.90
MBPTrack	51.60/71.80	39.30/56.40	28.90/45.00
P2P-Voxel	57.65/79.80	35.80/51.60	24.05/32.60
ReLiTrack	55.30/78.75	45.35/61.70	32.60/48.35

5.3 INTERPRETING THE TRANSFER PATTERN

The target dataset changes the ranking. P2P-Voxel has the highest KITTI-target mean in both categories, whereas ReLiTrack leads the Waymo- and nuScenes-target means for both metrics (Table 1(b)). Figure 1 connects this reversal to substantial changes in target support and inter-frame displacement. Success and Precision move together across these targets, showing that the transfer gap affects both oriented overlap and center localization.

Source choice matters as well. With nuScenes as the Vehicle/Car target, ReLiTrack obtains 50.2/61.3 from Waymo training but 29.5/47.6 from KITTI training. This same-target contrast points to source coverage and acquisition compatibility in addition to target difficulty. The effect is not simply monotonic in source size: M2-Track scores 50.1/62.3 on K→W and 46.2/55.8 on N→W. For pedestrians, the lower spatial support and changing body shape can further complicate correspondence; their direction-dependent results are less uniform than the vehicle results.

P2P-Voxel illustrates the importance of updating orientation across sensor regimes. Its W→K scores lead both categories, while W→N is substantially lower. The evaluated reproduction exports a fixed target yaw. Larger heading increments lower oriented overlap and rotate the following search crop away from the target, which also affects center localization through recursive feedback. The longer annotated interval in nuScenes makes this update especially important. Supplementary Section S5 analyzes this rotation behavior in detail.

Vehicle/Car transfer maps KITTI and nuScenes Car to Waymo's broader Vehicle label, so target shape and subtype composition change along with sensing. Pedestrians provide fewer spatially distinct surface points and deform across frames, producing a more demanding correspondence problem. ReLiTrack nevertheless raises the six-direction pedestrian mean by 4.48/5.20 points over the strongest baseline.

Table 1(b) separates target effects from the six-direction mean. The KITTI-target ranking favors P2P-Voxel, including Vehicle/Car Precision, where its 76.15 exceeds ReLiTrack's 73.30. On Waymo and nuScenes, ReLiTrack leads the target means for both categories and metrics. For Vehicle/Car, both incoming transfers to Waymo and both incoming transfers to nuScenes exceed the strongest baseline in Success and Precision. The same two-source consistency appears for Pedestrian on Waymo.

5.4 Ablation studies

Table 2 changes one observation, transition, memory, or fusion component at a time on Vehicle/Car W→N and W→K. All variants use identical training budgets, splits, initialization, and evaluation; focused sweeps appear in Supplementary Section S5.5.

Table 2: Representative component ablation on Vehicle/Car. Mean Δ gives the Success/Precision change from the full model, averaged over the two transfer directions.

Configuration	W→N	W→K	Mean Δ (S/P)
Full ReLiTrack	50.2/61.3	71.2/84.3	Ref.
Repeat points instead of masking	47.2/56.3	69.9/82.6	$-2.2/-3.4$
Without paired degradation	48.9/60.1	68.8/80.2	$-1.9/-2.7$
Fixed-step constant-velocity transition	40.5/53.3	55.8/70.2	$-12.6/-11.1$
Without permanent anchor	46.9/55.0	68.9/81.7	$-2.8/-4.5$
Fixed-average fusion	42.3/53.7	63.9/76.4	$-7.6/-7.8$

All five ablations reduce both metrics in both directions. Replacing the full transition with fixed-step constant velocity causes the largest mean loss, $12.6/11.1$ Success/Precision points. Its loss reaches $15.4/14.1$ points on W→K, compared with $9.7/8.0$ on W→N. Fixed-average fusion ranks second at $7.6/7.8$ points and is similarly harmful in both directions: $7.9/7.6$ on W→N and $7.3/7.9$ on W→K. The transition and fusion designs therefore account for the two largest changes in this comparison.

The observation and memory changes are smaller but consistent. Repeating points instead of masking lowers the mean by $2.2/3.4$ points, removing paired degradation by $1.9/2.7$, and removing the permanent anchor by $2.8/4.5$. Precision falls more than Success for all three, most clearly without the anchor on W→N (6.3 Precision points). This pattern links physical point support, source-side degradation, and anchored identity to center localization and the quality of the next recursive crop. Supplementary Section S5.5 places each comparison within its component family.

6 Conclusion

We introduced ReLiTrack for source-only cross-sensor 3D single-object tracking, treating acquisition change as coupled shifts in observation reliability and transition scale. The proposed factorization combines a history-only continuous-time motion prior with a support-aware appearance innovation, anchored memory, source-side range–density degradation, and correlation-aware fusion. Across all six directed transfers among KITTI, Waymo, and nuScenes, one unchanged source checkpoint attains the best mean Success and Precision for both Vehicle/Car and Pedestrian, improving on the strongest baseline means by 5.48/6.18 and 4.48/5.20 points, respectively. Target-conditioned results also show why a transfer matrix is more informative than one source–target pair: P2P-Voxel remains strongest on KITTI-target averages, whereas ReLiTrack leads on Waymo and nuScenes. The ablations identify continuous-time transition and covariance-intersection fusion as the largest contributors, while point masking, paired degradation, and the permanent anchor provide consistent gains. These results support a practical conclusion: preserving physical time and actual point support, then controlling how uncertain observations revise motion, makes recursive tracking more robust and its cross-sensor failures easier to diagnose.

REFERENCES

George Eskandar. An empirical study of the generalization ability of LiDAR 3D object detectors to unseen domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23815–23825, 2024.

Silvio Giancola, Jesus Zarzar, and Bernard Ghanem. Leveraging shape completion for 3D siamese tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1359–1368, 2019.

Qianjiang Hu, Daizong Liu, and Wei Hu. Density-insensitive unsupervised domain adaption on 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17556–17566, 2023.

Chengjie Huang, Vahdat Abdelzad, Sean Sedwards, and Krzysztof Czarnecki. SOAP: Cross-sensor domain adaptation for 3d object detection using stationary object aggregation pseudo-labelling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3352–3361, 2024.

Le Hui, Lingpeng Wang, Linghua Tang, Kaihao Lan, Jin Xie, and Jian Yang. 3d siamese transformer network for single object tracking on point clouds. In *Computer Vision – ECCV 2022*, pp. 293–310, 2022.

Yifan Jiao, Yunhao Li, Junhua Ding, Qing Yang, Song Fu, Heng Fan, and Libo Zhang. GSOT3D: Towards generic 3d single object tracking in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5469–5478, 2025.

Simon J. Julier and Jeffrey K. Uhlmann. A non-divergent estimation algorithm in the presence of unknown correlations. In *Proceedings of the American Control Conference*, volume 4, pp. 2369–2373, 1997. doi: 10.1109/ACC.1997.609105.

Jaeyeul Kim, Jungwan Woo, Jeonghoon Kim, and Sunghoon Im. Rethinking LiDAR domain generalization: Single source as multiple density domains. In *Computer Vision – ECCV 2024*, pp. 310–327, 2024.

Jiaming Liu, Yue Wu, Maoguo Gong, Qiguang Miao, Wenping Ma, Cai Xu, and Can Qin. M3SOT: Multi-frame, multi-field, multi-space 3D single object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3630–3638, 2024. doi: 10.1609/aaai.v38i4.28152.

Juntao Liu, Zikun Zhou, Zhuotao Tian, Guangming Lu, Jun Yu, and Wenjie Pei. PointRePar: Spatiotemporal point relation parsing for robust category-unified 3d tracking. In *International Conference on Learning Representations*, 2026.

Zhipeng Luo, Gongjie Zhang, Changqing Zhou, Zhonghua Wu, Qingyi Tao, Lewei Lu, and Shijian Lu. Modeling continuous motion for 3D point cloud object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4026–4034, 2024. doi: 10.1609/aaai.v38i5.28196.

Jiahao Nie, Zhiwei He, Xudong Lv, Xueyi Zhou, Dong-Kyu Chae, and Fei Xie. Towards category unification of 3d single object tracking on point clouds. In *International Conference on Learning Representations*, 2024.

Jiahao Nie, Fei Xie, Sifan Zhou, Xueyi Zhou, Dong-Kyu Chae, and Zhiwei He. P2P: Part-to-part motion cues guide a strong tracking framework for LiDAR point clouds. *International Journal of Computer Vision*, 133 (8):5326–5342, 2025. doi: 10.1007/s11263-025-02430-6.

Haozhe Qi, Chen Feng, Zhiguo Cao, Feng Zhao, and Yang Xiao. P2B: Point-to-box network for 3d object tracking in point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6329–6338, 2020.

Mengmeng Wang, Haonan Wang, Yulong Li, Xiangjie Kong, Jiaxin Du, Guojiang Shen, and Feng Xia. TrackAny3D: Transferring pretrained 3d models for category-unified 3d point cloud tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 28249–28259, 2025.

Qiangqiang Wu, Yan Xia, Jia Wan, and Antoni B. Chan. Boosting 3D single object tracking with 2D matching distillation and 3D pre-training. In *Computer Vision – ECCV 2024*, pp. 270–288, 2024a. doi: 10.1007/978-3-031-73254-6_16.

Qiao Wu, Kun Sun, Pei An, Mathieu Salzmann, Yanning Zhang, and Jiaqi Yang. 3D single-object tracking in point clouds with high temporal variation. In *Computer Vision – ECCV 2024*, pp. 279–296, 2024b.

Jie Xiao, Yinchao Ma, Yuyang Tang, Dengqing Yang, Jianpeng Yang, Xu Zhou, Qiao Li, Wenfei Yang, and Tianzhu Zhang. Generalizable structure-aware keypoint correspondence for category-unified 3d single object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28156–28166, 2026.

Qiangeng Xu, Yin Zhou, Weiyue Wang, Charles R. Qi, and Dragomir Anguelov. SPG: Unsupervised domain adaptation for 3d object detection via semantic point generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15446–15456, 2021.

Tian-Xing Xu, Yuan-Chen Guo, Yu-Kun Lai, and Song-Hai Zhang. CXTrack: Improving 3D point cloud tracking with contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1084–1093, 2023a.

Tian-Xing Xu, Yuan-Chen Guo, Yu-Kun Lai, and Song-Hai Zhang. MBPTrack: Improving 3d point cloud tracking with memory networks and box priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9911–9920, 2023b.

Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. ST3D: Self-training for unsupervised domain adaptation on 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10368–10378, 2021.

Yuxiang Yang, Hongjie Gu, Yingqi Deng, Zhekang Dong, Zhiwei He, and Jing Zhang. UAWTrack: Universal 3d single object tracking in adverse weather. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9336–9344, 2025. doi: 10.1609/aaai.v39i9.33011.

Jaejoon Yoo, SuBeen Lee, Yerim Jeon, Miso Lee, and Jae-Pil Heo. Temporally consistent long-term memory for 3D single object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 8388–8397, 2026.

Weichen Zhang, Wen Li, and Dong Xu. SRDAN: Scale-aware and range-aware domain adaptation network for cross-dataset 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6769–6779, 2021.

Kaijie Zhao, Haitao Zhao, Zhongze Wang, Jingchao Peng, and Zhengwei Hu. Object-preserving siamese network for single-object tracking on point clouds. *IEEE Transactions on Multimedia*, 26:3007–3017, 2024.

Kaijie Zhao, Haitao Zhao, Zhongze Wang, Jingchao Peng, and Lujian Yao. Enhancing point cloud tracking with spatio-temporal triangle optimization. *IEEE Transactions on Instrumentation and Measurement*, 2025a. doi: 10.1109/TIM.2025.3553981.

Kaijie Zhao, Haitao Zhao, Zhongze Wang, Lujian Yao, Jingchao Peng, and Zhengwei Hu. PBMA: Enhancing 3D point cloud tracking with point-to-box motion augmentation. *Expert Systems with Applications*, 281: 127548, 2025b. doi: 10.1016/j.eswa.2025.127548.

Chaoda Zheng, Xu Yan, Jiantao Gao, Weibing Zhao, Wei Zhang, Zhen Li, and Shuguang Cui. Box-aware feature enhancement for single object tracking on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13199–13208, 2021.

Chaoda Zheng, Xu Yan, Haiming Zhang, Baoyuan Wang, Shenghui Cheng, Shuguang Cui, and Zhen Li. Beyond 3d siamese tracking: A motion-centric paradigm for 3d single object tracking in point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8111–8120, 2022.

Changqing Zhou, Zhipeng Luo, Yueru Luo, Tianrui Liu, Liang Pan, Zhongang Cai, Haiyu Zhao, and Shijian Lu. PTTR: Relational 3D point cloud object tracking with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8531–8540, 2022.

Sifan Zhou, Yichao Cao, Jiahao Nie, Yuqian Fu, Ziyu Zhao, Xiaobo Lu, and Shuo Wang. CompTrack: Information bottleneck-guided low-rank dynamic token compression for point cloud tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13773–13781, 2026. doi: 10.1609/aaai.v40i16.38385.