Digital Intelligence PRESS

Journal of Discovery Core

https://doi.org/10.67541/jdc2612

Research Article **Volume 1, Issue 2 (Sep 2026)** *Open Access*

Deep Learning Based Dynamic Object Filtering for Semantic Information Assisted Visual LiDAR SLAM in Dynamic Scenes

Xianlong Zhao *

Swinburne College, Shandong University of Science and Technology, Jinan, Shandong, China

***Corresponding author:** Xianlong Zhao. Email: zxl041108@126.com

Abstract

To address the structural mismatch between semantic priors and motion evidence in tightly coupled visual-LiDAR SLAM, this paper proposes a dynamic object filtering architecture centered on evidence arbitration. Semantic segmentation outputs are transformed from binary rejection masks into continuous spatial modulation coefficients ranging from 0 to 1. These coefficients guide the generation of two complementary streams of motion evidence: visual optical flow residuals and multi-frame LiDAR occupancy consistency. Log-odds fusion then produces cross-modal confidence and its uncertainty. Using semantic instances as arbitration units, the method resolves conflicts within a sliding window of 6-8 frames by jointly considering mean confidence, temporal stability, and spatial concentration, with an upper decision delay bound of 3 frames. Finally, dynamic probabilities are embedded in the factor graph as information matrix modulation factors, replacing direct front-end rejection. Experiments on Semantic KITTI, KITTI raw, and Urban Nav show that, while maintaining a dynamic-point residual rate no higher than that of the hard-filtering baseline, the method increases the static semantic constraint retention rate from 0.42 to 0.78. Decision accuracy reaches 0.86 and 0.83 for conflict modes (b) and (c), respectively, and trajectory errors during transition intervals are substantially lower than those of the comparison variants. A breakdown of per-frame processing time identifies semantic segmentation as the main bottleneck. The system meets real-time constraints when the segmentation frame rate is at least 8 Hz, and the dynamic instance density does not exceed 15 instances per frame.

Keywords

Visual-LiDAR SLAM; Dynamic object filtering; Evidence arbitration; Semantic modulation; Tightly coupled optimization

Received: 13 Aug 2026; Revised: 10 Sep 2026; Accepted: 18 Oct 2026; Published: 30 Oct 2026
Copyright: ©2026 The Author(s). Published by Digital Intelligence Press Limited. This is an open access article under the CC BY 4.0 license.

1. INTRODUCTION

SLAM systems integrating vision and LiDAR have well-established observation models and optimization frameworks for static environments. Dynamic objects, however, invalidate both observation streams in different ways. On the visual side, they violate the fundamental assumption underlying feature correspondences: the inter-frame pixel displacement of a feature can no longer be explained solely by camera ego-motion [1],[2]. The resulting mismatches systematically bias fundamental and essential matrix estimates, and these biases propagate through the initial pose estimates into the entire optimization problem [3]. On the LiDAR side, dynamic objects introduce temporal inconsistencies in point cloud clusters. The same spatial location is occupied across frames by objects in different motion states, causing registration to mistake object displacement for platform displacement and producing trails and ghosting in the map. Crucially, these two forms of contamination are not merely

additive in a tightly coupled framework. Biased visual pose initialization alters the correspondence search region for LiDAR registration, while incorrect LiDAR registration in turn contaminates the three-dimensional constraints on visual features [4],[5]. Both errors reinforce one another through the shared pose state, creating coupled error amplification. Consequently, improvements from dynamic object filtering in only one stream can be offset by residual contamination in the other.

Existing dynamic filtering approaches fall into two main categories whose limitations form a structural contrast. Approaches based on semantic segmentation priors use semantic category membership as a hard criterion for motion and directly exclude observations of potentially dynamic classes, such as vehicles and pedestrians, from optimization [6],[7]. Although effective for genuinely moving objects, this strategy cannot distinguish objects that belong to dynamic semantic classes but are actually stationary, such as roadside parked vehicles, queues waiting at traffic lights, and stationary pedestrians. These objects could provide high-quality constraints, yet indiscriminate rejection sharply reduces the available constraints and can even compromise odometry observability in scenes with many potentially dynamic objects but little actual motion. Geometric motion detection instead relies on motion evidence itself and can capture moving objects missing by semantic detection. However, it is highly sensitive to LiDAR sparsity and occlusion: sparse regions fragment motion evidence, occlusions create false temporal inconsistencies, and pointwise or cluster wise detection is computationally demanding under real-time constraints [8],[9]. Most existing fusion schemes use the two approaches sequentially, first narrowing the geometric detection search space with a semantic mask and then rejecting observations according to geometric results. This reduces geometric processing costs but leaves the central issue unresolved. When semantic priors and motion evidence disagree, the later stage simply overrides the earlier one; the conflict itself is neither explicitly represented nor arbitrated. Stationary vehicles classified as dynamic by semantics but static by geometry, and missed objects classified as static by semantics but moving by geometry, thus produce systematic errors in opposite directions.

The central argument of this paper is that dynamic object filtering is fundamentally an evidence arbitration problem rather than a classification problem. Semantic priors and motion evidence carry uncertainty from different sources: segmentation quality limits the former, while sampling density and occlusion limit the latter. Their degree of agreement is informative about an object's motion state, but existing sequential schemes and hard criteria discard this information. We therefore construct a filtering architecture that uses semantic information as a structural prior and cross-modal motion consistency for verification. Its key design changes semantic labels from rejection instructions into attention modulation signals. Semantics no longer directly determines retention or rejection; instead, it modulates the spatial scope and thresholds used to compute motion evidence. This preserves opportunities to verify stationary objects in dynamic semantic classes and allows moving objects missed by semantics to obtain contradictory motion signals through the LiDAR stream. Disagreement between semantic and motion evidence is explicitly quantified as uncertainty and resolved by temporal arbitration at the semantic instance level [10]. The resulting continuous dynamic probabilities modulate the information matrices in tightly coupled optimization instead of being binarized into front-end rejection decisions. The contributions are threefold: a construction of cross-modal motion evidence that makes visual and LiDAR evidence complementary under semantic modulation rather than sequential; a semantic–motion conflict arbitration mechanism that replaces independent single-frame decisions with arbitration constrained by temporal consistency; and a dynamic-aware factor graph redesign that embeds dynamic probabilities as information matrix modulation factors, enabling back-end optimization to refine motion decisions. These contributions address three successive questions: which signals support a decision,

how conflicts are resolved, and how decisions enter optimization. Together, they provide an operational implementation of evidence arbitration.

2. SEMANTIC MODULATIONS FOR CROSS-MODAL DYNAMIC EVIDENCE GENERATION

2.1 Soft modulation of semantic priors

Let a semantic segmentation network output a class probability distribution $p_i = [p_i^1, \dots, p_i^C]$ for pixel i in an image, where C is the number of classes. Denote the set of potentially dynamic classes by $\mathcal{D}$, including vehicles, pedestrians, and cyclists, and the set of static classes by $\mathcal{S}$. Hard filtering constructs a binary mask $m_i^{\text{hard}} = \mathbb{1}[\arg\max_c p_i^c \in \mathcal{D}]$ and excludes pixels satisfying $m_i^{\text{hard}} = 1$ from all subsequent calculations. We instead construct a continuous modulation coefficient:

$$\alpha_i = 1 - \sigma\left(\beta \cdot \left(\sum_{c\in\mathcal{D}} p_i^c - \tau\right)\right) \tag{1}$$

Here, $\sum_{c\in\mathcal{D}} p_i^c$ is the aggregate probability that pixel i belongs to potentially dynamic classes, τ is the central modulation threshold, β is the slope coefficient, and $\sigma(\cdot)$ is the logistic function. The physical meaning of $\alpha_i \in (0,1)$ is the degree to which pixel i is allowed to contribute to motion evidence computation. When the dynamic-class probability is far above τ, α_i approaches 0, indicating that the pixel's motion evidence should be treated cautiously. When that probability is far below τ, α_i approaches 1, allowing full participation. In ambiguous boundary regions, α_i takes an intermediate value, preserving the evidence contribution while reducing its weight.

Differentiability and monotonicity are the key properties of this design: α_i varies continuously with the dynamic-class probability, avoiding the abrupt steps of hard masks at segmentation boundaries. For a LiDAR point cloud, point j is projected onto the image plane, and the mean of α_i within the projection neighborhood is used as its modulation coefficient α_j^{lidar}. Points projected outside the image are assigned coefficients using a moving average of their historical semantic observations.

2.2 Motion consistency measurement in the visual stream

Under semantic modulation, visual motion evidence is formed from residuals between optical flow and camera ego-motion compensated flow. For a feature whose modulation coefficient α_i exceeds the participation threshold, let its observed optical flow be f_i and its predicted ego-motion compensated flow, computed from initial adjacent-frame camera poses and depth priors, be $\hat{f}_i$. Define the residual as $r_i = \| f_i - \hat{f}_i \|$. Applying a fixed threshold directly to r_i causes systematic errors near semantic boundaries: depth discontinuities around segmentation boundaries increase prediction errors, and a fixed threshold mistakes this geometric uncertainty for motion [11],[12]. We therefore adaptively scale the decision threshold with the modulation coefficient:

$$\theta_i = \theta_0 \cdot \left(1 + \gamma(1 - \alpha_i)\right) \tag{2}$$

Here, θ_0 is the baseline threshold and γ is a relaxation coefficient. When α_i approaches 1, indicating a confidently static semantic label, $\theta_i \approx \theta_0$ and the criterion is strict. When α_i approaches

0, indicating a confidently dynamic semantic label, $\theta_i \approx \theta_0(1+\gamma)$ and the criterion is relaxed. The direction of this design is important: a relaxed threshold in semantically dynamic regions tolerates larger residuals without readily declaring motion, allowing stationary objects in dynamic semantic classes to be retained. Rejection is ultimately determined by the subsequent arbitration mechanism, not by a single-frame residual.

Residual decisions use adaptive segmentation based on the local residual distribution rather than a fixed global threshold. Within an image window centered on the feature, we compute the residual median and absolute deviation. The median plus three times the absolute deviation defines a local decision bound, which is used jointly with θ_i. A point is marked as visual motion evidence only when its residual exceeds both the local bound and the modulated threshold. Its evidence strength is defined as the normalized exceedance:

$$e_i^{\text{vis}} = \min\left(1, \frac{r_i - \max\left(\theta_i, \theta_i^{\text{local}}\right)}{\theta_i}\right) \tag{3}$$

$e_i^{\text{vis}} \in [0,1]$ provides a continuous output for the visual evidence stream, avoiding loss of information from binarization.

2.3 Point-level motion probability estimation in the LiDAR stream

Motion detection in LiDAR point clouds faces both sparsity and occlusion. A single point cloud does not supply point correspondences; clustering methods tend to fragment sparse regions, while object tracking introduces additional front-end dependencies [13],[14]. We design a point-level motion probability model based on multi-frame occupancy consistency. It relies on neither clustering nor tracking but directly measures inter-frame occupancy changes around each LiDAR point.

For LiDAR point j, denoted by x_j, define a spherical neighborhood $\mathcal{B}_j$ centered on the point with radius R_j. The radius is inversely scaled by the semantic modulation coefficient:

$$R_j = R_0 \cdot \left(1 + \kappa \alpha_j^{\text{lidar}}\right) \tag{4}$$

Here, R_0 is the baseline radius and κ is a scaling coefficient. Semantically dynamic regions, where α_j^{lidar} is small, receive smaller radii to reduce computation and increase local sensitivity. Semantically static regions, where α_j^{lidar} is large, receive larger radii to capture genuinely moving objects that semantic segmentation may have misclassified as static. This proactively addresses mode (c) conflicts.

Within the spherical neighborhood, point clouds from the most recent K frames are projected into the current frame's coordinate system to obtain the occupancy state sequences of the enclosed voxels. Let $o_v^k \in \{0,1\}$ indicate the occupancy of voxel v at frame k. The raw motion probability estimate for the point is the weighted mean occupancy flip frequency of neighboring voxels:

$$p_j^{\text{raw}} = \frac{1}{|\mathcal{V}_j|} \sum_{v \in \mathcal{V}_j} w_v \cdot \frac{1}{K-1} \sum_{k=2}^{K} \mathbb{1}\left[o_v^k \neq o_v^{k-1}\right] \tag{5}$$

Here, $\mathcal{V}_j$ is the set of voxels covered by the spherical neighborhood and w_v is the voxel weight, obtained by logarithmic normalization of the voxel's point count to suppress dominance by dense voxels. Occupancy flip frequency describes whether a spatial location repeatedly appears and disappears in

recent observations. Static structures remain occupied and therefore have low flip frequencies, whereas moving objects alternate occupancy states as their position's change [15].

Two biases in the raw estimate require correction: false flips caused by occlusion, when foreground objects temporarily hide background voxels, and missed detections caused by sparsity, when a moving object is not sampled in some frames. A visibility threshold is introduced: a voxel contributes to the statistics only if it has been observed at least $K_{\min}$ times within K frames, and occluded frames are excluded from the denominator [16]. The corrected motion probability is mapped to $[0,1]$ by temperature scaling:

$$p_j^{\text{lidar}} = \frac{1}{1+\exp\left(-a\cdot\left(p_j^{\text{raw}}-b\right)\right)} \tag{6}$$

a, b are calibration parameters that fit using static scene sequences without dynamic objects, so that p_j^{lidar} for static points is concentrated at low values.

2.4 Spatiotemporal alignment and confidence calibration of dual-stream evidence

Visual evidence is defined at the pixel level and LiDAR evidence at the point level. Their resolutions, viewpoints, and noise characteristics differ, so direct fusion can allow one stream to dominate. We achieve spatiotemporal alignment through bidirectional projection and aggregation. In the forward direction, LiDAR point x_j is projected onto the image plane. The modulation-weighted mean visual evidence within the 5×5 neighborhood of its projection $\pi(x_j)$ becomes the point's visual evidence $e_j^{\text{vis}\to\text{lidar}}$. In the reverse direction, image features are back projected into three-dimensional space using depth priors. The mean LiDAR evidence within a neighborhood whose radius equals the mean LiDAR point spacing becomes the feature's LiDAR evidence $e_i^{\text{lidar}\to\text{vis}}$ [17].

After alignment, reliability-weighted log-odds fusion replaces simple averaging:

$$L_f = \lambda_{\text{vis}}\cdot\log\frac{e^{\text{vis}}}{1-e^{\text{vis}}} + \lambda_{\text{lidar}}\cdot\log\frac{e^{\text{lidar}}}{1-e^{\text{lidar}}} \tag{7}$$

Here, L_f denotes the fused log odds and $\lambda_{\text{vis}}, \lambda_{\text{lidar}}$ denote stream reliability weights, determined by normalizing the inverse evidence variances measured in static reference scenes. A stream with lower variance receives a higher weight. Sigmoid map fused evidence to cross-modal motion confidence e_f in $[0,1]$. Its uncertainty σ_f, defined using disagreement between the streams, is output simultaneously:

$$\sigma_f = |e^{\text{vis}} - e^{\text{lidar}}|\cdot\left(1-|\,2e_f-1\,|\right) \tag{8}$$

Uncertainty is jointly determined by inter-stream disagreement and the distance of confidence from 0.5. It is low when both streams agree and are far from 0.5, and high when they disagree strongly or both lie near the decision boundary. This uncertainty is a core input to the arbitration mechanism in Section 3. High-uncertainty samples trigger delayed decisions and additional evidence accumulation rather than immediate decisions.

3. SEMANTIC-MOTION CONFLICT ARBITRATION AND DYNAMIC OBJECT DECISIONS

3.1 Formalization and classification of conflict modes

For instance, k, denote the semantic prior's dynamic tendency by $s_k \in [0,1]$, computed as the mean dynamic-class probability of its pixels; a high s_k indicates semantic membership in a potentially dynamic class. Denote cross-modal motion evidence by $m_k \in [0,1]$, computed as the spatially weighted mean of e_f within the instance. Using the neutral threshold τ_s, τ_m for each of s_k and m_k, instances can be divided into four conflict modes [18]:

$$C_k = \begin{cases} \text{(a) genuinely moving object,} & s_k \geq \tau_s, m_k \geq \tau_m \\ \text{(b) stationary semantically dynamic object,} & s_k \geq \tau_s, m_k < \tau_m \\ \text{(c) semantically missed moving object,} & s_k < \tau_s, m_k \geq \tau_m \\ \text{(d) reliable static background,} & s_k < \tau_s, m_k < \tau_m \end{cases} \tag{9}$$

In modes (a) and (d), semantic and motion evidence agree, making decisions relatively straightforward. Modes (b) and (c) involve contradictory evidence and are major sources of error in existing methods. Hard filtering inevitably rejects stationary vehicles incorrectly in mode (b), because its decisions depend only on s_k. Purely geometric methods in mode (c) cannot constrain the decision region when segmentation misses an object and are sensitive to sparse point clouds. Our soft modulation in Section 2.1 preserves a motion evidence computation path for mode (b) instances. Dual-stream fusion in Section 2.4 allows mode (c) instances to obtain LiDAR motion signals that contradict semantics and yield high uncertainty σ_f. Both conflicts thus become detectable disagreements rather than invisible errors, making arbitration possible. This mode classification determines evidence processing strategies rather than final one-step labels: modes (a) and (d) enter the fast decision path, whereas modes (b) and (c) require delayed accumulation.

3.2 Evidence accumulation and delayed decisions within a temporal sliding window

For an instance k entering the delayed path, maintain its cross-modal motion confidence time series $\left\{e_f^{(t)}\right\}_{t=t_0}^{t_0+W-1}$. The window length W depends on the mode: 8 for mode (b) and 6 for mode (c). Decisions use two window-level statistics rather than single-frame values: mean confidence μ_k and stability ν_k. Stability is defined as the normalized lag-one autocovariance of confidence within the window:

$$\mu_k = \frac{1}{W}\sum_t e_f^{(t)},\ \nu_k = \frac{1}{W-1}\sum_t \left(e_f^{(t)} - \mu_k\right)\left(e_f^{(t-1)} - \mu_k\right) \tag{10}$$

Introducing stability ν_k is central to this section. An instance with a high mean but low ν_k, indicating strong inter-frame oscillation, is more likely to exhibit false motion caused by transient segmentation errors or point cloud occlusion [19]. An instance with a moderate mean but high ν_k instead exhibits a persistent, consistent motion trend. The rejection condition therefore combines thresholds on mean confidence and stability:

$$\text{Reject} \iff \mu_k \geq \tau_\mu \wedge \nu_k \geq \tau_\nu \tag{11}$$

Here, τ_μ is the mean threshold, set to 0.62 for mode (b) and 0.55 for mode (c), with the lower value compensating for missing semantic priors. The stability threshold τ_ν is fixed at 0.45. Stability measures whether motion evidence is temporally coherent. A stationary vehicle may occasionally receive high single-frame confidence when another vehicle passes nearby, but its ν_k remains close to

zero, preventing rejection. A genuinely moving object maintains high ν_k through persistent displacement.

The upper decision delay bound is a real-time constraint that must be addressed. Before the window fills, if an instance's instantaneous confidence exceeds the emergency threshold τ_{emg}, set to 0.85, for N_{fast} consecutive frames, accumulation is bypassed and a decision is issued immediately. This handles sudden motion onset or high-speed cut-ins. Emergency, and regular paths share the same output interface, so downstream optimization does not distinguish the decision source.

3.3 Spatial constraints using semantic instances as arbitration units

Instance-level decisions require the spatial distribution of motion evidence within each instance to participate in the decision. Let instance k contain n_k evidence carriers, either pixels or LiDAR points, with evidence values $\left\{e_f^{(j)}\right\}_{j=1}^{n_k}$ and positions $\{u_j\}$. Define spatial concentration as the local consistency of evidence values within spatial neighborhoods:

$$\varrho_k = \frac{1}{n_k} \sum_j \frac{1}{|\mathcal{N}_j|} \sum_{l \in \mathcal{N}_j} \mathbb{1}\left[\left|e_f^{(j)} - e_f^{(l)}\right| < \delta\right] \tag{12}$$

Here, $\mathcal{N}_j$ is the spatial neighbor set of evidence carrier j, using pixel neighborhoods for images and point-distance neighborhoods for LiDAR, and δ is a consistency tolerance. A value of ϱ_k close to 1 indicates spatially continuous motion evidence; a value close to 0 indicates fragmented evidence. The arbitration rule is modified accordingly. When ϱ_k falls below the concentration threshold τ_ϱ, set to 0.55, the instance is classified as noise-dominated and retained even if μ_k exceeds the mean threshold. Conversely, when ϱ_k is high but μ_k is only moderate, between 0.45 and τ_μ, spatial concentration itself provides supplementary evidence that can trigger rejection.

This spatial constraint targets two degeneracies. First, ambiguous segmentation boundaries can introduce background pixels along instance edges. Strong motion evidence from these isolated pixels lowers ϱ_k, preventing false rejection. Second, a sparse LiDAR cloud may sample a moving object with only a few points. In this case, ϱ_k can remain high because those points are spatially concentrated, preventing missed motion detections. Thus, concentration separates evidence quantity from evidence spatial structure and reduces arbitration sensitivity to sampling density.

3.4 Dynamic probability masks and static constraint recovery from arbitration outputs

The final arbitration output is a dynamic probability $p_k^{\text{dyn}} \in [0,1]$ for each semantic instance, jointly mapped from the window mean, stability, and spatial concentration, rather than a binary mask:

$$p_k^{\text{dyn}} = \sigma(w_1\mu_k + w_2\nu_k + w_3\varrho_k - b_0) \tag{13}$$

Here, w_1, w_2, w_3 are fusion weights are calibrated according to the discriminative contributions of the three quantities on a validation set, with values 0.5, 0.3, and 0.2, respectively, and b_0 is a bias. The probability serves subsequent optimization through three tiers. Features of instances satisfying $p_k^{\text{dyn}} \geq 0.75$ contribute to residual construction with weight $1 - p_k^{\text{dyn}}$, reducing their influence rather than discarding them. Instances satisfying $0.35 < p_k^{\text{dyn}} < 0.75$ enter a pending buffer, where their

features temporarily participate with moderate weights while evidence continues accumulating in subsequent frames. Instances satisfying $p_k^{\mathrm{dyn}} \leq 0.35$ receive full weight.

The three-tier design primarily maintains odometry observability. In extremely dynamic scenes, such as dense pedestrian traffic, discarding all high-probability instances can abruptly leave too few static constraints for optimization and make pose estimation underdetermined. Down weighting retains gradient paths for high-probability instances with very small weights, suppressing contamination while preserving the rank of the constraint matrix. The pending buffer allows later frames to revise decisions for intermediate-probability instances, avoiding irreversible errors from instantaneous front-end decisions.

4. REDESIGN OF THE DYNAMIC-AWARE TIGHTLY COUPLED OPTIMIZATION ARCHITECTURE

4.1 Dynamic probability modulation of visual reprojection residuals

Let visual feature i be observed in keyframe l, with reprojection residual r_i^{vis} and information matrix Ω_i, typically the inverse observation noise covariance. Traditional dynamic SLAM omits residual construction for dynamic features, which is equivalent to setting their information matrices to zero [20],[21],[22]. We instead retain the residual terms and modulate their information matrices by the complement of the dynamic probability:

$$\tilde{\Omega}_i = \left(1 - p_{k(i)}^{\mathrm{dyn}}\right)^{\eta} \cdot \Omega_i \tag{14}$$

Here, $p_{k(i)}^{\mathrm{dyn}}$ is the dynamic probability of semantic instance $k(i)$ containing feature i, and η is a modulation sharpness exponent set to 1.5. The exponent $\eta > 1$ compresses the information matrices of high-dynamic-probability features more rapidly. When $p = 0.9$, the complement is 0.1; modulation by $\eta = 1.5$ reduces the weight to approximately 0.032, approaching rejection. When $p = 0.5$, the complement is 0.5 and the modulated weight is approximately 0.354, preserving a substantial contribution.

A key property of this modulation is gradient path preservation. As the information matrix approaches zero, the residual's Jacobian contribution to pose estimation also approaches zero, but the residual remains in the graph, and its gradient is not hard-truncated. When back-end pose updates significantly change the reprojection error of a high-probability feature, that change can still feed back into optimization through its small weight, providing indirect evidence of whether the feature is actually moving. Motion decisions thus become soft constraints continually tested during optimization rather than fixed decisions made beforehand. Table 1 presents how the modulation sharpness exponent effects information matrix compression.

Table 1. Sharpness settings and weight distributions for visual residual information matrix modulation

Dynamic probability p	$\eta = 1.0$ Weight	$\eta = 1.5$ Weight	$\eta = 2.0$ Weight	Equivalent treatment
0.20	0.80	0.72	0.64	Full participation
0.50	0.50	0.35	0.25	Moderate participation
0.75	0.25	0.13	0.06	Weak participation
0.90	0.10	0.03	0.01	Approximate rejection

Table 1 shows that $\eta = 1.5$ retains a weight of 0.35 at $p = 0.5$ while reducing it to 0.03 at $p = 0.9$, balancing continued participation of intermediate-probability instances against effective suppression of high-probability instances. With $\eta = 1.0$, compression at high probabilities is insufficient, since 0.10 may still introduce substantial contamination. With $\eta = 2.0$, the weight decays too early at $p = 0.5$, reaching 0.25 and limiting refinement through optimization. The choice of $\eta = 1.5$ balances these effects.

4.2 Dynamic probability weighting for LiDAR point cloud registration

The point-to-plane residual r_j^{lidar} and point-to-line residuals in LiDAR odometry also require dynamic probability weighting [23],[24]. However, evidence carriers are points rather than visual features, and point-level motion probabilities p_j^{lidar} have a more spatially continuous distribution. Direct pointwise weighting can assign inconsistent weights within the same rigid object, violating the rigid-body consistency assumption of registration [25]. We therefore use voxel-level aggregation. Let voxel v contain point set $\mathcal{P}_v$; its dynamic probability is the maximum point-level probability:

$$p_v^{\text{dyn}} = \max_{j \in \mathcal{P}_v} p_j^{\text{lidar}} \tag{15}$$

The maximum is used instead of the mean as a conservative choice. If a voxel contains a point with a high motion probability, the entire voxel should be treated cautiously; averaging can dilute it with many static points and underestimate motion risk. However, a single outlier may dominate the maximum, so an occupancy density threshold is introduced. The maximum rule applies only when the voxel point count $|\mathcal{P}_v|$ is at least τ_{occ}, set to 4. Otherwise, the voxel uses the mean to suppress isolated sparse outliers. The voxel weight then modulates the residual information matrix [26],[27]:

$$\tilde{\Omega}_v^{\text{lidar}} = \left(1 - p_v^{\text{dyn}}\right) \cdot \Omega_v^{\text{lidar}} \tag{16}$$

Here, Ω_v^{lidar} is the baseline information matrix shared by all point residuals in the voxel. We do not apply the visual sharpness exponent, because LiDAR point-level probabilities have already undergone temperature scaling and have relatively sharp distributions. Further exponentiation would prematurely suppress intermediate-probability voxels. Table 2 compares weight consistency under pointwise weighting, voxel averaging, and voxel maximum aggregation.

Table 2. Comparison of consistency characteristics for LiDAR residual weight allocation strategies

Allocation strategy	Intra-object weight variance	Sensitivity to point outliers	Robustness in sparse voxels	Intermediate-probability retention
Pointwise weighting	0.087	High	Low	0.48
Voxel mean	0.012	Low	Moderate	0.61
Voxel maximum (proposed)	0.019	Moderate	High	0.44

Table 2 shows that pointwise weighting achieves the highest retention, 0.48, but produces an intra-object weight variance of 0.087, undermining registration consistency. Voxel averaging has the lowest variance, 0.012, yet is overly sensitive to individual point anomalies and insufficiently robust in sparse voxels. The voxel maximum balances variance, 0.019, and robustness. Its slightly lower intermediate-probability retention, 0.44, can be compensated through the occupancy density threshold. This trade-off shows that LiDAR weighting aims to suppress contamination from moving points while maintaining rigid-body consistency, rather than simply maximizing retention.

4.3 Pose graph construction with joint semantic–motion constraints

The preceding sections modify sensor observation factors. This section introduces a factor absent from existing paradigms: a semantic instance-level motion consistency factor. Suppose semantic instance k is associated between keyframes t_a and t_b through instance feature matching or LiDAR cluster association. Denote its observed inter-frame relative pose by $T_k^{a\to b}$. Define the motion consistency residual in Lie algebra form as the difference between the instance relative pose and the platform relative pose:

$$r_k^{\mathrm{mo}} = \log\left(\left(T_{a\to b}^{\mathrm{body}}\right)^{-1} T_k^{a\to b}\right)^{\vee} \tag{17}$$

$T_{a\to b}^{\mathrm{body}}$ is the platform relative pose from t_a to t_b, and $\log(\cdot)^{\vee}$ is the SE (3) logarithmic map to a six-dimensional vector. For a static instance, r_k^{mo} should be close to zero; for a moving instance, the residual may be arbitrarily large. Instead of imposing a strong constraint on this residual, we modulate its information matrix using dynamic probability:

$$\tilde{\Omega}_k^{\mathrm{mo}} = \left(1 - p_k^{\mathrm{dyn}}\right) \cdot \Omega_0^{\mathrm{mo}} \tag{18}$$

Ω_0^{mo} is the baseline information matrix under the static assumption, calibrated from instance association accuracy. When p_k^{dyn} is low, this factor imposes a strong constraint, forcing consistency between platform and instance relative poses and effectively adding a static semantic instance as an extra pose observation. When p_k^{dyn} is high, the constraint is relaxed, allowing large instance displacement relative to the platform without penalty.

This factor changes motion decisions from a preprocessing step into a constraint relationship among optimization variables. Dynamic probability not only determines feature participation but directly configures pose graph edge weights. The motion decision for instance and platform pose estimation are therefore solved jointly within one optimization problem. If the front end incorrectly assigns a low dynamic probability to an actually moving instance, its motion consistency residual grows during optimization, and weighted information matrix feedback prompts the back end to revise that probability. Table 3 presents the factor's constraint strengths at different dynamic probabilities and their effects on pose graph connectivity.

Table 3. Constraint strengths and pose graph connectivity for semantic–motion consistency factors

Dynamic probability p	Factor weight	Constraint type	Added edges per instance	Graph connectivity gain
0.10	0.90	Strong static constraint	2	0.14
0.35	0.65	Moderate constraint	2	0.09
0.60	0.40	Weak constraint	1	0.04
0.85	0.15	Almost unconstrained	1	0.01

Table 3 shows that low-dynamic-probability instances provide a graph connectivity gain of up to 0.14, indicating that static semantic instances used as additional pose observations substantially increase pose graph redundancy. As dynamic probability rises, both the number of added edges and the connectivity decreases. High-probability instances approach isolated nodes contributing almost no constraints. This graded design adapts the graph's connectivity to scene dynamics rather than allowing global degradation in highly dynamic scenes.

4.4 Retaining dynamic information during marginalization

During sliding-window marginalization, states and features removed from the window are typically converted into prior factors, compressing their information into linear constraints on the remaining variables. For high-dynamic-probability instances, we prioritize marginalization while retaining probability priors. Their states and features are removed first to avoid contaminating subsequent optimization, but their dynamic probabilities remain as instance-level priors:

$$p_k^{\text{prior}} = \rho \cdot p_k^{\text{dyn,window}} + (1-\rho) \cdot p_k^{\text{prior,old}} \tag{19}$$

Here, $p_k^{\text{dyn, window}}$ is the instance's mean dynamic probability within the window, $p_k^{\text{prior, old}}$ is the prior retained from the previous round, and ρ is a forgetting factor set to 0.7 to favor current observations. When the instance reappears in subsequent frames, this prior directly initializes its dynamic probability instead of requiring evidence accumulation from scratch. Initial uncertainty depends on the number of prior accumulation rounds.

This strategy addresses a specific degeneracy: a dynamic object reenters the field of view after several frames of occlusion. Without a probability prior, the reappearance window contains no evidence, and the system must repeat the entire accumulation cycle, requiring 8 frames for mode (b). During this interval, the object's features participate at full weight and introduce brief but substantial contamination. Retaining the prior provides a dynamic probability near its historical level in the first reappearance frame, reducing renewed accumulation to 2–3 frames. Table 4 presents response characteristics for different occlusion durations.

Table 4. Responses to post-occlusion reappearance with retained dynamic probability priors

Occlusion duration (frames)	Reaccumulating without prior (frames)	Reaccumulating with prior (frames)	Initial prior	Prior uncertainty	Contamination exposure (frames)
3	8	2	0.71	0.09	2

Occlusion duration (frames)	Reaccumulating without prior (frames)	Reaccumulating with prior (frames)	Initial prior	Prior uncertainty	Contamination exposure (frames)
6	8	3	0.58	0.16	3
12	8	3	0.41	0.24	3
20	8	4	0.29	0.31	4

Table 4 shows that, as occlusion duration increases, the initial prior decreases from 0.71 to 0.29 and its uncertainty rises from 0.09 to 0.31, reflecting the declining relevance of historical evidence after longer occlusions. Even after a 20-frame occlusion, renewed accumulation with a prior requires only 4 frames rather than 8 without one, reducing contamination exposure from 8 frames to 4. The forgetting factor $\rho = 0.7$ allows the prior to decay naturally with occlusion duration instead of remaining frozen, preventing outdated priors from overwhelming new observations. Together with the weighting mechanisms in Sections 4.1–4.3, this strategy completes the dynamic-aware optimization architecture: front-end arbitration outputs continuous probabilities; the back end modulates visual and LiDAR observation factors, constructs semantic–motion consistency factors, and maintains decision continuity across windows through marginalization priors.

5. EXPERIMENTAL VALIDATION AND ABLATION ANALYSIS

5.1 Experimental setup and rationale for baseline selection

The evaluation comprises Semantic KITTI, KITTI raw, and Urban Nav [28]. They are selected for complementary coverage of conflict modes rather than dataset size. Semantic KITTI provides pointwise semantic annotations, enabling precise identification of stationary vehicle instances in mode (b). KITTI raw preserves transitions involving sudden object motion and post-occlusion reappearance, allowing the decision delay bound to be tested. Urban Nav supplies scenes with high dynamic density to stress-test constraint retention in optimization. Baselines are organized into three levels: fusion frameworks, LIO-SAM and LVI-SAM, provide external localization references; ablation controls, namely purely semantic hard-filtering and purely geometric motion detection, isolate the contributions of soft and semantic modulation; and a single-frame decision variant tests the necessity of temporal arbitration. All methods use the same semantic segmentation network weights and LiDAR preprocessing pipeline to exclude front-end differences as confounding factors.

To characterize transition behavior rather than average performance alone, results are reported using both the median and interquartile range (IQR). Transition intervals comprise F_t frames, set to 5, before and after dynamic motion onset. In addition to conventional absolute trajectory error, two metrics directly test this section's claims: the map dynamic-point residual rate ϵ_{res}, the proportion of map points classified as static that actually belong to moving objects; and the static semantic constraint retention rate ϵ_{keep}, the proportion of features on stationary objects in dynamic semantic classes correctly retained for optimization. Their joint distribution, rather than either individual value, forms the principal evidence. Table 5 summarizes key experimental parameters and the comparison dimensions of the baselines.

Table 5. Experimental settings and baseline comparison dimensions

Setting	Value	Comparison dimension	Number of baselines
Segmentation frame rate	10 Hz	Soft modulation vs hard-filtering	2
LiDAR channels	64	Dual-stream vs single-stream evidence	2
Arbitration window W	6 / 8	Temporal vs single-frame decisions	2
Transition interval F_t	5 frames	Optimization weighting vs front-end rejection	2
Repeated runs	3	External fusion framework references	2

Comparison dimensions are deliberately paired so that each baseline pair differs from the proposed method in only one design decision. Differences in the subsequent ablation chain can therefore be attributed to that decision rather than interacting with configuration changes. The semantic segmentation frame rate is fixed at 10 Hz to align with the LiDAR rate and prevent segmentation latency from being mistaken for arbitration delay.

5.2 Joint evaluation of localization accuracy and map cleanliness

In Figure 1, the horizontal axis is the map dynamic-point residual rate ϵ_{res} and the vertical axis is median absolute trajectory error. Each method is represented by scatter points and an IQR ellipse, with the transition-interval subset overlaid using hollow markers.

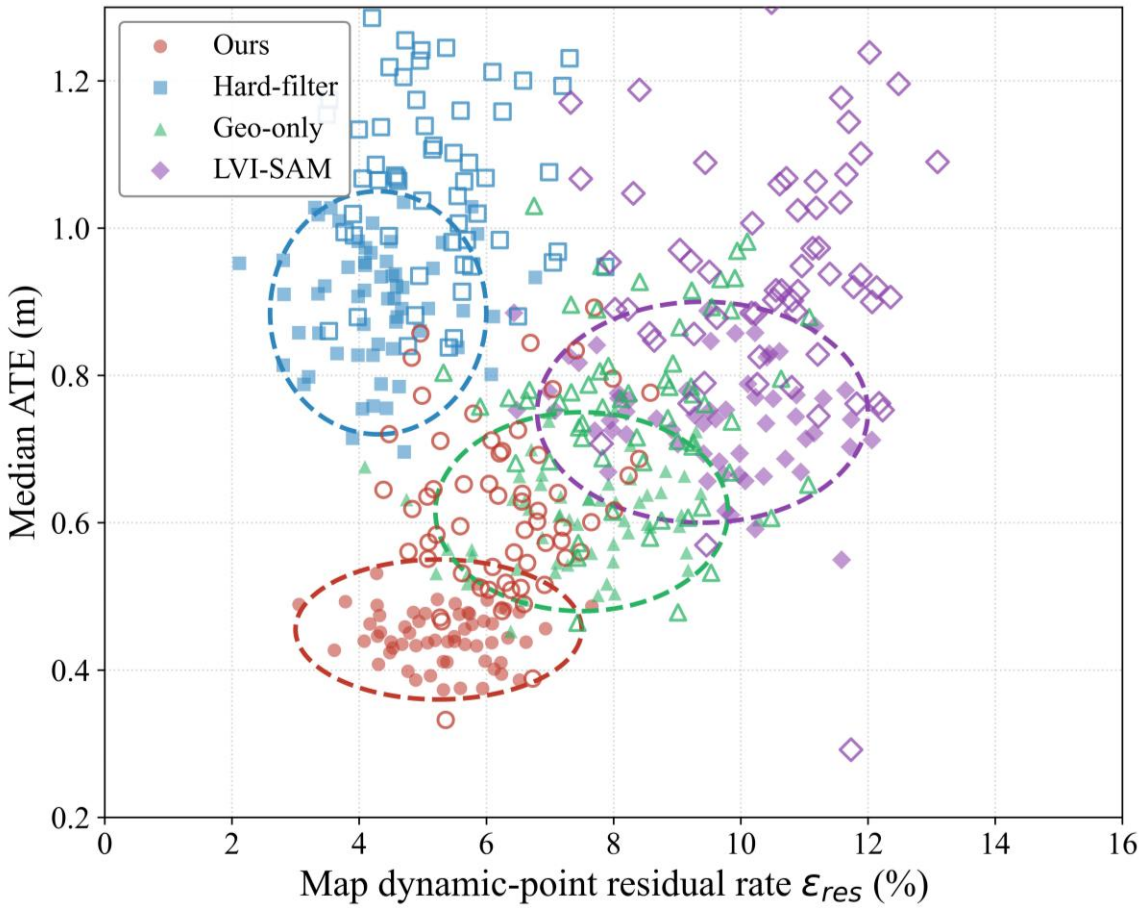


Figure 1. Joint distribution of trajectory error and map dynamic-point residual rate including transition subsets

Figure 1 exposes trade-offs that a single metric cannot reveal. The hard-filtering variant lies in the low residual-rate region but substantially higher on the error axis, indicating that it achieves map cleanliness by sacrificing constraints. The purely geometric variant shows the opposite pattern, with a higher residual rate and moderate error. The proposed method occupies the lower-left region and performs better in both dimensions. Its IQR ellipse also extends less in both directions than those of the two controls, indicating greater stability. The transition subset is especially diagnostic. During sudden object motion onset, the hard-filtering error distribution expands markedly toward the upper right, whereas the proposed method's transition points nearly overlap its overall distribution. This can be explained by the decision mechanisms. Hard-filtering rejects objects directly according to their semantic categories; because an object's label does not change at motion onset, the system does not respond until geometric residuals accumulate to a threshold. The proposed emergency path issues an early decision within 3 frames, and continuous optimization weighting suppresses transition-related error growth.

Figure 2 further shows error evolution over a representative sequence. Its horizontal axis is frame index, and its vertical axis is translation error within a sliding window. The three curves correspond to the proposed method, hard-filtering, and purely geometric detection. A shaded region marks the manually annotated onset of dynamic object motion.

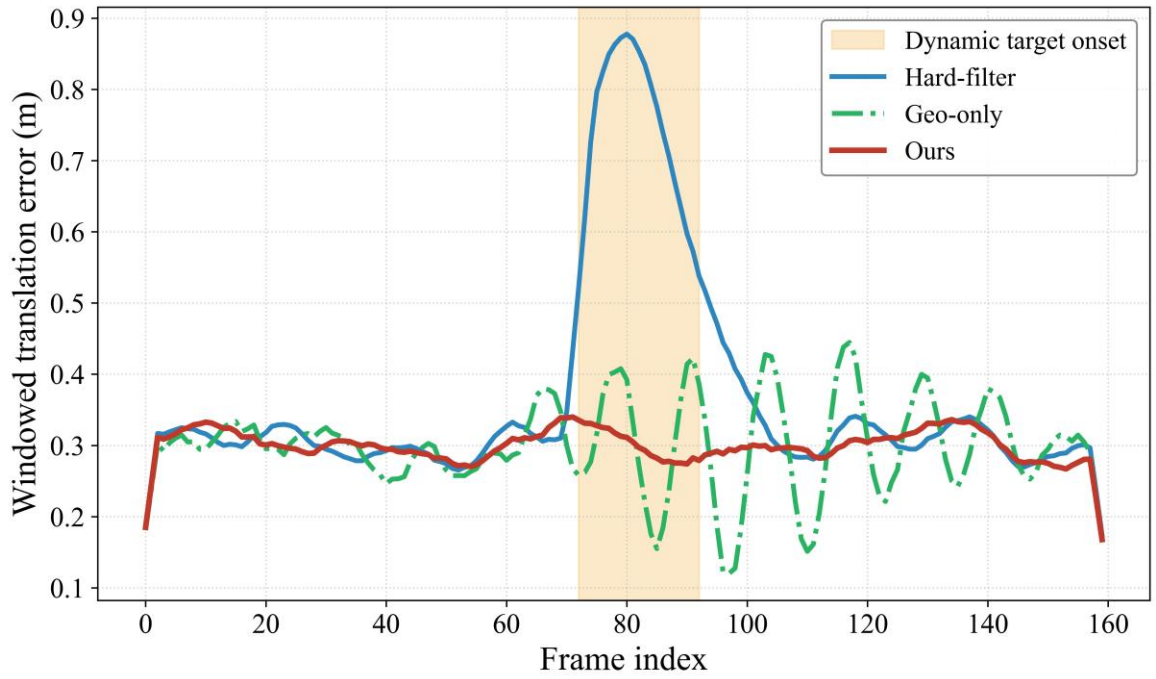


Figure 2. Translation error evolution during transitions in a representative sequence

In Figure 2, the hard-filtering curve rises abruptly after the shaded interval and takes many frames to recover. The purely geometric curve exhibits high-frequency fluctuations within the interval, while the proposed method has the smallest fluctuation amplitude. These shapes directly reflect the three decision mechanisms. The hard-filtering step arises because semantic labels lag actual motion;

geometric fluctuations arise from unstable inter-frame evidence under sparse LiDAR sampling; and the smoother proposed curve results from jointly constraining both degradations through bounded decision delay and optimization weighting. Figure 1 and Figure 2 provide complementary evidence concerning static-scene constraint retention and transition response speed.

5.3 Diagnostic experiments by conflict mode

Mode-specific diagnosis is the central test of the designs in Sections 2 and 3. Results are presented as decision accuracy distributions grouped by conflict mode. Figure 3 uses grouped box plots for modes (b) and (c), showing the proposed method, hard-filtering, purely geometric detection, and single-frame decisions side by side within each group.

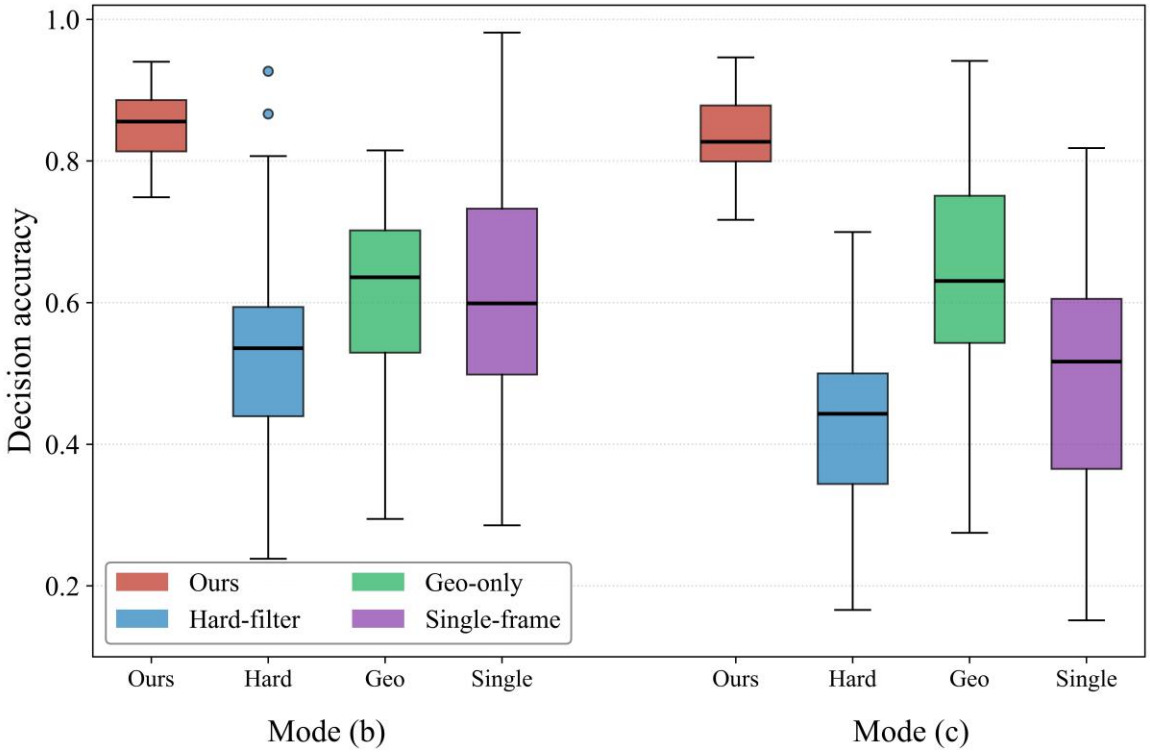


Figure 3. Grouped box plots of decision accuracy for conflict modes (b) and (c)

Figure 3 should be interpreted through box widths and outlier distributions as well as medians. For mode (b), the hard-filtering box shifts downward and has a long lower whisker, showing strong dependence on segmentation quality when judging stationary vehicles. Fluctuating confidence in segmentation causes many false rejections. For mode (c), the purely geometric box shifts upward but has numerous low-end outliers, indicating occasional correct decisions on semantically missed objects without reliable stability. The single-frame variant has broad boxes in both groups, supporting the claim that instantaneous noise dominates decisions without temporal accumulation. The proposed method has the narrowest boxes and highest medians in both groups, showing improved average accuracy and substantially lower decision variance.

Figure 4 presents the proposed method's outcomes for the two conflict modes as a confusion matrix heatmap. Rows represent ground-truth classes: stationary objects in dynamic semantic classes, genuinely moving objects, and moving objects missed by semantics. Columns represent arbitration

outputs: retained, pending, and rejected. Color intensity indicates sample proportion.

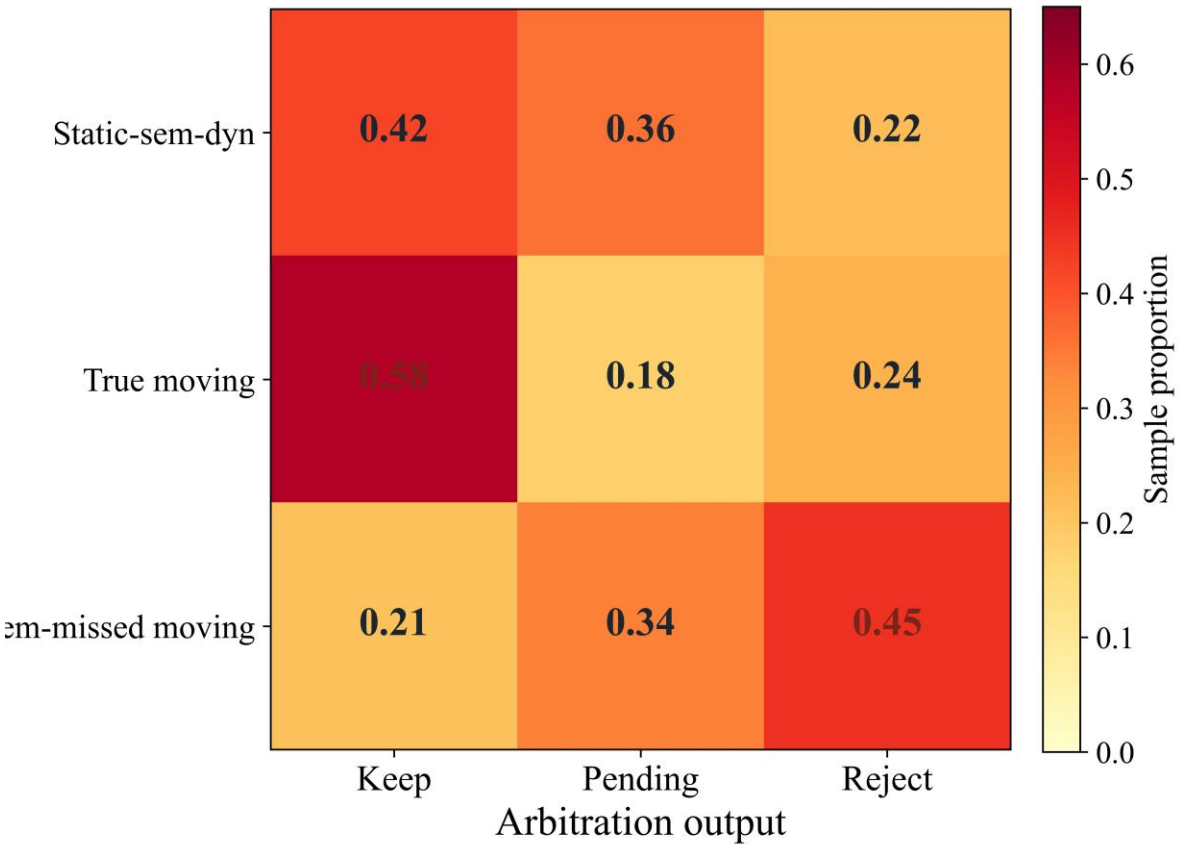


Figure 4. Confusion matrix of arbitration outputs for conflict modes using the proposed method

Figure 4 clarifies the role of the intermediate pending tier. A substantial proportion of mode (b) samples remain pending instead of being immediately retained, and some mode (c) samples remain pending instead of being immediately rejected. The mechanism therefore delays decisions when evidence is insufficient, matching the intended three-tier behavior in Section 3.4. Figure 5 reports the additional delay distribution introduced by arbitration. The horizontal axis is the number of frames from an object's first appearance to the final decision, the vertical axis is cumulative sample proportion, and the four curves represent four decision strategies.

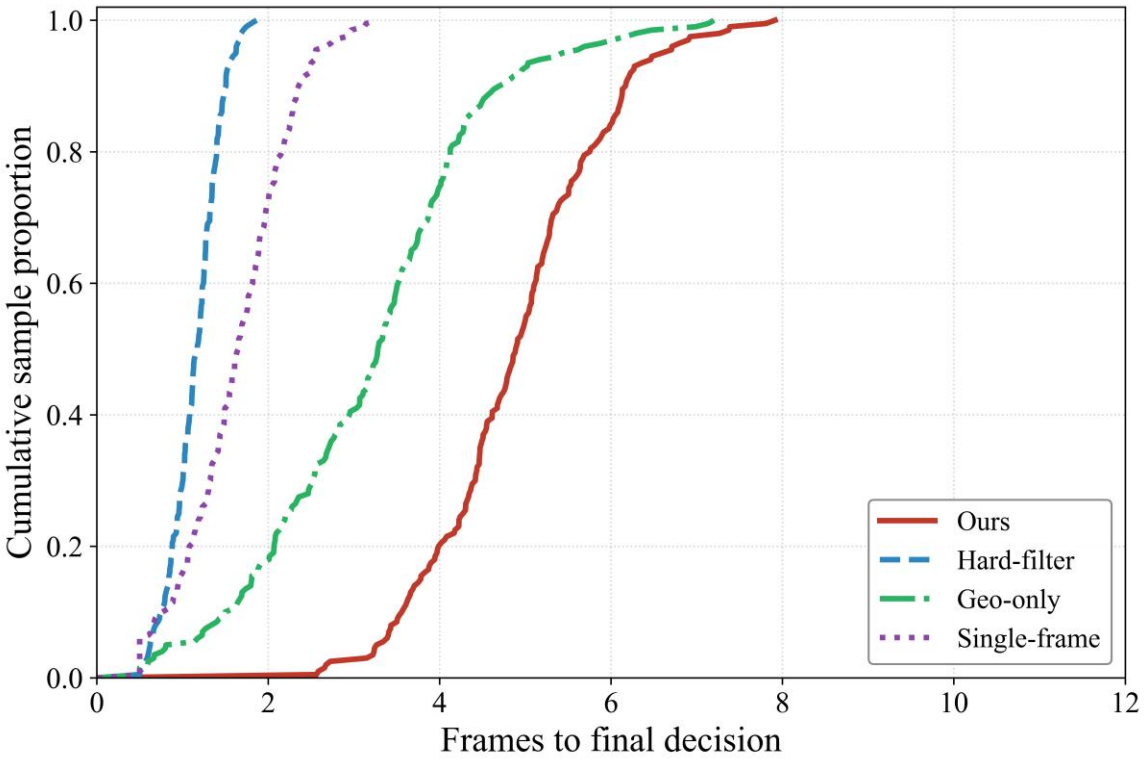


Figure 5. Cumulative distributions of decision delay under different decision strategies

In Figure 5, the proposed method's curve rises rapidly between 3 and 6 frames. The hard-filtering curve reaches a high cumulative proportion at 1 frame, but its accuracy cost is apparent in Figure 3. The single-frame variant has the lowest delay but unstable decisions, also shown in Figure 3. Together, Figure 3 and Figure 5 establish the delay-accuracy trade-off: low delay is undesirable if achieved at the cost of high variance. The position of the proposed curve indicates substantially lower decision variance within an acceptable delay range.

5.4 Ablation chain for individual design components

Ablation follows a dependency-ordered chain: removing semantic soft modulation, dual-stream evidence fusion, temporal arbitration, and weighting optimization. Each stage adds a removal to the previous stage, exposing the cumulative decline in marginal contributions. Figure 6 presents this chain as a waterfall plot. The horizontal axis is ablation level, and the vertical axis is performance retention relative to the complete method. Bar heights show the decrease attributable to each stage, while colored segments distinguish contributions from mode (b) errors, mode (c) errors, and trajectory error.

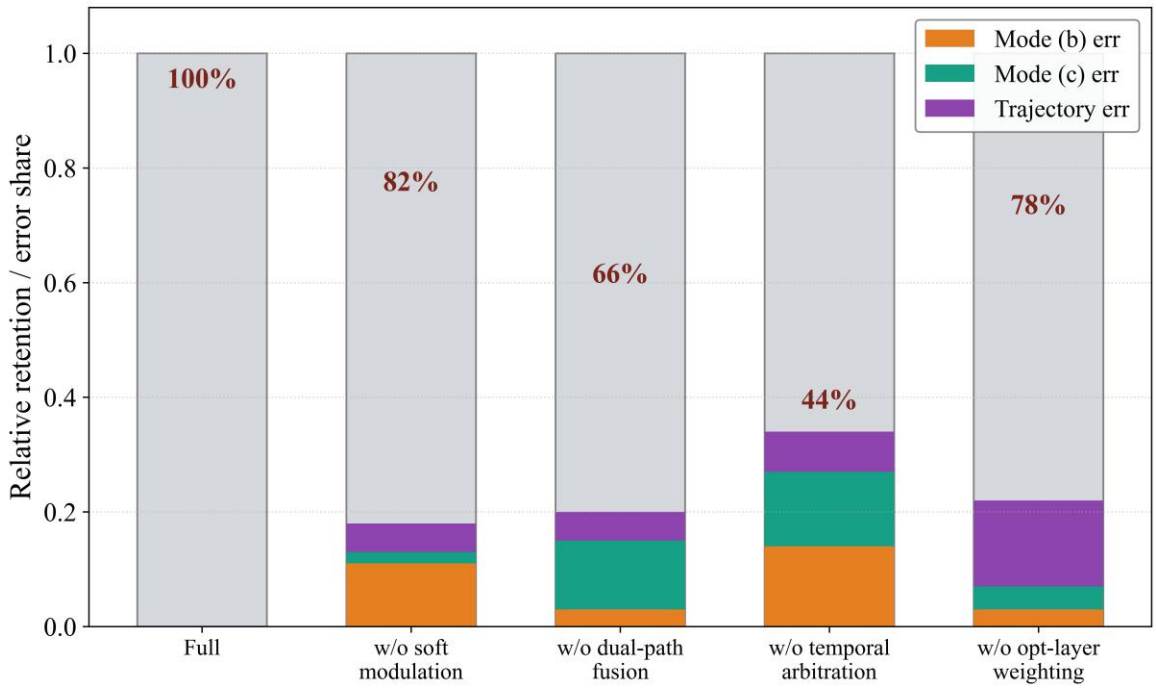


Figure 6. Waterfall plot of the ablation chain and decomposition of error sources

Figure 6 directly shows which conflict mode is most affected by removing each design. The soft-modulation removal stage is dominated by mode (b) errors because hard masks directly discard stationary vehicles. The dual-stream fusion removal stage is dominated by mode (c) errors because the LiDAR stream can no longer compensate for semantic misses. Removing temporal arbitration greatly expands both error components, indicating that single-frame decisions fail for both conflict types. Removing optimization weighting mainly increases the trajectory error component, with relatively small changes in conflict errors, showing that this design primarily supports constraint retention rather than motion decisions themselves.

Figure 7 complements this analysis with a radar chart comparing ablation variants in four dimensions: mode (b) accuracy, mode (c) accuracy, map cleanliness, and constraint retention.

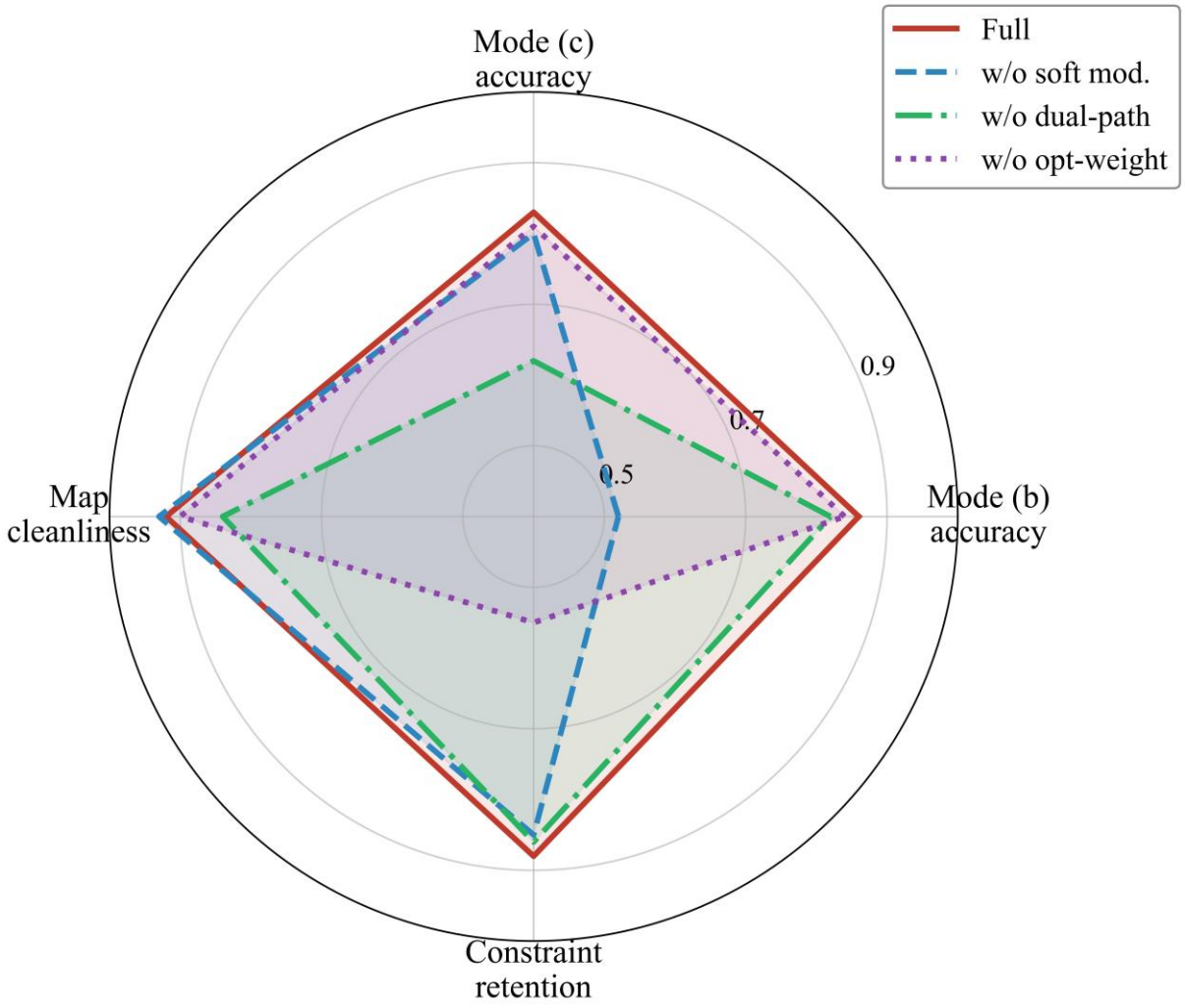


Figure 7. Radar chart of ablation variants in a four-dimensional evaluation space

In Figure 7, the complete method encloses the largest area across the four axes and has a balanced shape. Removing soft modulation creates an indentation along the mode (b) axis; removing dual-stream fusion creates one along mode (c); and removing optimization weighting creates one along constraint retention. The chart's geometry encodes the dimension affected by each design, allowing readers to identify component roles without comparing all numerical values. Figures 6 and 7 describe the same ablation results through cumulative degradation and functional dimensions, respectively, providing complementary perspectives.

5.5 Computational cost and real-time operating boundaries

Real-time analysis presents a deployable operating region through processing-time breakdowns and boundary conditions rather than a binary conclusion. Figure 8 is a stacked bar chart of module processing times. The horizontal axis lists semantic segmentation, visual evidence, LiDAR evidence, evidence fusion, arbitration, and optimization. The vertical axis is per-frame processing time, and bar segments distinguish between CPU and GPU time.

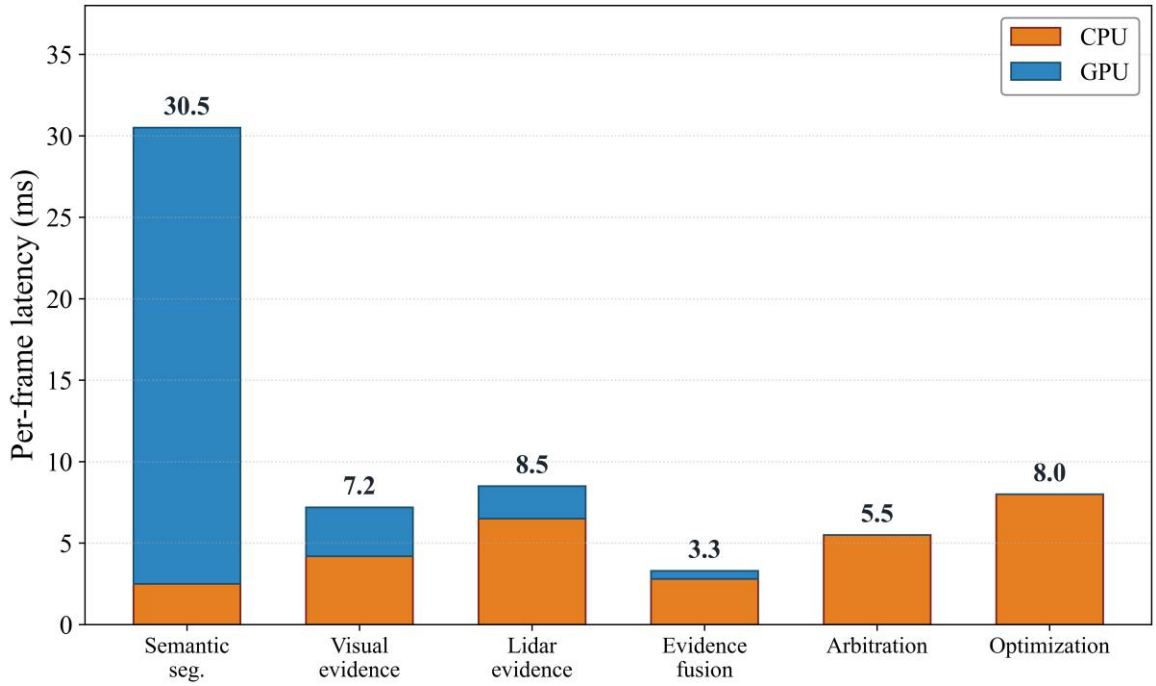


Figure 8. CPU and GPU breakdown of per-frame processing time by module

Figure 8 shows that semantic segmentation dominates GPU time. CPU time for arbitration and evidence fusion grows linearly with the number of dynamic instances but starts from a low baseline. LiDAR evidence computation depends on point cloud density through voxel aggregation. This breakdown identifies the external semantic segmentation dependency, rather than the newly introduced arbitration stage, as the real-time bottleneck. That finding determines the independent variables for the boundary surface in Figure 9. Its horizontal axis is dynamic object density, measured in instances per frame; its vertical axis is semantic segmentation frame rate; color intensity indicates whether the system meets the frame rate requirement; and a contour marks the boundary for 10 Hz processing.

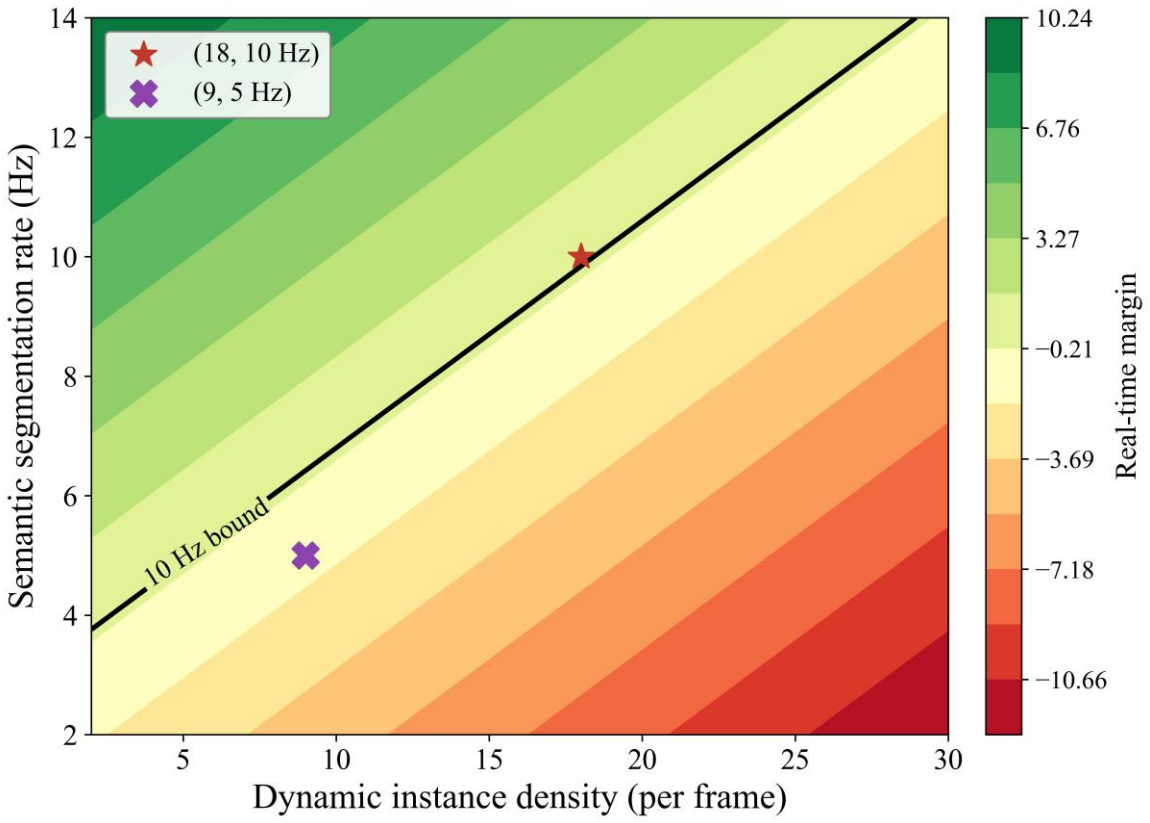


Figure 9. Real-time operating boundary over dynamic object density and semantic segmentation frame rate

Figure 9 shows that the system remains real-time at a segmentation rate of 10 Hz when dynamic density does not exceed 18 instances per frame. At 5 Hz, the boundary contracts to 9 instances per frame, indicating that segmentation inference frequency is the dominant real-time constraint. Figure 10 separately plots the relationship between arbitration window length W and decision delay. The horizontal axis is W; the two vertical axes show mean decision frames and emergency path activation ratio. The intersection of the dual-axis curves identifies the range from $W = 6$ to $W = 8$ as a compromise between delay and accuracy.

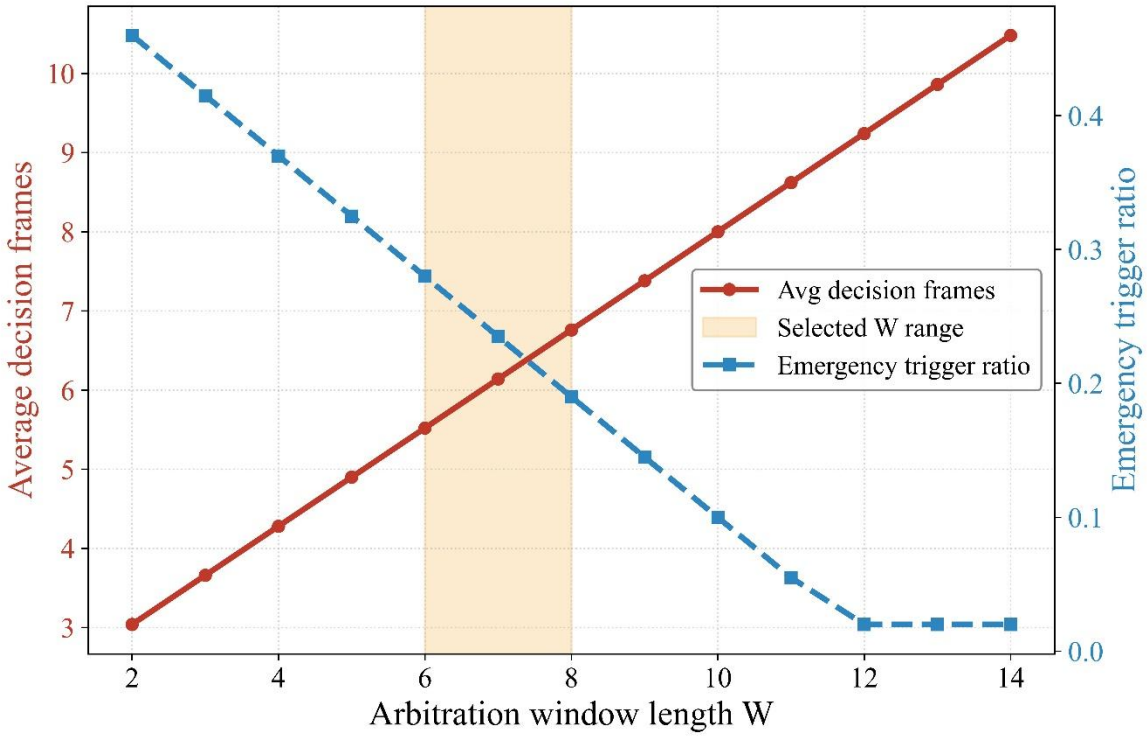


Figure 10. Effects of arbitration window length on decision delay and emergency path activation ratio

Figure 10 turns the choice of W in Section 5.1 from a stated hyperparameter into an evidence-based selection. If W is too small, stability ν_k cannot accumulate adequately; if W is too large, the response lag for mode (c) becomes unacceptable. Together, Figures 8-10 define the method's deployable region as urban scenes with semantic segmentation at no less than 8 Hz and no more than 15 dynamic instances per frame. Beyond this boundary, the system operates at a reduced frame rate rather than failing, while the emergency path preserves a minimum response capability.

6. CONCLUSIONS

This paper begins by reconsidering the root causes of dynamic object filtering failures rather than simply seeking more accurate segmentation or more sensitive motion detection. Existing semantic-prior rejection and geometric motion detection both implicitly treat dynamic decisions as one-step classification. The former directly uses category labels to decide retention, while the latter directly uses residual thresholds. Neither provides explicit rules for contradictory semantic and motion evidence. Our analysis identifies the absence of arbitration as a primary source of performance degradation: stationary objects in dynamic semantic classes are wrongly rejected, and genuinely moving objects missed by semantics are wrongly retained. Both reflect the failure of single-frame binary decisions when the two evidence sources disagree. We therefore reformulate dynamic object filtering as evidence fusion with delayed decisions and implement this through a complete architectural chain. Soft modulation converts semantic outputs into spatial attention weights; dual-stream fusion produces uncertain cross-modal motion confidence; temporal consistency supports conflict arbitration at the semantic instance level; and dynamic probabilities embed the resulting decisions in tightly coupled optimization through

information matrix modulation. Experimental validation tests the three claims along this chain rather than relying on superiority across every metric, allowing the contribution of each design decision to be independently traced and falsified.

The method's effective operating range must be honestly defined rather than broadly claimed to be universal. The arbitration mechanism uses semantic instances as its basic units. When instance segmentation is reliable, this design provides spatial consistency constraints and noise suppression. However, when severe occlusion fragments an instance, the same object is divided into several incomplete instances. Both within-instance spatial concentration and inter-frame association then degrade, causing arbitration to approach independent framewise decisions and weakening its temporal stability advantage. Similarly, although an emergency path imposes an upper bound on decision delay and enables a response within a limited number of frames when an object suddenly starts moving, this bound itself implies an inherent response-lag window. When a high-speed moving object suddenly cuts in, features within this window still participate in optimization with relatively high weights, introducing contamination that is limited but cannot be eliminated entirely. Accordingly, this paper does not claim applicability to all dynamic scenes. It explicitly defines the method's effective operating range as urban road scenes with available semantic segmentation and medium-to-high dynamic density. This is both the starting point of the design and the boundary within which its conclusions can be trusted.

Future work can extend evidence arbitration in two directions. The first is to replace manually designed conflict categories and fixed thresholds with a learnable evidence fusion strategy. The current four-mode classification and joint mean, stability, and concentration criteria depend on manually calibrated thresholds and weights. Learning the mapping from evidence to dynamic probability from a small, labeled dataset may support more complex conflicts while preserving interpretability. The second is to extend dynamic probabilities from optimization information matrix weights into persistent joint semantic-motion map labels. Currently, these probabilities support pose estimation only within the sliding window. Recording them together with uncertainty would let every map point retain historical evidence of having been dynamic, supporting long-term mapping and scene change detection. Dynamic decisions would then become reusable environmental attributes rather than by-products of online estimation. These directions extend the central argument through decision design and information persistence: effective dynamic object filtering depends on better organization of evidence rather than stronger isolated detectors.

Abbreviations

IQR, Interquartile Range.

Supplementary Material

Not applicable.

Appendix

Not applicable.

Ethics approval and consent to participate.

This study did not involve human participants, animal subjects, or any data requiring ethical approval. Therefore, ethics approval and consent to participate are not applicable.

Acknowledgements

The authors would like to thank the editors of this journal and all the anonymous reviewers who provided valuable comments on this work.

Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article.

Author contributions

All authors have read and agreed to the published version of the manuscript. The author's contributions are specified as follows: **X.Z.:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

Funding information

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability

The data that support the findings of this study are available upon request from the corresponding authors, **X.Z.**

Disclaimer

The views and opinions expressed in this article are those of the authors and are the product of professional research. It does not necessarily reflect the official policy or position of any affiliated institution, funder, agency, or that of the publisher. The authors are responsible for this article's results, findings, and content.

Declaration of AI and AI-assisted Technologies in the Writing Process

During the writing of this article, the author used ChatGPT for spelling and grammar checking. After using this tool, the author reviewed and edited the content as needed and assumes full

responsibility for the final published content.

REFERENCES

[1] Wang, H., Hu, Q., Li, B., Geng, Y., Su, C., Zhu, S., ... & Chen, W. (2026). RTS-SLAM: A Trajectory Consistency-Driven Multi-Constraint Dynamic Feature-Rejection Method for Visual SLAM in Dynamic Environments. Sensors, 26(9), 2846. https://doi.org/10.3390/s26092846

[2] Polukhin, A., Stirenko, S., Leier, M., Jervan, G., Rokovyi, O., Alienin, O., ... & Gordienko, Y. (2026). Depth-assisted sparse visual odometry for UAV-relevant synthetic RGB-D evaluation: A controlled geometric-backend ablation. Robotics, 15(7), 128. https://doi.org/10.3390/robotics15070128

[3] Zheng, S., Li, J., Wang, G., Ni, M., Palit, A., Montana, G., & Guan, Y. (2026). Person parametric physics-informed representation for mmwave-based human pose estimation. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 10(2), 1-34. https://doi.org/10.1145/3810219

[4] Zhang, Y. X., Gui, J., Yu, B., Cong, X., Gong, X., Tao, W., & Tao, D. (2026). Deep learning-based point cloud registration: A comprehensive survey and taxonomy. International Journal of Computer Vision, 134(4), 168. https://doi.org/10.1007/s11263-025-02723-w

[5] Cheng, X., Liu, X., Li, J., & Zhou, W. (2024). Deep learning-based point cloud registration: a comprehensive investigation. International Journal of Remote Sensing, 45(10), 3412-3442. https://doi.org/10.1080/01431161.2024.2343434

[6] Liu, M., Zou, Q., Long, J., Wang, Y., Lin, M., & Wang, F. (2025). SSE-SLAM: Semantic Visual SLAM Based on RGB-D Camera for High-Accuracy Pose Measurement in Dynamic Environments. IEEE Transactions on Instrumentation and Measurement. https://doi.org/10.1109/tim.2025.3621768

[7] Zhou, T., Porikli, F., Crandall, D. J., Van Gool, L., & Wang, W. (2022). A survey on deep learning technique for video segmentation. IEEE transactions on pattern analysis and machine intelligence, 45(6), 7099-7122. https://doi.org/10.1109/TPAMI.2022.3225573

[8] Li, X., Wang, L., Luo, H., & Li, G. (2026). RI-DVP: A Physics–Geometry Dual-Driven Framework for Static Map Construction in Sparse LiDAR Scenarios. Remote Sensing, 18(5), 821. https://doi.org/10.3390/rs18050821

[9] Chen, Z., Zhang, F., Zhu, C., Zhang, H., Chen, L., Zhu, H., & Yu, F. (2026). Enhancing PointPillars for 3-D Small Object Detection: A Multi-Dimensional Feature Fusion Approach. IEEE Sensors Journal. https://doi.org/10.1109/JSEN.2026.3680151

[10] Kojukhov, A., Levin, I., & Bovshover, A. (2025). From Subsumption to Semantic Mediation: A Generative Orchestration Architecture for Autonomous Systems. Algorithms, 18(12), 773. https://doi.org/10.3390/a18120773

[11] Sanket, N. J., Singh, C. D., Fermüller, C., & Aloimonos, Y. (2023). Ajna: Generalized deep uncertainty for minimal perception on parsimonious robots. Science Robotics, 8(81), eadd5139.

https://doi.org/10.1126/scirobotics.add5139
[12] Sun, Z., Xu, W., Shi, L., Cen, Y., Jin, Y., & Li, Y. (2026). UAGM: Uncertainty-Aware Geometric Modeling for Multi-Scenario 3-D Object Detection in Autonomous Vehicles. IEEE Transactions on Intelligent Transportation Systems. https://doi.org/10.1109/tits.2026.3685623
[13] Fang, Z., Zhou, S., Cui, Y., & Scherer, S. (2020). 3d-siamrpn: An end-to-end learning method for real-time 3d single object tracking using raw point cloud. IEEE Sensors Journal, 21(4), 4995-5011. https://doi.org/10.1109/jsen.2020.3033034
[14] Che, R., Hu, H., Wang, N., & Li, Z. (2026). Dense SLAM using 3D Gaussian Splatting with IMU and Geometric Constraints. IEEE Internet of Things Journal. https://doi.org/10.1109/JIOT.2026.3662499
[15] Tan, Z., Zhang, Z., Lyu, Y., & Pan, Q. (2026). EDOG-Map: Evidential Dynamic Occupancy Grid Mapping for Consumer-Grade UAV Navigation in Dynamic Environments. IEEE Transactions on Consumer Electronics. https://doi.org/10.1109/tce.2026.3725136
[16] Hirt, P. R., Holtkamp, J., Hoegner, L., Xu, Y., & Stilla, U. (2022). Occlusion detection of traffic signs by voxel-based ray tracing using highly detailed models and MLS point clouds of vegetation. International Journal of Applied Earth Observation and Geoinformation, 114, 103017. https://doi.org/10.1016/j.jag.2022.103017
[17] Chakraborty, D., & Dey, E. K. (2024). Segmentation of LiDAR point cloud data in urban areas using adaptive neighborhood selection technique. Plos one, 19(7), e0307138. https://doi.org/10.1371/journal.pone.0307138
[18] Zhou, B., Jiang, D., Dai, Y., & Hayat, T. (2023). Threshold dynamics and probability density function of a stochastic avian influenza epidemic model with nonlinear incidence rate and psychological effect. Journal of Nonlinear Science, 33(2), 29. https://doi.org/10.1007/s00332-022-09885-8
[19] Shen, S., Ito, T., & Hirose, T. (2025). Spatio-temporal gradient flow for efficient motion estimation in sparse point clouds. Journal of Robotics and Mechatronics, 37(6), 1327-1342. https://doi.org/10.20965/jrm.2025.p1327
[20] Wang, H., Hu, Q., Li, B., Geng, Y., Su, C., Zhu, S., ... & Chen, W. (2026). RTS-SLAM: A Trajectory Consistency-Driven Multi-Constraint Dynamic Feature-Rejection Method for Visual SLAM in Dynamic Environments. Sensors, 26(9), 2846. https://doi.org/10.3390/s26092846
[21] Morris, J., Wang, Y., Kliniewski, M., & Ila, V. (2025). Dynosam: Open-source smoothing and mapping framework for dynamic slam. IEEE Transactions on Robotics. https://doi.org/10.1109/tro.2025.3641813
[22] Li, Z., Jiang, C., Sun, Y., Su, H., & He, L. (2025). Precise feature removal method based on semantic and geometric dual masks in dynamic SLAM. Applied Sciences, 15(13), 7095. https://doi.org/10.3390/app15137095
[23] Xu, W., Liu, S., Chen, R., Du, S., & Wang, Y. (2026). Robust Multi-Modal Factor Graph Optimization for Distributed Collaborative LiDAR–Visual–Inertial SLAM. Applied Sciences, 16(10), 4677. https://doi.org/10.3390/app16104677

[24] Chen, J., Jia, K., & Wei, Z. (2025). Small but Mighty: A Lightweight Feature Enhancement Strategy for LiDAR Odometry in Challenging Environments. Remote Sensing, 17(15), 2656. https://doi.org/10.3390/rs17152656

[25] Gopalakrishnan, V., Dey, N., & Golland, P. (2026). PolyPose: Deformable 2D/3D Registration via Polyrigid Transformations. Advances in Neural Information Processing Systems, 38, 60633-60659. https://doi.org/10.48550/arXiv.2505.19256

[26] Zheng, J., Li, Y., Zhao, J., Chen, Y., Xu, J., & Xu, S. (2026). Residual Reweighting with Scan Normalization for LiDAR-Inertial Odometry in Degenerate Environments. IEEE Sensors Journal. https://doi.org/10.1109/JSEN.2026.3714871

[27] Deng, Z., Li, C., Lin, W., & Sun, L. (2026, June). GMM-LIO: Adaptive and Robust LiDAR-Inertial Odometry with Gaussian Mixture Model Voxel Map. In 2026 IEEE International Conference on Robotics and Automation (ICRA) (pp. 21276-21283). IEEE. https://doi.org/10.1109/icra57385.2026.11697299

[28] Liao, Y., Xie, J., & Geiger, A. (2022). Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3), 3292-3310. https://doi.org/10.1109/tpami.2022.3179507

APA Citation

Zhao, X. (2026). Deep Learning Based Dynamic Object Filtering for Semantic Information Assisted Visual LiDAR SLAM in Dynamic Scenes. *Journal of Discovery Core, 1*(2), 136-163. **https://doi.org/10.67541/jdc2612**