TRACK2RECON: PERSISTENT 3D OBJECT DETECTION AND TRACKING FOR DYNAMIC SCENE RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Instance-level dynamic 3D Gaussian Splatting (3DGS) for autonomous-driving scene reconstruction typically relies on per-frame 3D bounding boxes and persistent track IDs, whereas unlabeled driving logs contain only multi-camera images, LiDAR point clouds, calibration, and ego poses. Recovering persistent 3D objects from such data poses two main challenges: sparse and incomplete LiDAR observations make reliable association between 2D instances and 3D components difficult, while limited viewpoints and occlusion lead to underestimated object sizes and shifted centers. To address these issues, we present Track2Recon. Its hull-supported 2D–3D object association measures object-level spatial consistency using the projected envelope of an entire 3D component and combines it with support from actual LiDAR points to determine object assignment. Its complete-geometry recovery for partial observations estimates the full object extent from complementary observations across frames and uses visible-surface anchoring to complete missing geometry primarily toward unobserved regions, thereby jointly recovering object size and center. Track2Recon requires neither manual 3D annotations in the target domain nor additional supervised training. Experiments on PandaSet and Waymo demonstrate strong performance in 3D detection and persistent tracking, and the recovered 3D bounding boxes and track IDs can be directly used for instance-level dynamic scene reconstruction and editing. Our code is available at `https://anonymous.4open.science/r/iclr-code-7d42/`.

1 INTRODUCTION

High-fidelity, editable reconstruction of dynamic driving scenes typically represents traffic participants such as vehicles and pedestrians as object nodes with independent geometry, canonical spaces, and temporal trajectories. Reconstruction methods such as Street Gaussians and OmniRe rely on per-frame 3D bounding boxes and temporally consistent instance identities to support object removal, duplication, relocation, and trajectory editing (Yan et al., 2024; Chen et al., 2025). However, ordinary driving logs do not directly provide such information, while manual 3D annotation is costly.

Several approaches use 2D foundation models to assist 3D object parsing, including aggregating points inside masks, spatial clustering, and single-frame full 3D box search (Khurana et al., 2024; Tempfli et al., 2025; Chow et al., 2025). For persistent object reconstruction, however, two key challenges remain. **Object assignment:** Because of viewpoint differences between cameras and LiDAR, occlusion, and depth projection, a single mask may contain 3D points from multiple objects, and local projection hits alone are insufficient to determine whether an entire 3D component truly belongs to that instance. **Partial-observation geometry:** A single LiDAR frame often observes only one side of an object, so the visible extent does not reflect its full physical extent. Symmetrically expanding the size around the original center would also move visible surfaces supported by actual observations; therefore, size and center must be recovered jointly.

Track2Recon combines component-level projection envelopes with LiDAR support to resolve object assignment, then estimates complete object extents from multi-frame observations and uses visible-face anchoring to determine completion direction. After recovering the primary geometry,

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.

it repairs identity fragments, conservatively re-estimates geometry from the original point clouds, and recovers missing observations only when current-frame evidence is sufficient. Geometry and identity reinforce each other: more stable geometry improves cross-frame association, while more continuous tracks provide additional observations for geometric recovery. Our main contributions are as follows:

- We recover per-frame 3D boxes and persistent track IDs from unlabeled driving logs without manual 3D annotations or supervised target-domain training, and experiments demonstrate that our method achieves strong performance in detection, tracking, and downstream reconstruction.
- We propose hull-supported 2D–3D object association that determines object assignment using both projection consistency of complete components and support from actual LiDAR points.
- We propose complete-geometry recovery for partial observations that combines multi-frame physical-extent estimation with visible-face anchoring to jointly recover object size and center.

2 Related Work

Object-level dynamic reconstruction. 3DGS provides an efficient explicit representation for driving scenes (Kerbl et al., 2023). Street Gaussians and OmniRe model dynamic instances in independent canonical spaces but rely on external 3D boxes and persistent trajectories; background contamination, identity merging, and trajectory fragmentation can impair cross-frame aggregation (Yan et al., 2024; Chen et al., 2025). Zhang et al. (2025) reduce dependence on 3D trackers through 2D tracking, 3D fusion, and motion learning, while DIAL-GS incorporates instance discovery into reconstruction optimization (Su et al., 2026). In contrast, we recover persistent 3D objects before reconstruction and output explicit boxes and trajectories for existing reconstructors.

Object discovery without manual 3D boxes. CM3D generates pseudo 3D boxes from LiDAR points inside masks using class-specific size priors; VESPA adds spatial clustering and class priors to suppress projection background and compensate incomplete size; OV-SCAN associates 3D clusters before full-box search; SAM3D performs zero-shot BEV detection with SAM (Khurana et al., 2024; Tempfli et al., 2025; Chow et al., 2025; Zhang et al., 2023a). These methods remain vulnerable to projection ambiguity, cluster fragmentation, and single-frame partial observations, complicating object assignment and full-size recovery. In contrast, we determine assignment from complete-component projections and LiDAR support, and recover instance-specific full extents from multi-frame observations. Auto4D uses corner-aligned resizing, while 3DAL and DetZero refine boxes from sequences (Yang et al., 2021; Qi et al., 2021; Ma et al., 2023); MODEST, OYSTER, and LISO learn from unlabeled LiDAR data (You et al., 2022; Zhang et al., 2023b; Baur et al., 2024).

3D tracking and 2D identities. AB3DMOT, SimpleTrack, MCTrack, and CenterPoint/PubTracker associate per-frame 3D detections over time, so identity continuity is sensitive to detection stability and geometric quality (Weng et al., 2020; Pang et al., 2021; Wang et al., 2025; Yin et al., 2021). 2D foundation models provide richer identity cues but remain affected by occlusion, viewpoint changes, and 2D–3D inconsistency. We therefore use frozen SAM 3 identities (Carion et al., 2026) only as auxiliary evidence and require consistency with 3D position, size, and current LiDAR support.

3 Method

3.1 Problem Formulation and Information Flow

Track2Recon recovers persistent 3D objects from offline driving logs for dynamic reconstruction. Full-sequence access provides complementary geometric evidence across viewpoints and supports conservative repair of occlusion-induced trajectory fragments.

The inputs are synchronized multi-camera images I_t^k, camera calibration $(K_k, T^t_{C_k \leftarrow L})$, LiDAR point clouds P_t, and ego poses $T^t_{W \leftarrow L}$. Each output object is described by its center $\mathbf{c}_{t,j}$, complete

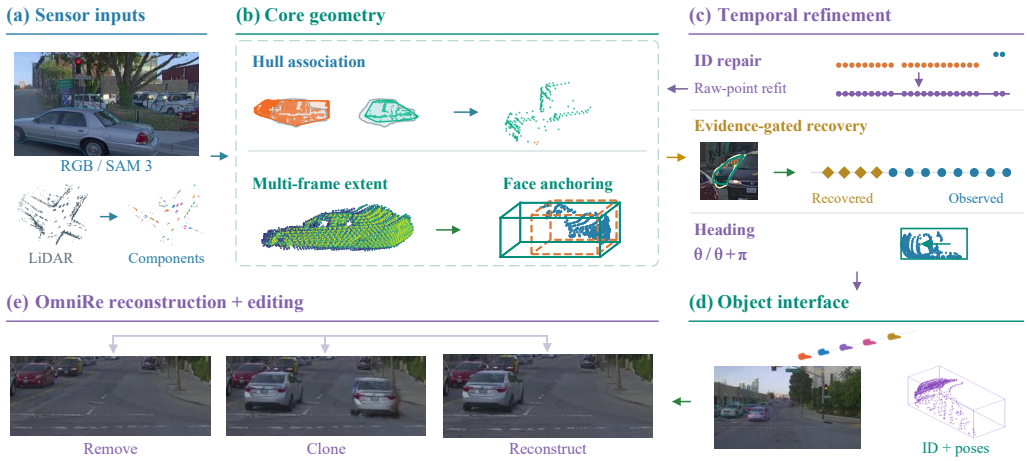

Figure 1: **Overview of Track2Recon.** Inputs are multi-camera images, LiDAR point clouds, camera calibration, and ego poses. Association determines 3D component assignment; cross-frame observations and visible-face constraints recover complete geometry; identity maintenance and observation recovery produce persistent tracks. The resulting 3D boxes and track IDs directly support OmniRe reconstruction and editing.

size $\mathbf{d}_{t,j} = (l, w, h)$, directed heading $\theta_{t,j}$, class $\kappa_{t,j}$, detection confidence $q_{t,j}$, and persistent track ID $z_{t,j}$.

As shown in Fig. 1, the system associates 2D instances with 3D components, recovers size and center along preliminary trajectories, stitches identity fragments using motion and instance-source evidence, and performs one geometry re-estimation. Fig. 2 summarizes association and complete-geometry recovery.

3.2 Hull-Supported 2D–3D Object Association

A 2D mask defines the visible object region, but viewpoint differences, occlusion, and depth projection may mix LiDAR points from different 3D components; conversely, small fragments can still achieve high hit ratios. We therefore combine local hit evidence with object-level consistency of the complete component.

Frozen SAM 3 provides per-camera instance masks. After projection filtering, ground removal, and 3D clustering, LiDAR points form candidate components C. For camera k, valid component points are projected and enclosed by a convex hull:

$$E_C^k = \mathrm{Hull}\{\pi_k(p) \mid p \in C,\ p_z^{C_k} > 0,\ \pi_k(p) \in \Omega_k\}\,. \tag{1}$$

Here, π_k is the camera projection and Ω_k the image domain; the hull provides an object-level region for mask comparison.

Object-level consistency and actual-point support. For a candidate component and non-empty mask M, correspondence uses region-level coverage r and envelope IoU J, together with point-level hit count n and hit ratio ρ. The region terms measure 2D consistency, while n and ρ enforce LiDAR support and reject matches based on sparse local hits or unsupported envelopes.

Candidates must form a non-degenerate hull, pass the n, ρ, and r gates, and satisfy $J > 0$. They are ranked lexicographically by J, r, n, and ρ, prioritizing region consistency. Within each camera, only mutual component–mask best matches are retained, with fixed-ID tie breaking. Support from one camera is sufficient; evidence from others is attached to the same 3D component to avoid duplicates.

Point-set-consistent component splitting. For spatially merged components, multiple instance masks split the parent cluster. The split is accepted only if the children satisfy point-count, size, and geometric constraints and form a complete, non-overlapping partition of the original point set; otherwise, the parent is retained.

3.3 Complete Geometry Recovery from Partial Observations

Reliable association determines object assignment, but limited-view LiDAR observes only partial surfaces. Visible span therefore underestimates full size, while symmetric expansion moves observed surfaces. We decompose recovery into two coupled problems: cross-frame observations determine *how much to complete*, and visible faces determine *where to complete*.

Multi-frame extent estimation and compensation. Let $\mathbf{c}^0$ and $\mathbf{d}^0$ denote the input center and size for this stage, including earlier visible-face compensation. After stabilizing the horizontal axes, $R_t = R(\theta_t)$ denotes the object-local frame, and raw points are transformed by $R_t^\top (p_{xy} - \mathbf{c}^0_{t,xy})$. Along axis u, the boundaries are $a_{t,u}$ and $b_{t,u}$, with visible span $s_{t,u} = b_{t,u} - a_{t,u}$.

For track z, V_z contains frames passing frontend filtering. Geometric reliability w_t combines 2D–3D overlap, actual-point hit ratio, and object-point count. The first two use the maximum over associated cameras; the point-count term grows with the square root, and both it and the final weight are clipped. Unlike detection confidence, w_t measures support for geometry estimation.

For vehicle tracks with at least three real observation frames, we first obtain a robust track-level extent estimate from a reliability-weighted high quantile of the spans:

$$\begin{aligned} d^{\text{multi}}_{z,u} &= Q^w_\alpha\left(\{s_{t,u}\}_{t\in V_z}\right), \\ d'_{t,u} &= \max\!\left(d^0_{t,u}, d^{\text{multi}}_{z,u}\right) + \epsilon_d, \qquad u \in \{x, y\}. \end{aligned} \tag{2}$$

Here, Q^w_α is the weighted high-quantile operator, obtained by interpolation over sorted cumulative weights, and ϵ_d is a small size margin. Compared with averaging, this statistic emphasizes more complete observations and reduces scale compression from partial views.

In addition to span statistics, we retain the boundary locations of each frame in the object-local coordinate system. For axis u, let the robust lower and upper boundaries in frame t be $\ell_{t,u}$ and $r_{t,u}$, respectively, and fuse them separately across frames using reliability weights:

$$\begin{aligned} \ell_{z,u} &= Q^w_\beta\left(\{\ell_{t,u}\}_{t\in V_z}\right), \\ r_{z,u} &= Q^w_{1-\beta}\left(\{r_{t,u}\}_{t\in V_z}\right), \\ d^{\text{bdry}}_{z,u} &= r_{z,u} - \ell_{z,u}. \end{aligned} \tag{3}$$

Boundary fusion lets opposite surfaces observed in different frames contribute separately to $\ell_{z,u}$ and $r_{z,u}$, so complementary views can recover the full extent even when no single frame covers the object. The final size combines this result with the robust span estimate and existing size constraints.

For sufficiently observed stationary vehicles, cross-frame boundaries are additionally fused under stable world-frame alignment and accepted only if size and center remain consistent.

Visible-face anchoring and center correction. The cross-frame extent determines object scale but not the complete box position in the current frame. We preserve the visible boundary supported by LiDAR and complete the missing extent mainly toward the unobserved side rather than expanding symmetrically about the original center. Let $\mathbf{m} = (\mathbf{a} + \mathbf{b})/2$ be the local midpoint of the visible span, $\mathbf{s} = \mathbf{b} - \mathbf{a}$ its horizontal span, and $\mathbf{o} = R^\top(\mathbf{o}^W_{xy} - \mathbf{c}^0_{xy})$ the sensor position in the same coordinates. The hard-anchored center is

$$\mathbf{c}^H_{xy} = \mathbf{c}^0_{xy} + R\left[\mathbf{m} - \operatorname{sgn}(\mathbf{o} - \mathbf{m}) \odot \frac{\mathbf{d}'_{xy} - \mathbf{s}}{2}\right]. \tag{4}$$

Here, $\odot$ denotes element-wise multiplication and $\operatorname{sgn}(0) = 0$. The sign term preserves the LiDAR-supported visible boundary and shifts the center toward the unobserved side by half of the missing extent, so size recovery and center correction jointly satisfy the visible-face constraint.

Equation 4 computes the center offset directly from the current extent-recovery result in the local coordinate system defined by $\mathbf{c}^0$.

Hard-anchor correction under viewpoint uncertainty. Hard anchoring must also account for uncertainty induced by the viewing direction. Let the normalized line-of-sight direction from the LiDAR sensor to the target be $\hat{\mathbf{v}}$ and the current object-axis direction be $\mathbf{a}$. When $|\hat{\mathbf{v}}^\top \mathbf{a}|$ is small, the line of sight is nearly perpendicular to that axis. In this grazing-view regime, determining the visible side is uncertain, and applying the full half-span compensation can cause excessive center shifts.

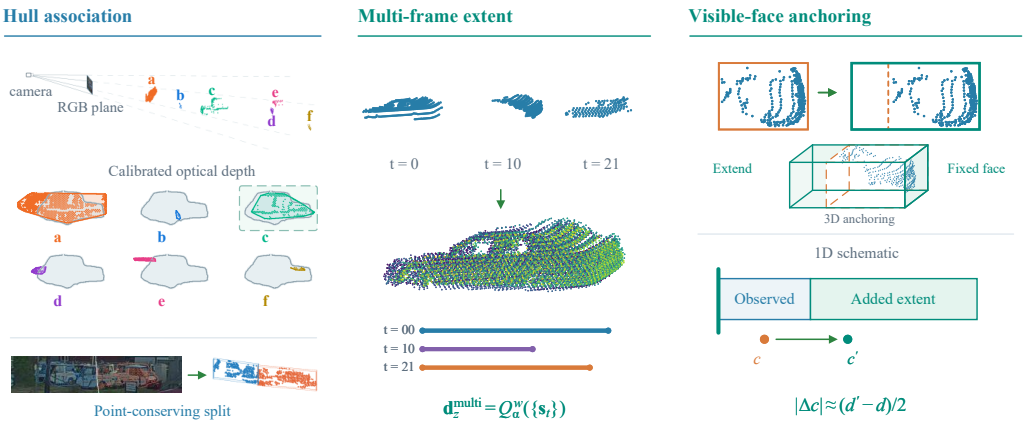

Figure 2: **Core modules.** Hull-supported association combines complete-component projections with LiDAR support for object assignment while preserving the original point set during splitting. Complete-geometry recovery fuses cross-frame boundary evidence to estimate missing extent, then uses visible faces to determine completion direction and jointly recover size and center.

Therefore, after hard anchoring and size-prior updating, we continuously attenuate the compensation according to the magnitude of the line-of-sight projection onto the object axis.

Specifically, we first estimate the actual visible span from robust quantile ranges of the current-frame raw points along axis $\mathbf{a}$, and use it to estimate the unobserved portion compensated by hard anchoring. Let the original hard-anchor displacement be $\Delta\mathbf{c}_{\text{hard}}$, and define the viewpoint-confidence weight

$$\begin{aligned} \gamma(\mathbf{a}) &= \min\left(\frac{|\hat{\mathbf{v}}^\top \mathbf{a}|}{\tau_{\text{view}}}, 1\right), \\ \Delta\mathbf{c}_{\text{rev}} &= -(1-\gamma)\Delta\mathbf{c}_{\text{hard}}, \\ \Delta\mathbf{c}_{\text{net}} &= \Delta\mathbf{c}_{\text{hard}} + \Delta\mathbf{c}_{\text{rev}} = \gamma\,\Delta\mathbf{c}_{\text{hard}}. \end{aligned} \tag{5}$$

where τ_{view} controls saturation of the viewpoint confidence. The reverse correction attenuates the hard-anchor displacement; ignoring temporal smoothing, the resulting net displacement is $\gamma\Delta\mathbf{c}_{\text{hard}}$.

Thus, $\gamma = 1$ preserves full hard-anchor compensation when the line of sight has sufficient axial support, whereas grazing views reduce γ toward zero and progressively withdraw the compensation. This avoids excessive center motion under weak viewpoint evidence. The correction is applied only to vehicle tracks with sufficient valid observations.

Height, pedestrians, and robust temporal estimation. Height recovery uses an independent ground-contact constraint. Candidate height is estimated from center-to-ground distance only for reliable ground fits consistent with low object points and the input-box bottom. Cross-frame fusion uses observation and ground-fit reliability, limits upward growth, and never shrinks the current height. Vehicle bottoms align to valid ground estimates, whereas pedestrians receive only partial correction.

For pedestrians, horizontal size uses conservative trajectory-level fusion with class-dependent lower bounds. Because smaller targets exhibit greater observation variability, we do not apply vehicle-style planar visible-face anchoring.

Finally, per-frame geometric estimates are coordinated by a robust temporal model:

$$\min_{\{x_t\}} \sum_t w_t\, \rho_\delta(x_t - \tilde{x}_t) + \lambda \sum_t \|\Delta_t^2 x\|^2, \tag{6}$$

The first term is a reliability-weighted Huber loss, while the second suppresses trajectory jitter using the actual frame intervals. We solve the problem by iteratively reweighted least squares and use separate smoothing parameters for horizontal center and height. The physical axis uses the double-angle representation $(\cos 2\theta, \sin 2\theta)$ to eliminate front–back flips modulo π.

3.4 Persistent Identity, Observation Recovery, and Object Interface

Global identity stitching under geometric constraints. After complete-geometry recovery, a temporally ordered identity graph connects vehicle trajectory fragments caused by occlusion or local association failures; pedestrians retain their previous IDs. Candidate edges combine motion, size, undirected axis, and 2D identity evidence, with 3D motion and geometry as primary constraints and 2D sources as auxiliary support. The connection cost is

$$C_{ij} = \lambda_m D_{ij}^{\mathrm{mot}} + \lambda_s D_{ij}^{\mathrm{size}} + \lambda_\theta D_{ij}^{\mathrm{axis}} + \lambda_t \Delta t_{ij} - \lambda_q S_{ij}^{\mathrm{src}}, \tag{7}$$

The motion residual is normalized by frame interval and observation reliability; size differences use log-ratios of length, width, and height; axis differences are computed modulo π; and S_{ij}^{src} measures support from shared 2D sources. Existing-identity consistency provides only auxiliary evidence and never forces a merge.

A 2D source is defined by camera, class, 2D track ID, and a fixed source tag, with reliability determined by projected overlap, actual-point hit ratio, and source score. Candidate edges must pass temporal-gap, displacement, size, and motion gates; longer-gap cross-identity connections additionally require shared-source support. Hungarian assignment uses the null match as baseline and retains only valid edges, giving each node at most one predecessor and one successor.

Limited feedback from raw observations. After identity stitching, the system regroups the original point clusters by the new track IDs and performs one additional geometry pass. For existing vehicles, the re-estimated centers and sizes are fused with the initial results, providing one-time identity-to-geometry feedback without iterative refinement.

Observation recovery and directed heading. Short-gap interpolation is separate from evidence-based recovery. For internal gaps and trajectory boundaries, recovery requires consistent motion, foreground LiDAR support, and compatible 2D evidence in the current frame, and stops when support is insufficient. Recovered boxes inherit stabilized size and undirected axis with bounded center adjustment.

Directed heading is estimated after geometry and identity. Vehicles choose between θ and $\theta + \pi$: reliable motion determines direction, while low-speed or stationary cases use a frozen shape branch calibrated with motion pseudo-labels as auxiliary evidence.

Confidence and object interface. Boxes with real observations inherit confidence from the accepted 2D source; across cameras, the maximum frozen source score is retained through later geometry, identity, and heading updates. Geometric reliability is used only for geometry. Interpolated and recovered frames receive lower confidence, with boundary-extension scores decaying with distance from the nearest real observation.

Frames sharing a persistent identity form an object $\mathcal{O}_z$ containing class, per-frame 3D boxes, track ID, world poses, and confidence scores. This interface replaces the manual boxes and trajectories required by object-level reconstructors; we use OmniRe to evaluate downstream usability.

4 Experiments

4.1 Experimental Setup and Evaluation Protocol

Datasets and baselines. We evaluate 3D detection on PandaSet (Xiao et al., 2021) and Waymo Open Dataset (Sun et al., 2020), and tracking and dynamic reconstruction on PandaSet. Detection baselines are CM3D, VESPA, OV-SCAN, and SAM3D; tracking baselines are IS-Fusion detections tracked with the PubTracker implementation from CenterPoint (IS-Fusion + PubTracker) (Yin et al., 2024; 2021) and VESPA + MCTrack (Tempfli et al., 2025; Wang et al., 2025). IS-Fusion uses

Table 1: **Cross-dataset 3D detection results.** Each metric is an equal-weight macro-average over Vehicle and Pedestrian. Best values are shown in bold.

Dataset	Method	P@2m (%)↑	R@2m (%)↑	mAP@0.5 (%)↑	3D IoU↑	Center (m)↓	Size Err. (%)↓
PandaSet	CM3D	72.37	56.68	3.72	0.3970	0.7819	12.0411
	VESPA	79.13	61.14	8.31	0.4171	0.4800	17.0834
	OV-SCAN	65.11	**73.12**	15.94	0.4570	0.5530	10.0514
	SAM3D	42.73	33.42	12.87	0.5882	0.4263	16.6568
	Track2Recon-Online	91.52	66.31	27.86	0.6493	0.3122	9.8473
	Track2Recon-Offline	**92.28**	68.08	**42.41**	**0.7075**	**0.2447**	**7.6216**
Waymo	CM3D	56.82	62.21	3.28	0.3529	1.0045	15.7166
	VESPA	64.19	79.33	14.71	0.3859	0.4288	21.1863
	OV-SCAN	66.92	**90.40**	22.36	0.4403	0.4915	14.7507
	SAM3D	14.42	57.31	14.18	0.6470	0.3368	12.9710
	Track2Recon-Online	68.74	71.36	34.29	0.6162	0.3761	12.4973
	Track2Recon-Offline	**69.38**	75.68	**47.04**	**0.6528**	**0.3302**	**10.3346**

nuScenes-pretrained weights, while VESPA detections are converted to world coordinates before MCTrack with frame-rate adaptation.

Data splits and inference settings. Detection uses 30 randomly selected PandaSet sequences and 20 Waymo sequences for Vehicle and Pedestrian. Track2Recon uses complete logs offline, while baselines retain their original inference modes. All methods share the same splits, input ranges, class mappings, and evaluation implementation.Track2Recon-Online restricts all processing to observations available up to the current frame, whereas Track2Recon-Offline uses complete-log information for cross-frame geometry recovery and global identity repair.

Detection metrics. AP uses a 3D IoU threshold of 0.5. For other metrics, predictions with score≥ 0.3 are matched one-to-one within each frame and class by Hungarian assignment with XY center distance ≤ 2 m. Precision@2m and Recall@2m follow these matches; 3D IoU, Center Error, and Size Error are computed on matched pairs, with Size Error defined as the mean absolute relative error over length, width, and height. Metrics are computed per class and equally macro-averaged over Vehicle and Pedestrian.

Table 2: **PandaSet 3D multi-object tracking results.**

Method	HOTA (%)↑	DetA (%)↑	AssA (%)↑	IDF1 (%)↑	MOTA (%)↑	IDS↓	Frag↓
IS-Fusion + PubTracker	33.32	23.49	52.58	31.97	-56.31	116	469
VESPA + MCTrack	19.36	18.52	21.15	12.15	-46.73	1057	437
Track2Recon-Online	47.83	38.46	64.92	55.71	35.18	76	132
Track2Recon-Offline	**54.71**	**43.11**	**72.99**	**65.34**	**44.01**	**39**	**68**

Tracking metrics. Tracking uses oriented 3D IoU as detection similarity. HOTA (Luiten et al., 2021) computes global GT–prediction identity alignment from sequence-level co-occurrence and then performs per-frame Hungarian matching. HOTA, DetA, and AssA are averaged over $\alpha \in \{0.05, 0.10, \ldots, 0.95\}$; IDF1 (Ristani et al., 2016) uses global identity matching. MOTA, IDS, and Frag follow CLEAR MOT (Bernardin & Stiefelhagen, 2008) with oriented 3D IoU ≥ 0.5.

Reconstruction metrics. For reconstruction, OmniRe training and evaluation settings are fixed, with only the input 3D bounding boxes and track IDs replaced. PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018) are reported separately for full images and dynamic regions.

4.2 Cross-Dataset 3D Detection

Table 1 shows that Track2Recon-Offline achieves 42.41% and 47.04% mAP@0.5 on PandaSet and Waymo, outperforming the best baselines by 26.47 and 24.68 points, respectively. It also gives the best 3D IoU, center error, and size error, while OV-SCAN has higher recall. The Online version already improves mAP over the baselines, and Offline processing further improves detection and geometry.

4.3 Persistent Identity and Tracking Quality

Table 2 shows that Track2Recon-Offline improves HOTA and IDF1 over IS-Fusion + PubTracker by 21.39 and 33.37 points, while reducing IDS from 116 to 39 and Frag from 469 to 68. These



Figure 3: **Qualitative BEV detection comparison.** From left to right, the columns compare CM3D, VESPA, OV-SCAN, SAM3D, Track2Recon, and GT across representative scenes.

results indicate stronger cross-frame identity consistency, and Offline processing further improves all tracking metrics over the Online version.

4.4 Core Module Ablations

Table 3 reports stage-wise effects of hull-supported association, multi-frame extent recovery, and visible-face anchoring on detection, geometry, and tracking.

Hull-supported association. Component-level projection consistency increases 3D IoU from 0.5584 to 0.5749 and reduces center and size errors, showing that complete-component envelopes provide stronger assignment evidence than local point hits.

Table 3: **Stage-wise comparison of core modules.** Hull-supported association, multi-frame extent recovery, and visible-face anchoring are added progressively to point-cloud initialization.

Configuration	mAP@0.5 (%)↑	3D IoU↑	Center (m)↓	Size Err. (%)↓	HOTA (%)↑	IDF1 (%)↑
Point-only initialization	23.76	0.5584	0.2911	10.9748	41.88	44.46
+ Hull-supported association	23.41	0.5749	0.2796	10.4931	42.73	45.66
+ Multi-frame extent recovery	35.78	0.6848	0.2613	**7.5734**	51.13	56.97
+ **visible-face anchoring**	**42.41**	**0.7075**	**0.2447**	7.6216	**54.71**	**65.34**

Table 4: **OmniRe reconstruction results with different object inputs.** Bold indicates the best result among automatic object inputs.

Object input	Full image			Dynamic region		
	PSNR (dB)↑	SSIM↑	LPIPS↓	PSNR (dB)↑	SSIM↑	LPIPS↓
GT 3D boxes + track IDs	27.4692	0.8762	0.1426	26.5569	0.8488	0.0715
IS-Fusion + PubTracker	26.4902	0.8398	0.1816	23.6280	0.7551	0.1450
VESPA + MCTrack	26.2338	0.8242	0.1725	22.5388	0.7002	0.1570
Track2Recon	**27.2228**	**0.8703**	**0.1509**	**25.7279**	**0.8154**	**0.0930**

4.5 Downstream OmniRe Reconstruction and Editing

Table 4 shows that Track2Recon inputs outperform both detector–tracker baselines and approach GT-input reconstruction. Full-image PSNR is within 0.2464 dB of GT; in dynamic regions, compared with IS-Fusion + PubTracker, PSNR and SSIM improve while LPIPS decreases. These gains reflect better geometry and identity consistency for cross-frame aggregation.

Per-frame 3D bounding boxes and persistent track IDs organize dynamic instances as independent object nodes in canonical space, thereby supporting editing operations such as object removal, duplication, and relocation (Fig. 1). The reconstruction and editing results demonstrate the downstream effect of connecting Track2Recon outputs to an existing object-level reconstructor.



Figure 4: **OmniRe reconstruction comparison.** From left to right, the inputs are IS-Fusion + PubTracker, VESPA + MCTrack, Track2Recon, and GT objects; zoomed regions show dynamic-object contours and separation between neighboring instances.

5 Conclusion

Track2Recon recovers per-frame 3D boxes and persistent identities from unlabeled driving logs without manual 3D annotations or supervised target-domain training. Its component-level projection and LiDAR support resolve 2D–3D assignment, while cross-frame extent estimation and visible-face anchoring recover size and center from partial observations. Results on PandaSet and Waymo show strong detection performance; PandaSet tracking and downstream OmniRe experiments show that the recovered objects support persistent reconstruction and editing.

AI Use Statement

Generative AI assisted writing, language polishing, related-work discovery, and research discussion.

References

Stefan Andreas Baur, Frank Moosmann, and Andreas Geiger. LISO: Lidar-only self-supervised 3D object detection. In *Computer Vision – ECCV 2024*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/11661_ECCV_2024_paper.php.

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

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=r35clVtGzw.

Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo de Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gojcic, Sanja Fidler, Marco Pavone, Li Song, and Yue Wang. OmniRe: Omni urban scene reconstruction. In *International Conference on Learning Representations*, 2025. URL https://ziyc.github.io/omnire/.

Adrian Chow, Evelien Riddell, Yimu Wang, Sean Sedwards, and Krzysztof Czarnecki. OV-SCAN: Semantically consistent alignment for novel object discovery in open-vocabulary 3D object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2503.06435.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 226–231. AAAI Press, 1996.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

Mehar Khurana, Neehar Peri, James Hays, and Deva Ramanan. Shelf-supervised cross-modal pretraining for 3D object detection. In *Conference on Robot Learning*, 2024. URL https://arxiv.org/abs/2406.10115.

Jonathon Luiten, Aljoša Ošep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. HOTA: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 129:548–578, 2021. doi: 10.1007/s11263-020-01375-2. URL https://link.springer.com/article/10.1007/s11263-020-01375-2.

Tao Ma, Xuemeng Yang, Hongbin Zhou, Xin Li, Botian Shi, Junjie Liu, Yuchen Yang, Zhizheng Liu, Liang He, Yu Qiao, Yikang Li, and Hongsheng Li. DetZero: Rethinking offboard 3D object detection with long-term sequential point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6736–6747, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Ma_DetZero_Rethinking_Offboard_3D_Object_Detection_with_Long-term_Sequential_Point_ICCV_2023_paper.html.

Ziqi Pang, Zhichao Li, and Naiyan Wang. SimpleTrack: Understanding and rethinking 3D multi-object tracking. *arXiv preprint arXiv:2111.09621*, 2021. URL https://arxiv.org/abs/2111.09621.

Charles R. Qi, Yin Zhou, Mahyar Najibi, Pei Sun, Khoa Vo, Boyang Deng, and Dragomir Anguelov. Offboard 3D object detection from point cloud sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6134–6144, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Qi_Offboard_3D_Object_Detection_From_Point_Cloud_Sequences_CVPR_2021_paper.html.

Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Computer Vision – ECCV 2016*

Workshops, pp. 17–35, 2016. doi: 10.1007/978-3-319-48881-3_2. URL https://link.springer.com/chapter/10.1007/978-3-319-48881-3_2.

Chenpeng Su, Wenhua Wu, Chensheng Peng, Tianchen Deng, Zhe Liu, and Hesheng Wang. DIAL-GS: Dynamic instance aware reconstruction for label-free street scenes with 4D gaussian splatting. In *IEEE International Conference on Robotics and Automation*, 2026. URL https://arxiv.org/abs/2511.06632.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo Open Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2446–2454, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Sun_Scalability_in_Perception_for_Autonomous_Driving_Waymo_Open_Dataset_CVPR_2020_paper.html.

Levente Tempfli, Esteban Rivera, and Markus Lienkamp. VESPA: Towards un(human)supervised open-world pointcloud labeling for autonomous driving. *arXiv preprint arXiv:2507.20397*, 2025. URL https://arxiv.org/abs/2507.20397.

Xiyang Wang, Shouzheng Qi, Jieyou Zhao, Hangning Zhou, Siyu Zhang, Guoan Wang, Kai Tu, Songlin Guo, Jianbo Zhao, Jian Li, and Mu Yang. MCTrack: A Unified 3D Multi-Object Tracking Framework for Autonomous Driving. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2025. URL https://github.com/megvii-research/MCTrack.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861. URL https://ieeexplore.ieee.org/document/1284395.

Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 3D multi-object tracking: A baseline and new evaluation metrics. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020. URL https://arxiv.org/abs/1907.03961.

Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, Yunlong Wang, and Diange Yang. PandaSet: Advanced sensor suite dataset for autonomous driving. *arXiv preprint arXiv:2112.12610*, 2021. URL https://arxiv.org/abs/2112.12610.

Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians: Modeling dynamic urban scenes with gaussian splatting. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2401.01339.

Bin Yang, Min Bai, Ming Liang, Wenyuan Zeng, and Raquel Urtasun. Auto4D: Learning to label 4D objects from sequential point clouds. *arXiv preprint arXiv:2101.06586*, 2021. URL https://arxiv.org/abs/2101.06586.

Junbo Yin, Jianbing Shen, Runnan Chen, Wei Li, Ruigang Yang, Pascal Frossard, and Wenguan Wang. IS-Fusion: Instance-scene collaborative fusion for multimodal 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3D object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Yurong You, Katie Luo, Cheng Perng Phoo, Wei-Lun Chao, Wen Sun, Bharath Hariharan, Mark Campbell, and Kilian Q. Weinberger. Learning to detect mobile objects from LiDAR scans without labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1130–1140, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/You_Learning_To_Detect_Mobile_Objects_From_LiDAR_Scans_Without_Labels_CVPR_2022_paper.html.

Dingyuan Zhang, Dingkang Liang, Hongcheng Yang, Zhikang Zou, Xiaoqing Ye, Zhe Liu, and Xiang Bai. SAM3D: Zero-shot 3D object detection via segment anything model. *arXiv preprint arXiv:2306.02245*, 2023a. URL https://arxiv.org/abs/2306.02245.

Lunjun Zhang, Anqi Joyce Yang, Yuwen Xiong, Sergio Casas, Bin Yang, Mengye Ren, and Raquel Urtasun. Towards unsupervised object detection from LiDAR point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9317–9328, 2023b. URL https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_Towards_Unsupervised_Object_Detection_From_LiDAR_Point_Clouds_CVPR_2023_paper.html.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. doi: 10.1109/CVPR.2018.00068. URL https://arxiv.org/abs/1801.03924.

Ruida Zhang, Chengxi Li, Chenyangguang Zhang, Xingyu Liu, Haili Yuan, Yanyan Li, Xiangyang Ji, and Gim Hee Lee. Street gaussians without 3D object tracker. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2412.05548.

A Local Mechanisms and Module Coordination

A.1 Component-Level Cross-Modal Disambiguation

Point hits inside a 2D mask provide local correspondences, while the projected envelope of a component describes the spatial relationship between the entire candidate and the 2D instance. Track2Recon combines both: the envelope measures whether the component can explain the instance at the object scale, while actual LiDAR hits verify the physical support for the correspondence, thereby distinguishing candidates with similar local projections but different object assignments.

This division of labor allows frontend clustering to focus on generating coarse candidates, while subsequent association further determines object assignment. Observations from multiple cameras that refer to the same component are merged into unified evidence; splitting of merged components reallocates the original points of the parent cluster while preserving mutual exclusivity and completeness of the point set.

A.2 Multi-Frame Extent Recovery and Visible-Face Anchoring

Single-frame LiDAR mainly covers the visible surface of an object, so its observed span only provides a lower bound on the full extent. Track2Recon aligns observations of the same instance across frames into the object-local coordinate system and fuses boundary evidence at both ends of each physical axis. Surfaces observed from different viewpoints and times can therefore jointly recover the object's full extent without requiring complete coverage in any single frame.

Extent recovery determines the completion scale, while visible-face anchoring further determines the placement of the complete box. It preserves the visible-side boundary supported by actual points and completes the missing extent mainly toward the unobserved side, jointly correcting size and center. The geometric stage uses undirected principal axes modulo π, handled separately from vehicle front–back direction estimation.

A.3 Persistent Identity and Raw-Observation Feedback

Global identity stitching combines motion continuity, size consistency, undirected principal axes, and reliable 2D identity sources to establish connections between trajectory fragments that satisfy the association conditions, while leaving other candidates disconnected. Identity stitching first updates track IDs while keeping existing 3D boxes unchanged.

When the repaired identities aggregate more real observations, the system reorganizes the original LiDAR point clusters by the new IDs, re-estimates complete geometry, and conservatively fuses it with the initial results. This feedback is executed once, allowing the observation gains from identity stitching to further improve geometric recovery and forming a coordinated update between geometry and identity.

A.4 Observation Recovery and Directed Heading

Internal-gap recovery combines motion consistency between the real observations on both sides with 2D instance evidence and LiDAR support in the current frame. Boundary extension verifies these signals frame by frame and terminates when support becomes insufficient. Recovered states inherit the stabilized size and undirected principal axis, while the center is constrained by the current point cloud.

Directed heading is estimated after geometry and identity recovery are complete. When motion is sufficiently strong and stable, motion direction is used; at low speed or while stationary, multi-frame shape information assists the decision. The two candidate directions differ by π. During evaluation, the prediction–GT matching is fixed and the directed-heading error is reported separately.

B Implementation and Reproducibility Details

B.1 Coordinate Systems and Information Flow

World coordinates are used to remove ego motion and establish trajectories, object-local coordinates describe physical extent and shape, and image coordinates verify instance support. Information updates are clearly separated across stages: projected envelopes are used for cross-modal consistency checks, identity stitching updates track IDs, identity feedback re-estimates geometry from LiDAR points within each trajectory, and the directed-heading module selects the front–back orientation.

B.2 Candidate Generation and Initial Box Fitting

Candidate components. The frontend constructs 3D candidates from multi-camera instance masks and calibrated LiDAR point clouds. The point cloud is first lightly voxelized and then spatially clustered into coarse components; final geometric estimation uses the original-resolution point cloud. Clustering provides an initial separation of different objects or depth layers, while local under-segmentation, over-segmentation, and boundary noise are further handled by subsequent 2D–3D association and temporal observations. Neighborhood scales are set by dataset and class and kept fixed for all sequences under the same setting.

Boundary-supported rectangle. The initial horizontal box is fitted using a boundary-supported rectangle. For a candidate axis $\phi \in [0, \pi/2)$, points are transformed to $R_\phi^\top p_{xy}$, and axis-wise extrema $\mathbf{a}, \mathbf{b}$ are used to obtain the span $\mathbf{s} = \mathbf{b} - \mathbf{a}$. A point within 0.15 m of an edge contributes a support interval of length 0.24 m along that edge; the interval is centered at the projected point and clipped to the edge bounds. The fitting objective maximizes the ratio between the total union length of support intervals on all four edges and the rectangle perimeter. Ties are resolved in order by selecting the candidate with smaller area, greater support length, more support points, and a direction closer to the minimum-area rectangle.

B.3 Observation Weighting and Geometric Stabilization

Observation reliability. The observation weight combines 2D–3D overlap quality J_t, actual-point support ratio ρ_t, and object point count N_t:

$$w_t = \phi(J_t, \rho_t, N_t), \tag{A1}$$

Here, $\phi(\cdot)$ increases monotonically with all three evidence terms and clips extreme observations to limit their weights. Center estimation combines a robust data term with temporal smoothing; the undirected axis is represented by $(\cos 2\theta, \sin 2\theta)$ to handle axial changes modulo π. Complete size is estimated from weighted high-quantile extents across frames to reduce scale compression caused by local visibility.

Local ground constraints are enabled only when point-cloud support is sufficient and compatible with the current geometry. For small-scale objects such as pedestrians, smaller-magnitude corrections are applied to size and center.

Vehicle size priors. For persistently under-observed objects, the reference size for a passenger car is set to $4.2 \times 1.9 \times 1.5$ m (length, width, height). Lower bounds on the initial box size are $2.8 \times 1.2 \times 0.9$ m for passenger cars, $3.5 \times 1.5 \times 1.2$ m for trucks, and $6.0 \times 2.0 \times 2.0$ m for Waymo buses.

Pedestrian size priors. The lower bound on the initial pedestrian size is $0.5 \times 0.4 \times 1.2$ m. Subsequent pedestrian completion uses a horizontal-size lower bound of 0.9×0.75 m. These priors provide lower bounds for under-observed objects, while the final size may increase further when supported by instance observations.

B.4 Evidence-Constrained Observation Recovery

Internal-gap recovery combines motion predictions from the real observations before and after the gap and requires the current image and LiDAR to both provide object support. Track boundaries are extended by frame-by-frame search: extension stops when the 2D instance evidence, actual-point hits, or component consistency clearly deteriorates. Recovered boxes inherit the stabilized size and undirected axis, and their centers are adjusted within bounded ranges supported by the current point cloud.

B.5 Directed Heading and Shape Calibration

Antisymmetric shape features. The physical-box recovery stage estimates an undirected principal axis modulo π. Directed heading then determines the vehicle front–back direction. For vehicles with insufficient motion information, the system aggregates real multi-frame point clouds in object-local coordinates and describes shape differences between the two ends using statistics such as longitudinal occupancy, local height, and lateral span. Let the local shape descriptor be $f(P, \theta)$; the antisymmetric feature is defined as

$$g(P, \theta) = \frac{f(P, \theta) - f(P, \theta + \pi)}{2}, \qquad g(P, \theta + \pi) = -g(P, \theta). \tag{A2}$$

This construction causes the feature to change sign under a front–back flip and therefore directly corresponds to the two candidate directions. Shape-based discrimination is used when the track has sufficiently reliable shape observations; otherwise, the original undirected axis is retained.

One-time calibration. The heading model uses stable motion directions from development trajectories as pseudo-labels, quality-weights multi-frame shape features, and trains a lightweight linear classifier to distinguish the two directions of the principal axis. The model is frozen at inference time and only selects whether to apply a π flip, leaving the center, size, detection score, track ID, and number of boxes unchanged.

C Supplementary Comparison with Related Methods

2D projection and object assignment. CM3D aggregates LiDAR points inside a mask and uses mask erosion to reduce background interference near boundaries. VESPA addresses points behind the target that project into the foreground mask using spatial clustering and dominant connected-component denoising. OV-SCAN explicitly constructs candidate relations between 2D proposals and 3D components. These methods address projection ambiguity from different perspectives. Track2Recon further checks consistency with the 2D instance at the scale of the complete component and verifies the correspondence using actual point hits.

Visible surfaces and complete extent. The point-set center in CM3D tends to shift toward the sensor-visible side. VESPA compensates under-observed boxes using class sizes and the relative ego direction. OV-SCAN estimates complete boxes using class box anchors and single-frame geometric search. Track2Recon recovers instance-specific physical extents from complementary cross-frame observations of the same instance and then determines the completion direction from the currently visible face, coupling scale estimation with center placement.

Persistent objects and dynamic reconstruction. Object-level dynamic reconstruction requires instance identities and physical extents that are consistent over time. Overly wide boxes can include background or neighboring instances; identity merging can aggregate observations from different objects into the same canonical space; and trajectory fragmentation reduces the cross-frame evidence available to each object. Track2Recon jointly recovers object assignment, complete geometry, and persistent identity, and exposes the final 3D boxes and track IDs through a unified object interface for existing dynamic-scene reconstructors.

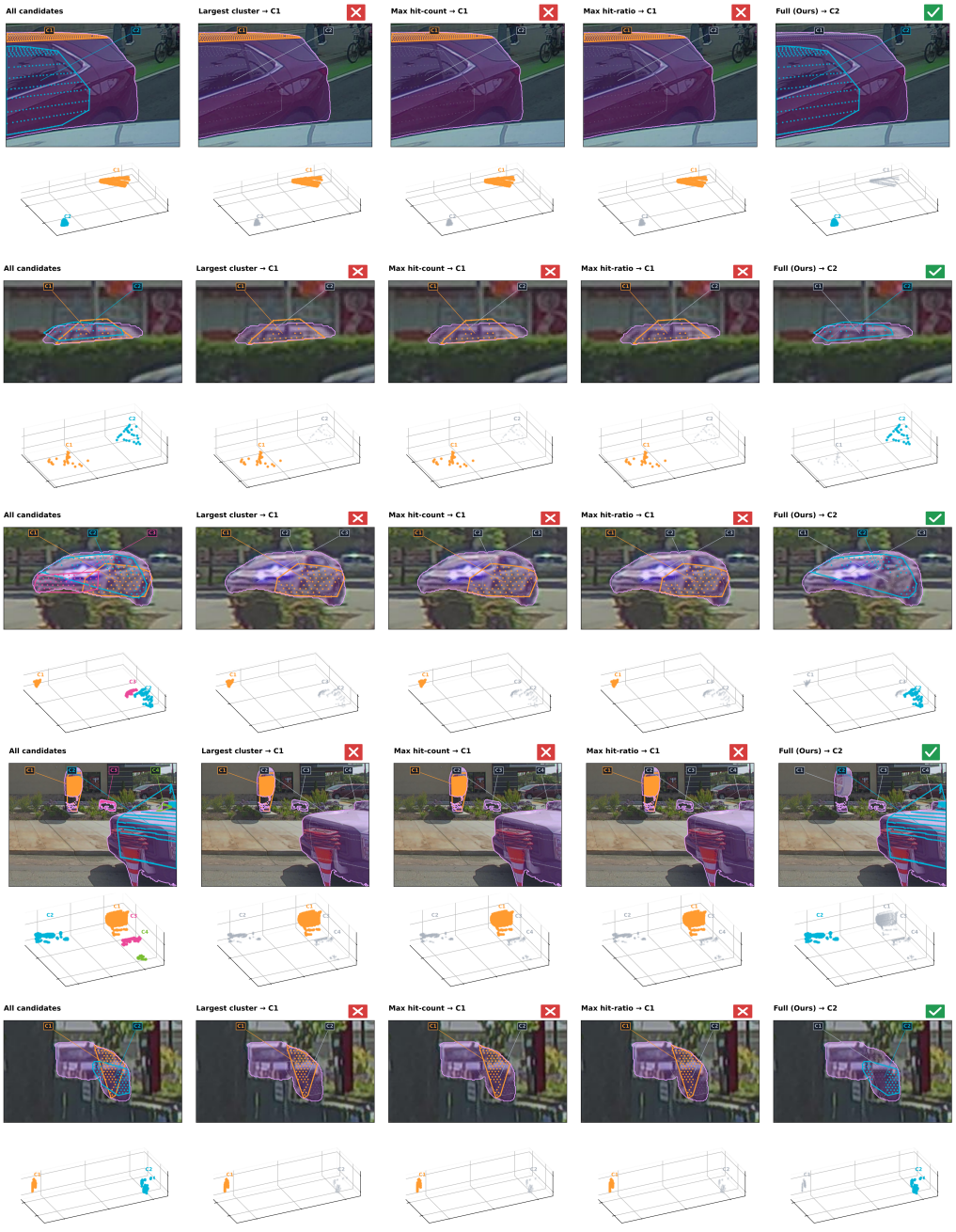

Figure A1: **Visualization of association-rule ablations.** The five rows show five scene cases, and the five columns show all candidates, Largest cluster, Max hit-count, Max hit-ratio, and Full (Ours), respectively. A red cross marks a rule selecting the wrong component, while a green check marks the correct component.

D SUPPLEMENTARY EXPERIMENTS AND ANALYSIS

D.1 DETAILED CORE ABLATION ANALYSIS

Multi-frame extent recovery. Cross-frame extent and boundary fusion reduces size error from 10.4931% to 7.5734%, raises 3D IoU from 0.5749 to 0.6848, and increases mAP from 23.41% to 35.78%. Complementary frame-wise boundaries recover opposite sides without requiring any single frame to fully observe the object.

Visible-face anchoring. Visible-face anchoring reduces center error from 0.2613 m to 0.2447 m, raises mAP to 42.41%, and brings HOTA/IDF1 to 54.71%/65.34%, improving both spatial accuracy and temporal consistency.

D.2 END-TO-END COMPARISON OF ASSOCIATION RULES

Table A1: **End-to-end comparison of association rules.** The complete system uses each association rule in turn and reports detection, tracking, and dynamic-region reconstruction metrics.

Association rule	Detection and tracking			Dynamic-region reconstruction		
	mAP@0.5 (%)↑	HOTA (%)↑	IDF1 (%)↑	PSNR (dB)↑	SSIM↑	LPIPS↓
Largest cluster	29.76	36.88	40.66	24.1264	0.7805	0.1193
Max hit-count	30.97	38.93	38.01	24.3186	0.7987	0.1278
Max hit-ratio	25.27	34.63	36.87	23.9462	0.7561	0.1114
Full	**42.41**	**54.71**	**65.34**	**25.7279**	**0.8154**	**0.0930**

We further compare four association rules in the complete pipeline: Largest cluster selects candidates by component size, Max hit-count by the number of actual point hits, Max hit-ratio by hit ratio, and Full jointly uses component-level projection consistency and actual-point support. Fig. A1 shows the candidate components and assignment choices of each rule in five cases, and Table A1 summarizes detection, tracking, and dynamic-region reconstruction results.

The Full association rule achieves the best results across all three tasks. Rules based only on component size or local point hits have difficulty stably determining object assignment when a single mask contains components at different depths or when components are fragmented or merged. Component-level consistency organizes local point evidence into object-level correspondence, and its benefits further propagate to trajectory construction and dynamic-region reconstruction.

Table A2: **3D tracking results under different association rules.**

Association rule	DetA (%)↑	AssA (%)↑	MOTA (%)↑	IDS↓	Frag↓
Largest cluster	26.65	53.66	15.08	148	310
Max hit-count	28.30	51.65	9.34	124	406
Max hit-ratio	26.40	48.27	-6.66	171	577
Full	**43.11**	**72.99**	**44.01**	**39**	**68**

Table A2 reports more complete tracking metrics. Full outperforms the three simplified rules in DetA, AssA, and MOTA, while reducing IDS and Frag to 39 and 68, respectively. More accurate initial object assignment improves not only single-frame geometry but also cross-frame identity continuity.

D.3 DIRECTED-HEADING EVALUATION UNDER FIXED GEOMETRY

Directed-heading error is computed with a 2π period, while the principal-axis error of the physical box is computed with a π period. This experiment fixes the original prediction–GT matching and keeps the number of objects, centers, sizes, and track identities unchanged, comparing vehicle front–back direction errors before and after heading calibration.

As shown in Table A3, the directed-heading error for all matched PandaSet objects decreases from 54.20° to 20.85°, and that on the held-out set decreases from 58.75° to 22.54°; on Waymo, the error decreases from 50.13° to 32.68°. With physical geometry and matching fixed, multi-frame shape information effectively improves discrimination of vehicle front–back direction.

Table A3: **Vehicle directed-heading evaluation under fixed geometry and matching.** The left columns report matching scale and geometric metrics, while the right columns compare errors before and after heading calibration.

Evaluation set	Matches	Fixed geometry			Directed heading error (°)↓	
		Center (m)↓	Size Err. (%)↓	Axis Err. (°)↓	Before calibration	After calibration
PandaSet (all)	4227	0.2465	6.8277	5.37	54.20	20.85
PandaSet (held-out)	3000	0.2585	6.9097	6.12	58.75	22.54
Waymo	536	0.4094	9.3179	2.65	50.13	32.68

D.4 Additional Implementation Details

DBSCAN (Ester et al., 1996) generates 3D candidates. PandaSet uses $\epsilon = 0.45$ m and min-samples$= 5$ for both classes; Waymo uses the same setting for vehicles and $\epsilon = 0.30$ m, min-samples$= 3$ for pedestrians. Geometry is estimated from the original LiDAR points. Local association uses world-frame constant-velocity prediction with a maximum observation gap of 2. Multi-frame recovery uses 2D–3D consistency, actual-point support, point count, and weighted high-quantile size estimation. Recovery requires current-frame 2D and LiDAR evidence; low-motion heading uses multi-frame shape cues. Other parameters are fixed across sequences except class/dataset clustering and scale settings.

D.5 Supplementary Visualizations

D.5.1 Local Consecutive Frames

Figure A2 compares 3D boxes and track identities across consecutive frames, showing localization stability and identity continuity within a local temporal window.



Figure A2: **Local consecutive-frame tracking comparison.** Columns compare IS-Fusion + PubTracker, VESPA + MCTrack, Track2Recon, and GT across consecutive frames.

D.5.2 Full-Sequence Comparisons

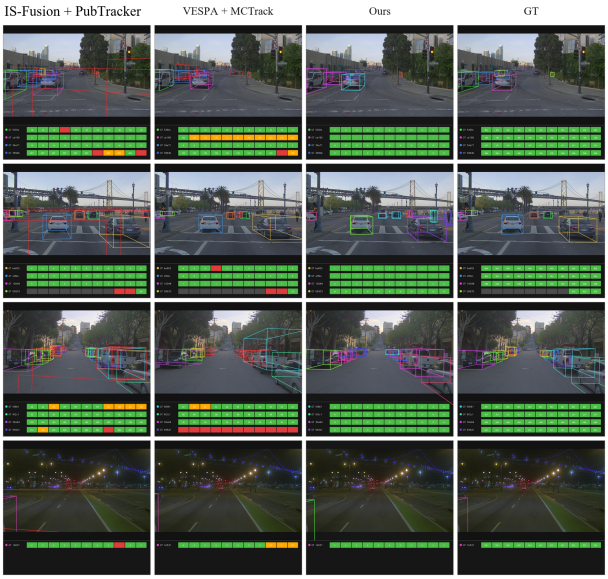

Figure A3: **Full-sequence PandaSet tracking.** Columns compare IS-Fusion + PubTracker, VESPA + MCTrack, Track2Recon (Ours), and GT; rows show four scenes.



Figure A4: **Additional full-sequence tracking comparisons.** Rows show four scenes; columns compare IS-Fusion + PubTracker, VESPA + MCTrack, Track2Recon (Ours), and GT.

D.6 Detection Evaluation Sequences

PandaSet (30): `001, 004, 006, 008, 011, 012, 015, 016, 021, 024, 028, 035, 041, 042, 044, 053, 063, 065, 070, 084, 088, 089, 101, 106, 109, 120, 122, 123, 149, 158.`

Waymo (20; short IDs): `100508, 102036, 102062, 119714, 474717, 106762, 117240, 128796, 132384, 143481, 148697, 152217, 152706, 153495, 163453, 177619, 342571, 398895, 881121, 938501.`