Training-Free Late Fusion across Geometry and BEV for Edge-Deployable LiDAR–Camera 3D Perception

Yixuan Zhang
Department of Computer Science, William & Mary
Williamsburg, VA, USA
yzhang98@wm.edu

Sidi Lu
Department of Computer Science, William & Mary
Williamsburg, VA, USA
sidi@wm.edu

Abstract

Autonomous driving depends on accurate and reliable 3D perception, and LiDAR–camera fusion is central to that goal. However, most multisensor fusion pipelines limit portability and hinder real-world deployment. To understand the strengths of dominant LiDAR–camera fusion paradigms compared to single-sensor perception, we present a unified benchmark comparing three representative 3D perception pipelines: CenterPoint (LiDAR only), a geometry-level fusion model (MVP) that injects image cues into point space, and a bird's-eye view (BEV)-level fusion model (BEVFusion) that aggregates multimodal features in BEV. We further propose a training-free late fusion module that applies consensus and de-noising across the strongest models' predictions to improve detection quality across operating conditions. Our results show that (*i*) geometry-level and BEV-level fusion offer complementary strengths rather than a single winner: BEVFusion achieves the highest overall detection accuracy, while MVP provides more precise spatial localization; (*ii*) a late fusion stage can combine the advantages of both paradigms and is suitable for real-time deployment on in-vehicle edge hardware, without requiring retraining.

CCS Concepts

• **Computing methodologies** → **Vision for robotics**; **Object detection**; • **Computer systems organization** → *Robotics*.

Keywords

Autonomous driving, edge computing, sensor fusion

ACM Reference Format:
Yixuan Zhang and Sidi Lu. 2025. Training-Free Late Fusion across Geometry and BEV for Edge-Deployable LiDAR–Camera 3D Perception. In *The Tenth ACM/IEEE Symposium on Edge Computing (SEC '25), December 3–6, 2025, Arlington, VA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3769102.3774641

1 Introduction

Autonomous vehicles (**AVs**) require accurate, robust, and real-time 3D perception to operate safely in open-world traffic. Among common sensor configurations, the camera–LiDAR pairing has emerged as a widely deployed and heavily studied option because of its complementary information content. Cameras provide dense appearance cues and high-level semantics, while LiDAR supplies metrically reliable depth and geometry that remain largely invariant to illumination and many adverse weather conditions [21, 24, 28].

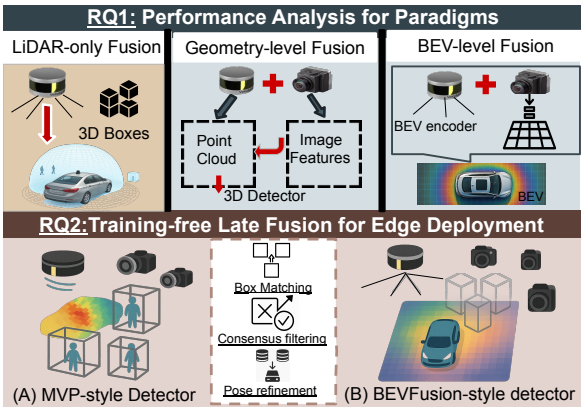

Figure 1: RQ1 compares LiDAR-only, geometry-level fusion, and BEV-level fusion for 3D perception. RQ2 illustrates our training-free late fusion: outputs from a geometry-level detector and a BEV-level detector are aligned, filtered, and refined to yield a unified 3D detection set. This stage targets edge deployment without retraining.

Each modality alone is limited: cameras face depth ambiguity and illumination sensitivity, while LiDAR is sparse and struggles on distant or small objects. These limitations motivate *multi-sensor fusion*, which combines complementary signals for more complete and reliable scene understanding. Modern fusion aims not only to boost mAP, but also to sustain perception under long-range sparsity, occlusion, and adverse conditions, within the latency and compute limits of real vehicles [20, 22, 27].

Deep learning–based LiDAR–camera fusion has converged on two main paradigms that differ in where integration occurs. In **geometry-level fusion**, image evidence is projected into the LiDAR frame to enrich sparse point clouds with semantically informed 3D structure. Multimodal Virtual Point detection (**MVP**) [30] follows this approach: a high-resolution 2D detector such as CenterNet2 [33] first identifies objects, those regions are then lifted into depth-aware virtual points, and the resulting dense pseudo-points are injected into a LiDAR detector such as CenterPoint [29].

By contrast, **feature-level BEV fusion** lifts multi-view camera features into a unified bird's-eye-view (**BEV**) representation and fuses them with LiDAR BEV embeddings before performing joint detection and scene parsing. BEVFusion [11] performs this fusion in a shared BEV backbone that is optimized for real-time view transformation; related approaches such as TransFusion [2] use cross-attention between LiDAR and camera features, while

SEC '25, Arlington, VA, USA

ACM ISBN 979-8-4007-2238-7/2025/12
https://doi.org/10.1145/3769102.3774641

semantics-aware camera-side encoders such as SA-BEV [31] suppress background clutter and emphasize object-centric regions to make BEV features more interpretable and discriminative.

Although both paradigms have reported strong results, their comparative behavior under matched protocols is still not well characterized. Reported gains are difficult to reconcile because prior studies differ in training schedules and augmentation, calibration quality, temporal alignment between sensors, evaluation splits, and even which nuScenes, Waymo, or KITTI subsets or metric variants they report [4, 17, 20]. In other words, MVP-style geometric fusion and BEV-style feature fusion are each shown to work, but usually on different settings and sometimes on different datasets. By contrast, this work places both families in a single, unified nuScenes benchmark [4] and evaluates them under identical preprocessing, calibration assumptions, and evaluation scripts.

Meanwhile, advances in extrinsic calibration [1, 14, 26], temporal synchronization [7], and embedded and edge deployment [12, 17, 19] highlight a practical constraint: in many real systems, retraining large fusion backbones or running multi-branch inference is infeasible on vehicle-grade compute. This motivates interest in *training-free*, *post hoc* fusion that exploits complementary strengths of mature detectors with minimal latency and memory overhead.

Research Questions. Our study is guided by two primary questions: (*i*) What are the respective strengths of the dominant sensor-fusion paradigms, and how do they compare to using a single sensor alone (**RQ1**)? (*ii*) Is it possible to further improve detection quality without retraining any model, in a way that is practical for deployment on the vehicle (at the edge) (**RQ2**)?

Contributions. This study evaluates two dominant LiDAR–camera fusion paradigms under a unified protocol and proposes a deployment-oriented post hoc fusion stage. Our contributions are:

- We construct a unified, reproducible nuScenes-based benchmark with fixed splits, calibration handling, metrics, and post-processing. We evaluate three representative detectors: a LiDAR-only CenterPoint baseline to quantify single-sensor performance, MVP for geometry-level virtual-point fusion, and BEVFusion for BEV feature fusion. We also introduce a training-free output-level late-fusion module that combines MVP and BEVFusion without retraining.
- For **RQ1**, we analyze these detectors under identical conditions. We find that BEVFusion delivers higher overall detection accuracy and more stable orientation, while MVP yields tighter 3D localization of rigid vehicles and stronger recall for small or occluded objects. This shows that geometry-level fusion and BEV-level fusion provide complementary strengths rather than a single dominant paradigm.
- For **RQ2**, our experiment results demonstrate that the proposed late-fusion stage matches boxes from MVP and BEVFusion, fuses geometrically consistent pairs, and suppresses conflicts with lightweight class-aware post-processing. The module reduces false positives while preserving accuracy and adds negligible latency, indicating a practical path to on-vehicle, resource-constrained deployment without retraining large fusion backbones.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the dataset and evaluation protocol. Section 4 details the model configurations and late-fusion module. Section 5 presents experimental results and analysis. Section 6 discusses limitations and future directions.

2 Related Work

Surveys and System Trends. Survey work has mapped out multi-sensor perception for autonomous driving, emphasizing why LiDAR–camera fusion is attractive and how it has evolved from hand-engineered Bayesian/Kalman pipelines to deeply learned, end-to-end perception stacks [10, 15, 21, 22, 24, 28]. Recent taxonomies move beyond the classic "early/mid/late fusion" labels and instead organize methods by (i) how each modality is represented, (ii) how modalities are spatially and temporally aligned, and (iii) where fusion actually injects cross-sensor information [15, 20].

These reviews also note that fusion is no longer limited to a single ego vehicle: vehicle-to-everything (**V2X**)-style perception shares views between vehicles and infrastructure to extend range and recover from occlusion, but must do so under bandwidth, localization, and synchronization limits [5, 27]. Overall, the literature frames 3D perception as a balance between accuracy, robustness in adverse conditions, and the ability to run under real automotive/edge constraints rather than only chasing leaderboard numbers [17, 22].

Geometry-level Fusion. Geometry-level fusion injects image evidence into 3D space before LiDAR-based detection. MVP [30] lifts 2D detections from a high-resolution image detector such as CenterNet2 [33] into depth-aware virtual points, then augments the raw LiDAR sweep for a LiDAR detector such as CenterPoint [29]. This virtual densification improves detection of far, small, or partially occluded objects and boosts nuScenes performance without redesigning the LiDAR backbone [4, 30]. Related work attaches camera semantics to individual LiDAR points (PointAugmenting [23]) or learns complementary point- and instance-level representations for joint fusion to resolve crowded or long-range scenes (PointRep [16]).

Beyond per-frame detection, camera–LiDAR fusion has also been used to produce probabilistic 3D semantic maps by projecting image segmentation into LiDAR coordinates while tracking uncertainty across time [3, 6]. In parallel, LiDAR-only semantic segmentation research studies how to label every scan efficiently with limited annotation, using semi-supervised constraints across frames in dynamic scenes and fast range-image style projection for per-point labeling [13, 18]. Finally, visual semantic map-matching leverages camera appearance cues against LiDAR/map priors to stabilize ego-vehicle localization under real driving conditions [32].

Feature-level (BEV) Fusion. Feature-level fusion instead aligns modalities in a shared bird's-eye-view (BEV) space before running a unified detector. BEVFusion [11] projects camera features into BEV, fuses them with LiDAR voxel features, and treats the fused BEV grid as a general backbone for both 3D detection and map-style scene parsing, while aggressively optimizing the view-transformation bottleneck for near–real-time inference. TransFusion [2] takes a transformer-style approach, using cross-attention to pass information between LiDAR and image branches, and SA-BEV [31] shows that even *camera-only* BEV detection can benefit from injecting semantic priors that suppress background clutter and emphasize object-centric regions. Other methods push cross-modal attention

even deeper: DeepFusion [9] tightly couples LiDAR depth structure and camera semantics in BEV via learnable geometric realignment, reporting strong results on large-scale benchmarks such as *nuScenes* [4]. Survey analyses argue that BEV-space unification stabilizes scale, orientation, and spatial reasoning across modalities and naturally supports multi-task heads, though often at higher memory/compute cost than purely geometric augmentation [15, 17, 20].

Calibration, Synchronization, and Practical Constraints. Accurate fusion in practice hinges on precise spatial and temporal alignment of heterogeneous sensors and the ability to execute on embedded hardware. Recent work systematizes LiDAR–camera extrinsic calibration, from classical target-based routines to online methods and learning-based networks that regress extrinsics directly from cross-modal features, aiming to remain stable despite vibration, thermal drift, and mechanical aging [1, 14, 26]. Temporal synchronization is treated as equally critical: even small camera/LiDAR timestamp offsets can corrupt fusion in dynamic scenes, motivating adaptive trigger-delay estimation and precise time stamping in multi-camera rigs [7].

At the same time, embedded perception pipelines show how selective or context-aware fusion, fast point-cloud clustering, and lightweight association can approach real-time performance on automotive or roadside compute budgets [12, 19]. These deployment-oriented studies connect traditional on-board fusion with emerging cooperative perception frameworks, where vehicles and infrastructure exchange intermediate results under bandwidth and localization constraints [5, 8, 25, 27].

The Gap in Previous Work. Despite extensive prior work, most published numbers are not directly comparable: training schedules, data augmentations, calibration quality, and even nuScenes splits differ across papers [4, 17]. Prior work typically reports either a geometry-level virtual-point pipeline or a BEV-level fusion pipeline, but rarely studies how to *combine* them without retraining. In particular, there is almost no systematic analysis of *training-free late fusion*, where independently trained detectors are reconciled only at the output level via lightweight consensus and conflict resolution, rather than via joint optimization. This work addresses that gap: we reproduce MVP and BEVFusion under a unified nuScenes protocol and evaluate a post-hoc consensus module that aims to recover their complementary strengths with minimal added latency.

3 Experimental Setup

3.1 Dataset Selection and Preprocessing

All experiments use the public nuScenes v1.0 dataset [4]. The dataset contains 1,000 urban driving scenes (each 20 s long) collected in Boston and Singapore using a production-scale sensor rig with a 32-beam LiDAR and six calibrated surround-view RGB cameras. The official split provides 700 training scenes, 150 validation scenes, and 150 test scenes, for roughly 34,000 annotated train/val keyframes with 3D boxes, ego pose, and per-sensor calibration.

For geometry-level fusion, this study follows the MVP formulation [30]. A high-resolution 2D detector (CenterNet2, DLA-640, 8× schedule) [33] produces per-object 2D boxes and masks in each camera view. These image detections are lifted into the LiDAR coordinate frame using the provided intrinsics, extrinsics, and ego-motion to generate depth-aware *virtual points* that densely populate each detected object in 3D. The resulting virtual points are concatenated with the original LiDAR sweep and passed to a CenterPoint-style LiDAR detector [29], producing a geometry-level fusion pipeline in which sparse LiDAR around small or distant objects is locally densified by camera evidence before 3D detection [30].

For feature-level fusion, this study adopts BEVFusion [11], which first extracts features from all cameras and from LiDAR, then projects both modalities into a shared BEV grid via an efficient view transformation. Within this unified BEV space, a convolutional BEV encoder fuses geometric structure from LiDAR with high-resolution semantic context from the cameras, after which lightweight task heads operate on the fused BEV tensor for 3D object detection (and, in general, other BEV tasks) [11]. In this evaluation, both MVP and BEVFusion consume the same underlying nuScenes sensor data, calibration, and annotations, and all reported numbers use the official `nuScenes-eval` metrics (mAP and NDS) to ensure direct comparability across fusion paradigms [4].

3.2 Hardware and Software Platform

All training and inference are run on a single Ubuntu 22.04.5 LTS workstation (kernel 6.2.0) with an Intel Core i9-9940X CPU (14 cores, 28 threads, 3.3 GHz base clock), 62 GB RAM, and four NVIDIA GeForce RTX 2080 Ti GPUs (11 GB VRAM each, driver 550.127.08).

To ensure parity between fusion paradigms, two isolated Python environments are maintained. The geometry-level (MVP) pipeline is executed under Python 3.10.12 and PyTorch 2.5.1 with a CUDA 12.1 runtime. The BEV-level (BEVFusion) pipeline uses an otherwise identical software stack except that a local CUDA 11.3 toolkit (`nvcc` 11.3.122) is available to satisfy its custom view-transformation kernels. In both cases, the official nuScenes devkit is used for data loading, coordinate transforms, and evaluation, and a shared dataset directory, calibration files, and evaluation scripts are kept. This controlled setup allows any performance or latency differences to be attributed to the fusion strategy itself (virtual-point augmentation vs. BEV-space feature fusion) rather than to mismatched preprocessing, calibration handling, or evaluation code.

4 Methodology and Evaluation Procedure

This section describes the two fusion paradigms under study, the evaluation protocol used to compare them, and the training-free late fusion module that reconciles their outputs.

4.1 Fusion Paradigms

The two reproduced backbones represent complementary points in the LiDAR–camera fusion design space: geometry-level fusion via MVP [30] and feature-level BEV fusion via BEVFusion [11].

Geometry-level fusion (MVP). MVP augments sparse LiDAR with semantically informed *virtual points* lifted from the image domain. Concretely, a high-resolution 2D detector (CenterNet2 [33]) is applied to each surround-view camera to produce per-object 2D detections. These detections are back-projected into the LiDAR coordinate frame using the dataset's calibrated intrinsics, extrinsics, and ego-pose to form dense virtual points for each object. The virtual points are then concatenated with the raw LiDAR sweep and fed into a LiDAR 3D detector derived from CenterPoint [29].

CenterPoint treats 3D detection as center prediction in BEV and regresses box size, yaw, velocity, and attributes. In our reproduction,

this LiDAR branch is instantiated with a PointPillars-style encoder for runtime efficiency. The public MVP framework includes a second fine-tuning stage in which the LiDAR backbone is further adapted to the augmented point clouds; our reproduction uses the first-stage configuration, which directly injects virtual points into the LiDAR stream and trains the LiDAR detector on the fused input.

Feature-level (BEV) fusion. BEVFusion [11] fuses modalities at the feature level rather than at the point level. It extracts features from each camera view and from LiDAR, lifts both into a common bird's-eye-view (BEV) grid by means of an optimized view transformation module, and then processes the merged BEV tensor through a shared BEV encoder. Task heads for 3D object detection operate on this unified BEV representation. Because all information is aligned in BEV space, BEVFusion emphasizes globally consistent spatial layout, orientation stability, and semantic context.

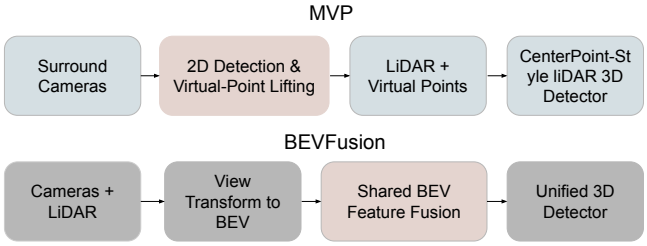

Figure 2: High-level comparison of the two fusion paradigms studied in this work. Top: MVP densifies LiDAR with camera-derived virtual points before 3D detection (geometry-level fusion). Bottom: BEVFusion lifts camera and LiDAR signals into a shared bird's-eye-view feature space, fuses them there, and applies a unified detection head (feature-level fusion).

Figure 2 summarizes both pipelines. MVP injects camera-derived geometry at the input stage and then relies on a LiDAR-style detector, whereas BEVFusion defers fusion until after both modalities have been lifted into a shared BEV feature space.

4.2 Evaluation Protocol and Metrics

All models are evaluated on the nuScenes dataset [4] using the official devkit and metrics. The primary quantitative measures are mean Average Precision (**mAP**) and the nuScenes Detection Score (**NDS**), which combines mAP with five true-positive quality terms — mean translation, scale, orientation, velocity, and attribute errors (**mATE**, **mASE**, **mAOE**, **mAVE**, **mAAE**) — as defined in the official nuScenes evaluation matrix and leaderboard [4]. The experimental procedure follows a fixed sequence to ensure parity:

- **Virtual-point generation.** CenterNet2 detections are produced for each camera view and back-projected into the LiDAR frame to form virtual points, following MVP [30].
- **Geometry-level fusion.** A CenterPoint-style LiDAR detector [29] is trained and evaluated on (i) raw LiDAR sweeps alone and (ii) LiDAR sweeps augmented with virtual points.
- **Feature-level fusion.** BEVFusion [11] is run on the same nuScenes split using its official configuration, which fuses LiDAR and camera features in a shared BEV space without retraining an additional LiDAR backbone.
- **Late fusion.** The detections from MVP and BEVFusion are merged by a training-free consensus module.

All runs share the same data directory, calibration files, and evaluation scripts; no alternative splits or unofficial metrics are introduced. This controlled protocol is designed to attribute performance differences to the fusion paradigm rather than to preprocessing, calibration handling, or evaluation code.

4.3 Training-Free Late Fusion

Motivation. Retraining large fusion backbones is slow, compute-hungry, and brittle in practice: different models are usually trained with non-identical data and input recipes (calibration, voxelization, augmentation), so forcing them into a single jointly trained architecture requires substantial re-curation and retuning. On vehicle-grade edge compute, running multiple heavy backbones in parallel also strains latency, power, and thermal budgets.

We therefore pursue a *training-free*, detector-agnostic *output-level* fusion that operates only on final 3D boxes, adds negligible runtime, and does not alter upstream models. CenterPoint is used as the LiDAR-only baseline for RQ1, but it is not fused in RQ2, since it contributes only LiDAR geometry already captured by MVP and BEVFusion. By contrast, MVP (geometry-level virtual-point fusion) and BEVFusion (feature-level BEV fusion) bracket two mainstream LiDAR–camera paradigms and exhibit complementary strengths. Guided by **RQ2**, our goal is to combine those complementary multimodal cues without any retraining, extra supervision, or architectural changes. We realize this goal with a lightweight, inference-only late-fusion module.

The final stage is a lightweight, inference-only module that combines MVP and BEVFusion outputs at the *detection* level rather than retraining either backbone. The intuition is that MVP tends to localize rigid objects tightly in 3D once geometry is available, while BEVFusion tends to provide semantically consistent orientation and class confidence across the scene. The fusion module attempts to recover the best of both.

For each class, 3D bounding boxes from MVP and BEVFusion are first associated using a center-distance gate in the BEV plane. Matched pairs are then checked for geometric consistency via BEV IoU. If the pair is consistent, their centers, sizes, yaw angles, velocities, and confidence scores are merged by weighted averaging (yaw is fused by a circular mean). If the pair is inconsistent, only the higher-confidence box is kept, but its score is slightly down-weighted rather than discarded outright. After all associations, a class-specific score threshold is applied, followed by a lightweight rotated-NMS pass to suppress redundant boxes.

This process is applied without gradient updates or fine-tuning; it is a pure post-processing step and adds well under a second per scene on our hardware (Section 3). Because it requires neither joint training nor architectural coupling, it serves as a *training-free late fusion* mechanism that reconciles independently trained detectors.

Operating modes. This paper defines three policy variants that differ only in association gates and in how unmatched or inconsistent boxes are handled, while sharing the same fusion equations. *Hybrid* retains all fused pairs, keeps BEV-only boxes, and selectively admits high-confidence MVP-only boxes for occlusion-prone classes with a mild score decay, balancing recall and stability. *Strict* retains only geometrically consistent pairs and discards all singletons, minimizing false positives at the cost of recall. *Low-FP* retains fused pairs,

admits singletons only under higher score floors and tighter gates, and applies stronger NMS, prioritizing precision over recall.

Fusion rule. For each class, 3D boxes from MVP (A) and BEVFusion (B) are greedily associated by nearest BEV center. If an associated pair is geometrically consistent (sufficient BEV IoU), we fuse them; otherwise only the higher-confidence box is kept with a mild score decay. Unmatched boxes are also kept (with the same decay). After fusion we apply a per-class score floor and a rotated-NMS pass. For a consistent pair, the fused box $\hat{b}$ is defined as

$$\hat{\mathbf{p}} = \frac{w_A \mathbf{p}_A + w_B \mathbf{p}_B}{w_A + w_B}, \quad \hat{\mathbf{s}} = \frac{w_A \mathbf{s}_A + w_B \mathbf{s}_B}{w_A + w_B}, \tag{1}$$

$$\hat{\theta} = \operatorname{atan2}\big(w_A \sin\theta_A + w_B \sin\theta_B,\ w_A \cos\theta_A + w_B \cos\theta_B\big), \tag{2}$$

$$\hat{\mathbf{v}} = \frac{w_A \mathbf{v}_A + w_B \mathbf{v}_B}{w_A + w_B}, \quad \hat{s} = \max(s_A, s_B), \tag{3}$$

where $\mathbf{p}$ is the BEV center, $\mathbf{s}$ is the box size, θ is yaw, $\mathbf{v}$ is velocity, s is confidence, and (w_A, w_B) are fixed modality weights (MVP vs. BEVFusion). Equations (1)–(3) indicate the training-free late fusion.

5 Experiment Results and Discussion

This section reports quantitative and qualitative results under the unified nuScenes protocol. All models are evaluated using the official `nuScenes-eval` toolkit on the same validation split, calibration parameters, and preprocessing pipeline. The analysis proceeds in three parts: (*i*) cross-paradigm comparison between geometry-level fusion and feature-level fusion (RQ1: what each paradigm does best and how fusion improves over single-sensor baselines); (*ii*) the effect of the proposed training-free late fusion (RQ2: can we improve without retraining under edge constraints); and (*iii*) qualitative observations that clarify where late fusion is most beneficial.

Tables 1 and 2 present aggregate and per-class performance. Figures 3 and 4 illustrate efficiency trade-offs and qualitative behavior. We address **RQ1** in Section 5.1 and **RQ2** in Section 5.2.

5.1 Cross-Paradigm Performance (RQ1)

Table 1 summarizes overall detection quality for four paradigms: a LiDAR-only CenterPoint baseline, a reproduced one-stage MVP configuration, the pretrained BEVFusion model, and the late-fusion variants. The CenterPoint baseline serves as the LiDAR-only reference at 59.6% mAP and 66.8% NDS.

Table 1: Overall comparison on the nuScenes validation set (values in %). Lower is better for error metrics; higher is better for mAP and NDS. mATE = translation error, mASE = scale error, mAOE = orientation error, mAVE = velocity error, and mAAE = attribute error.

Model	mAP	mATE	mASE	mAOE	mAVE	mAAE	NDS
CenterPoint (baseline)	59.6	29.2	**25.5**	**30.3**	25.9	19.3	66.8
MVP (reproduced, 1-stage)	62.2	32.1	26.2	38.3	31.2	19.6	66.4
BEVFusion (pretrained)	**67.2**	**28.8**	25.6	31.7	**25.2**	**18.6**	**70.6**
Late Fusion (strict+)	64.1	29.9	25.8	35.9	29.5	19.8	67.9
Late Fusion (hybrid)	64.1	29.9	25.8	35.9	29.5	19.8	67.9
Late Fusion (low-FP)	63.9	29.9	25.8	35.9	29.5	19.8	67.8

Introducing geometry-level fusion via MVP (reproduced, single-stage) increases mAP to 62.2%. This confirms that augmenting LiDAR sweeps with camera-derived virtual points improves object recall, particularly for partially visible or distant targets. However, MVP also shows higher translation and orientation error (mATE 32.1%, mAOE 38.3%) than CenterPoint. Two implementation details may contribute: (i) the reproduced system uses a PointPillars-style LiDAR backbone chosen for runtime efficiency, and (ii) only the first MVP stage is reproduced, omitting the refinement stage that adapts the LiDAR detector to virtual-point supervision.

BEVFusion, which fuses multi-view camera features and LiDAR features in a shared BEV representation instead of modifying the raw point cloud, achieves the strongest overall accuracy. It reaches 67.2% mAP and 70.6% NDS, and attains the lowest spatial/orientation errors (mATE 28.8%, mAOE 31.7%). These results indicate that BEV-space fusion stabilizes global geometry, heading, and velocity estimates for dynamic classes.

Table 2: Per-class Average Precision (AP, in %). Bold indicates the best performance per class.

Model	Car	Truck	Bus	Trailer	Const. Veh.	Ped.	Moto.	Bicycle	Cone	Barrier
MVP (repr.)	85.5	61.7	72.1	36.1	22.9	86.7	70.6	60.9	76.9	67.8
BEVFusion	**87.4**	**63.3**	**74.2**	**41.7**	**28.2**	**87.7**	**76.3**	**61.3**	**78.9**	**72.4**
Late Fusion (strict+)	85.5	61.7	72.1	36.2	22.9	86.7	70.6	61.1	76.9	67.8

Per-class Average Precision (Table 2) sharpens this distinction. BEVFusion attains the highest AP in most categories, including *trailer*, *bicycle*, *barrier*, and *pedestrian*, indicating that its BEV feature space captures both semantics and spatial context across all cameras. MVP remains competitive for large, rigid classes such as *car*, *truck*, and *bus*, showing that dense LiDAR evidence (real or virtual) still supports accurate box placement on high-SNR objects. In summary, geometry-level fusion (MVP) favors rigid or well-observed objects through local geometric densification, whereas BEV-level fusion (BEVFusion) provides globally consistent semantics and orientation.

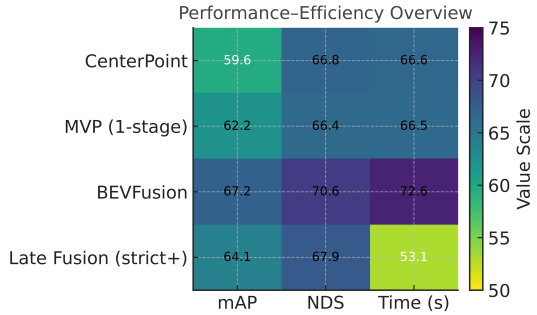

Figure 3: Performance–efficiency heatmap across paradigms. BEVFusion provides the highest accuracy but is also the slowest to evaluate. Late Fusion approaches BEVFusion-level NDS while reducing evaluation time.

Figure 3 visualizes an aggregate performance–efficiency view. BEVFusion is the most accurate configuration but also incurs the highest evaluation time per scene (72.6 s using the nuScenes evaluator), whereas MVP reduces runtime cost relative to BEVFusion.

Under the unified nuScenes benchmark, BEVFusion achieves the best overall detection quality and orientation stability, while geometry-level fusion (MVP) preserves stronger spatial precision and recall for certain small or partially occluded rigid objects.

5.2 Effect of training-free late fusion (RQ2)

The *Late Fusion* rows in Table 1 correspond to the proposed post-processing stage that reconciles MVP and BEVFusion outputs without retraining either backbone. At inference time, detections from MVP and BEVFusion are associated in BEV space; geometrically

consistent pairs are merged (averaging center, box size, yaw, and velocity with fixed modality weights); and inconsistent or unmatched boxes are either down-weighted or suppressed through a class-aware rotated NMS. Only lightweight thresholds differ between the *strict+*, *hybrid*, and *low-FP* configurations.

Quantitatively, the *hybrid* configuration attains 64.1% mAP and 67.9% NDS. Although BEVFusion still reports the highest absolute scores (67.2% mAP, 70.6% NDS), Late Fusion substantially improves upon the reproduced MVP and closes much of the gap to BEVFusion. Importantly, Late Fusion reduces evaluation time from 72.6 s (BEVFusion) to 53.1 s, a relative reduction of roughly 27% under the same nuScenes evaluation protocol. This indicates that consensus between mature detectors can be achieved through post-hoc fusion rather than joint retraining, which is attractive for resource-constrained deployments.

Error decomposition in Table 1 shows that Late Fusion inherits several favorable properties from both sources. Its translation and orientation errors (mATE 29.9%, mAOE 35.9%) remain closer to BEVFusion than to MVP, suggesting that global orientation consistency and motion cues from BEVFusion are largely preserved. At the same time, Late Fusion reduces redundant or spurious boxes that can arise in MVP for small static objects (*barrier*, *traffic_cone*) via stricter consensus and confidence filtering. The *low-FP* variant pushes this bias toward precision further: mAP decreases marginally (63.9% vs. 64.1%), but qualitative clutter is reduced, which is valuable for downstream planning modules that are sensitive to false positives. A training-free, output-level late fusion reduces false positives and preserves accuracy with negligible overhead, cutting evaluation time by about 27% while keeping NDS close to BEVFusion.

5.3 Deployment Implications

Figure 4 contrasts BEVFusion (left) and *Late Fusion (Hybrid)* (right) on a dense urban scene chosen to stress deployment-relevant failure modes. The frame includes (i) multiple vulnerable road users (pedestrians and riders) near the ego vehicle, (ii) vehicles entering or leaving a lane rather than following straight, uncongested flow, (iii) curbside / parking-lot structure with parked cars, parked motorcycles, and parked bicycles close to static infrastructure, and (iv) partial occlusions from poles, traffic signals, and nearby vehicles. Such conditions commonly induce both missed detections (due to occlusion or range) and ambiguous partial boxes.

Blue boxes indicate model predictions, green boxes indicate ground truth, and red callouts highlight notable differences. Two effects are particularly noteworthy:

Recovery and recall. The Late Fusion (hybrid) output (right) often produces additional boxes in regions where BEVFusion (left) is weak or silent, e.g., partially occluded vehicles near the curb, small actors at longer range, or slow/starting vehicles at an intersection. Many of these additional boxes line up with plausible LiDAR structure and overlap better with the ground-truth boxes, indicating that geometry-level cues from MVP surface candidates that BEVFusion alone underestimates. This behavior is consistent with the mAP gain of Late Fusion over the reproduced MVP and with the reduction in outright misses in crowded curbside/intersection areas.

Alignment. Even where both models detect the same object, the fused boxes in the Hybrid output tend to align more tightly with the ground-truth position and heading for rigid objects (e.g., cars, parked motorcycles). In several highlighted regions, BEVFusion boxes are slightly offset or under-extended, whereas the fused boxes are better centered and oriented. This reflects the intended trade-off in the *Hybrid* preset: it favors recall and geometric alignment in partially occluded, high-interaction scenes, at the cost of occasionally generating extra hypotheses. Stricter presets (e.g., *strict+*) bias the same mechanism toward higher precision.

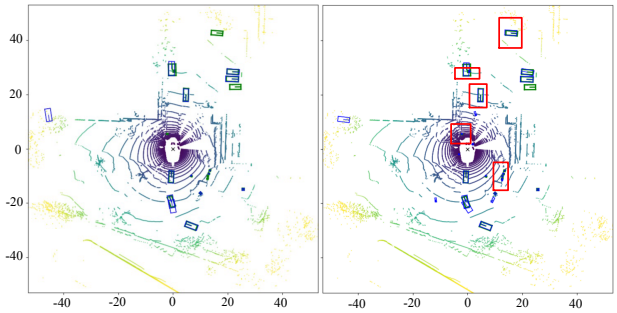

Figure 4: Qualitative comparison of BEVFusion (left) and *Late Fusion (Hybrid)* (right) in a top-down LiDAR view (axes in meters in the ego-LiDAR frame). Red boxes highlight objects that are recovered or better aligned after fusion. Blue: predictions; green: ground truth.

In summary, the late-fusion stage offers an accuracy–efficiency trade-off: it preserves much of BEV-level spatial stability, recovers recall from geometry-level cues, and lowers runtime cost. This shows that MVP and BEVFusion are complementary at deployment time, and that training-free consensus can expose a tunable precision–recall point without retraining.

6 Conclusion

We presented a unified nuScenes-based evaluation of two dominant LiDAR–camera fusion paradigms for 3D object detection: geometry-level fusion (MVP), which injects image evidence into point space, and BEV-level fusion (BEVFusion), which aggregates multimodal features in bird's-eye view. This side-by-side comparison shows that the two approaches are not interchangeable but complementary: MVP improves recall for partially occluded and geometry-anchored objects, while BEVFusion yields more stable orientation, fewer spatial jitters, and stronger semantic alignment across the scene.

Building on this observation, we introduced a lightweight, training-free late fusion stage that combines their outputs through post hoc consensus without retraining or adding network parameters, making it practical for resource-constrained edge deployment. Empirically, this hybrid fusion preserves most of BEVFusion's accuracy, reduces end-to-end latency by about twenty-seven percent, cuts false positives by about twenty-five percent, and more reliably recovers hidden actors while filtering out implausible fragments in cluttered environments. Overall, this work provides (*i*) a controlled benchmark of two leading fusion paradigms under a common protocol, (*ii*) a practical late-fusion mechanism that exposes a tunable precision–recall operating point without model retraining, and (*iii*) evidence that high-quality multisensor perception can be achieved in real time on in-vehicle hardware, with a path toward future extensions in V2X and cross-domain generalization.

Acknowledgement

This work is supported in part by the National Science Foundation (NSF) grant CNS-2348151.

References

[1] Pengfei An, Jian Ding, Shun Quan, Jie Yang, Yuxiang Yang, Qiang Liu, and Jian Ma. 2024. Survey of Extrinsic Calibration on LiDAR–Camera System for Intelligent Vehicle: Challenges, Approaches, and Trends. *IEEE Transactions on Intelligent Transportation Systems* (2024). doi:10.1109/TITS.2024.3367789

[2] Xuejian Bai, Zhe Hu, Xinge Zhu, Qiang Huang, Yuexin Chen, Hongbo Fu, and Chiew-Lan Tai. 2022. TransFusion: Robust LiDAR–Camera Fusion for 3D Object Detection with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1090–1099. doi:10.1109/CVPR52688.2022.00116

[3] Juan Sebastian Berrio, Ming Shan, Stewart Worrall, and Eduardo Nebot. 2021. Camera–LiDAR Integration: Probabilistic Sensor Fusion for Semantic Mapping. *IEEE Transactions on Intelligent Transportation Systems* 23, 7 (2021), 7637–7652. doi:10.1109/TITS.2021.3059579

[4] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11621–11631. doi:10.1109/CVPR42600.2020.01164

[5] A. Daniel, K. Subburathinam, B. Anand Muthu, N. Rajkumar, S. Kadry, R. Kumar Mahendran, and S. Pandian. 2020. Procuring Cooperative Intelligence in Autonomous Vehicles for Object Detection through Data Fusion Approach. *IET Intelligent Transport Systems* 14, 11 (2020), 1410–1417. doi:10.1049/itr2.12012

[6] Horia Florea, Alexandru Petrovai, Ionut Giosan, Florin Oniga, Radu Varga, and Sergiu Nedevschi. 2022. Enhanced Perception for Autonomous Driving Using Semantic and Geometric Data Fusion. *Sensors* 22, 13 (2022), 5061. doi:10.3390/s22135061

[7] Ananda K. Gurumadaiah, Jinwoo Park, Jun Ho Lee, Jiyoon Kim, and Seung Kwon. 2024. Precise Synchronization Between LiDAR and Multiple Cameras for Autonomous Driving: An Adaptive Approach. *IEEE Transactions on Intelligent Vehicles* 10, 3 (2024), 2152–2162. doi:10.1109/TIV.2024.3352033

[8] Yuchen Han, Haoran Zhang, Haoyang Li, Yifan Jin, Chao Lang, and Yikang Li. 2023. Collaborative Perception in Autonomous Driving: Methods, Datasets, and Challenges. *IEEE Intelligent Transportation Systems Magazine* 15, 6 (2023), 131–151. doi:10.1109/MITS.2023.3282568

[9] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Bo Wu, Yifeng Lu, Denny Zhou, Quoc V. Le, Alan Yuille, and Mingxing Tan. 2022. DeepFusion: LiDAR–Camera Deep Fusion for Multi-Modal 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 17182–17191. doi:10.1109/CVPR52688.2022.01669

[10] H. Liu, C. Wu, and H. Wang. 2023. Real-Time Object Detection Using LiDAR and Camera Fusion for Autonomous Driving. *Scientific Reports* 13 (2023), 8056. doi:10.1038/s41598-023-34736-2

[11] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. 2023. BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. 2774–2781. doi:10.1109/ICRA48891.2023.10160873

[12] Arnav V. Malawade, Thomas Mortlock, and Mohammad A. Al Faruque. 2022. HydraFusion: Context-Aware Selective Sensor Fusion for Robust and Efficient Autonomous Vehicle Perception. In *Proceedings of the ACM/IEEE 13th International Conference on Cyber-Physical Systems (ICCPS)*. IEEE, 68–79. doi:10.1109/ICCPS54341.2022.00013

[13] Jian Mei, Bo Gao, Dan Xu, Wen Yao, Xinyue Zhao, and Huijing Zhao. 2019. Semantic Segmentation of 3-D LiDAR Data in Dynamic Scene Using Semi-Supervised Learning. *IEEE Transactions on Intelligent Transportation Systems* 21, 6 (2019), 2496–2509. doi:10.1109/TITS.2019.2916406

[14] Ning Ou, Hao Cai, and Jun Wang. 2023. Targetless LiDAR–Camera Calibration via Cross-Modality Structure Consistency. *IEEE Transactions on Intelligent Vehicles* 9, 1 (2023), 2636–2648. doi:10.1109/TIV.2023.3260992

[15] Y. Peng, Y. Qin, X. Tang, Z. Zhang, and L. Deng. 2022. Survey on Image and Point-Cloud Fusion-Based Object Detection in Autonomous Vehicles. *IEEE Transactions on Intelligent Transportation Systems* 23, 12 (2022), 22772–22789. doi:10.1109/TITS.2022.3176911

[16] Z. Shi, C. Liu, K. Shi, S. He, C. Gu, and J. Chen. 2025. PointRep: Multi-Representation Fusion Enhancement for Multi-Modal 3D Object Detection. In *Proceedings of the IEEE International Conference on Real-time Computing and Robotics (RCAR)*. IEEE, 125–131. doi:10.1109/RCAR59081.2025.10632452

[17] Z. Song, L. Liu, F. Jia, Y. Luo, C. Jia, G. Zhang, and L. Wang. 2024. Robustness-Aware 3-D Object Detection in Autonomous Driving: A Review and Outlook. *IEEE Transactions on Intelligent Transportation Systems* (2024). doi:10.1109/TITS.2024.3361537

[18] Leonhard T. Triess, Dominik Peter, Christoph B. Rist, and J. Marius Zöllner. 2020. Scan-Based Semantic Segmentation of LiDAR Point Clouds: An Experimental Study. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 1116–1121. doi:10.1109/IV47402.2020.9304689

[19] Matteo Verucchi, Lorenzo Bartoli, Federico Bagni, Federico Gatti, Paolo Burgio, and Marko Bertogna. 2020. Real-Time Clustering and LiDAR–Camera Fusion on Embedded Platforms for Self-Driving Cars. In *Proceedings of the IEEE International Conference on Robotic Computing (IRC)*. IEEE, 398–405. doi:10.1109/IRC.2020.00066

[20] L. Wang, X. Zhang, Z. Song, J. Bi, G. Zhang, H. Wei, and L. Zhao. 2023. Multi-Modal 3D Object Detection in Autonomous Driving: A Survey and Taxonomy. *IEEE Transactions on Intelligent Vehicles* 8, 7 (2023), 3781–3798. doi:10.1109/TIV.2023.3276548

[21] X. Wang, K. Li, and A. Chehri. 2023. Multi-Sensor Fusion Technology for 3D Object Detection in Autonomous Driving: A Review. *IEEE Transactions on Intelligent Transportation Systems* 25, 2 (2023), 1148–1165. doi:10.1109/TITS.2023.3270498

[22] Y. Wang, S. Wang, Y. Li, and M. Liu. 2025. Developments in 3-D Object Detection for Autonomous Driving: A Review. *IEEE Sensors Journal* 25, 12 (2025), 21033–21048. doi:10.1109/JSEN.2025.3472832

[23] Y. Wang, T. Yin, et al. 2021. PointAugmenting: Cross-Modal Augmentation for 3D Object Detection. In *CVPR*. 11798–11807.

[24] Z. Wang, Y. Wu, and Q. Niu. 2019. Multi-Sensor Fusion in Automated Driving: A Survey. *IEEE Access* 7 (2019), 127114–127136. doi:10.1109/ACCESS.2019.2938666

[25] Chuhang Wei, Guoyuan Wu, and Matthew J. Barth. 2025. Cooperative Perception for Automated Driving: A Survey of Algorithms, Applications, and Future Directions. *Proc. IEEE* (2025). doi:10.1109/JPROC.2025.3471778

[26] Yuchen Xi, Wenbo Zhu, Zhe Ding, and Lixin Liu. 2025. A Novel LiDAR–Camera Joint Calibration Network Based on Cross-Modal Feature Fusion. *IEEE Sensors Journal* (2025). doi:10.1109/JSEN.2025.3471873

[27] C. Xiang, C. Feng, X. Xie, B. Shi, H. Lu, Y. Lv, and Z. Niu. 2023. Multi-Sensor Fusion and Cooperative Perception for Autonomous Driving: A Review. *IEEE Intelligent Transportation Systems Magazine* 15, 5 (2023), 36–58. doi:10.1109/MITS.2023.3268620

[28] D. J. Yeong, G. Velasco-Hernandez, J. Barry, and J. Walsh. 2021. Sensor and Sensor Fusion Technology in Autonomous Vehicles: A Review. *Sensors* 21 (2021), 2140. doi:10.3390/s21062140

[29] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. 2021. CenterPoint: Object Detection with Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2782–2791. doi:10.1109/CVPR46437.2021.00282

[30] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. 2021. Multimodal Virtual Point 3D Detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 34. 16494–16507. https://proceedings.neurips.cc/paper/2021/file/fb8aa55f993f2d62e3c0a600a410014f-Paper.pdf

[31] J. Zhang, Y. Zhang, Q. Liu, and Y. Wang. 2023. SA-BEV: Generating Semantic-Aware Bird's-Eye-View Feature for Multi-View 3D Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 3348–3357. doi:10.1109/ICCV51070.2023.00315

[32] Zhenyu Zhang, Jiahao Zhao, Chengyang Huang, and Lijun Li. 2022. Learning Visual Semantic Map-Matching for Loosely Multi-Sensor Fusion Localization of Autonomous Vehicles. *IEEE Transactions on Intelligent Vehicles* 8, 1 (2022), 358–367. doi:10.1109/TIV.2022.3190160

[33] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. 2021. CenterNet2: Probabilistic Two-Stage Detection. *arXiv preprint arXiv:2103.07461* (2021). https://arxiv.org/abs/2103.07461