Journal of Science, Innovation & Social Impact

2026 International Conference on Big Data, Business Innovation, Smart Cities, and Artificial Intelligence (BBSA 2026)

Article

Benchmarking Robustness-Efficiency Trade-offs of Camera-LiDAR Fusion Strategies for 3D Object Detection Under Environmental Corruptions

Yuhan Li [1,*]

[1] Computer Science, Northeastern University, Boston, MA, USA
* Correspondence: Yuhan Li, Computer Science, Northeastern University, Boston, MA, USA

Abstract: Multi-modal sensor fusion combining camera and LiDAR data has become the dominant paradigm for 3D object detection in autonomous driving. The selection among early fusion, late fusion, and deep fusion strategies involves complex trade-offs between detection accuracy, computational efficiency, and robustness under adverse conditions. This paper presents a systematic benchmarking study that quantitatively evaluates these trade-offs on the nuScenes dataset under diverse environmental corruptions and calibration perturbations. Six representative fusion algorithms spanning three fusion categories are evaluated across 10 corruption types at three severity levels, with simultaneous measurement of detection accuracy (mAP, NDS), computational resource consumption (latency, GPU memory), and degradation patterns under spatial misalignment and temporal desynchronization. The results reveal that deep fusion approaches achieve 2.8%--5.1% higher NDS than early fusion under clean conditions, while late fusion strategies demonstrate 12.3%--18.7% lower mean Corruption Error under LiDAR-degraded scenarios. Calibration perturbation analysis shows that soft-association mechanisms reduce mAP degradation by 41.2% compared to hard-association approaches at 0.5-meter spatial misalignment. These findings provide evidence-based guidance for engineering teams selecting sensor fusion configurations under real-world deployment constraints.

Keywords: multi-modal sensor fusion; robustness benchmarking; 3D object detection; autonomous driving perception

Received: 28 February 2026
Revised: 16 April 2026
Accepted: 30 April 2026
Published: 06 May 2026

1. Introduction

1.1. Background and Motivation

The perception pipeline of autonomous driving vehicles relies on complementary sensor modalities to achieve reliable environmental understanding. Camera sensors capture high-resolution semantic information, while LiDAR sensors provide precise geometric measurements through three-dimensional point clouds. The nuScenes benchmark [1], aggregating data from six cameras, five radars, and one LiDAR across 1,000 driving scenes in Boston and Singapore, has established the standard evaluation protocol for multi-modal 3D object detection. The Waymo Open Dataset [2] extends this landscape with 1,150 scenes across multiple U.S. cities, incorporating hierarchical difficulty stratification. Feng et al [3]. systematically categorized multi-modal fusion approaches into early, middle, and late fusion paradigms. The present study adopts the terminology of early, late, and deep fusion, where "deep fusion" corresponds to Feng et al.'s "middle fusion" category, referring to methods that integrate multi-modal features at intermediate representation stages rather than at the raw-input or decision level.

The question of which fusion strategy to deploy under specific operational conditions remains inadequately addressed. Dong et al [4]. introduced corruption benchmarks (KITTI-C, nuScenes-C, Waymo-C) encompassing 27 corruption types and evaluated 24 detection models, revealing that fusion approaches exhibit asymmetric robustness profiles depending on whether image or point-cloud corruptions dominate. Kong et al [5]. proposed standardized robustness metrics---mean Corruption Error (mCE) and mean Resilience Rate (mRR)---and benchmarked 34 perception models across eight corruption types, demonstrating that architecturally similar models with comparable clean-condition accuracy can exhibit dramatically different robustness characteristics.

1.2. Research Scope and Contributions

1.2.1. Research Scope

This study conducts a controlled benchmarking evaluation of six representative Camera-LiDAR fusion algorithms spanning three fusion categories (early, late, and deep fusion) on the nuScenes validation set. The evaluation encompasses three dimensions: detection accuracy under both clean and corrupted conditions, computational resource utilization across GPU platforms, and sensitivity to calibration errors and temporal desynchronization. The corruption protocol incorporates 10 environmental perturbation types at three severity levels, covering weather-related degradations, sensor noise, and geometric distortions. The study does not propose novel architectures or training procedures; its contribution lies exclusively in systematic evaluation methodology and quantitative analysis.

1.2.2. Key Contributions

The principal contributions of this work are threefold. The study establishes a unified evaluation protocol that simultaneously measures accuracy, robustness, and computational efficiency for Camera-LiDAR fusion algorithms under identical experimental conditions. The benchmarking results reveal previously unreported asymmetries in how fusion strategies degrade under different corruption categories, with quantitative evidence that the optimal fusion strategy depends strongly on the anticipated operational environment. The calibration sensitivity analysis provides deployment-relevant data showing the relationship between sensor alignment precision and detection performance across fusion categories, enabling engineering teams to establish calibration tolerance specifications based on empirical measurements.

2. Related Work

2.1. Multi-Modal Fusion Strategies for 3D Object Detection

2.1.1. Fusion Strategy Taxonomy

Multi-modal fusion strategies for 3D object detection are categorized by the stage at which information from different sensors is combined. Early fusion approaches merge raw or minimally processed sensor data before feature extraction. PointPainting [6] represents the canonical early fusion method, projecting LiDAR points into image semantic segmentation outputs and appending class probability scores to each point before passing them to a downstream LiDAR detector. Late fusion strategies process each modality through independent pipelines, combining outputs at the decision level. CenterPoint [7], achieving 65.5 NDS on nuScenes as the dominant LiDAR-only baseline, provides the reference architecture upon which many late fusion approaches build. In this study, the late fusion representative CenterPoint+ImgFeat extends CenterPoint by adding an independent image branch that extracts 2D detection scores from a ResNet-50 backbone; these image-derived confidence scores are combined with CenterPoint's LiDAR-based detection scores through weighted score-level fusion at the output stage, without modifying the LiDAR feature extraction pipeline.

2.1.2. Representative Deep Fusion Approaches

Deep fusion methods integrate multi-modal features at intermediate representation stages, enabling cross-modal feature interaction during encoding. TransFusion [8]

introduced a transformer-based soft-association mechanism that adaptively fuses object queries with image features through cross-attention, demonstrating superior robustness to sensor misalignment compared to hard-association methods relying on calibration-matrix-based projection. DeepFusion [9] explored feature-level interaction using cross-attention in 3D space, providing standardized latency comparisons across fusion configurations on identical V100 hardware. BEVFusion from MIT (Liu et al [10]., hereafter BEVFusion-MIT) unified multi-modal features in bird's-eye view space, achieving 70.2% mAP and 72.9% NDS on nuScenes test with optimized BEV pooling that reduced view transformation latency by over 40×. A concurrent variant by Liang et al [11]. from Peking University (hereafter BEVFusion-PKU) demonstrated that independent modality streams with BEV-space concatenation improve PointPillars by 18.4% mAP and CenterPoint by 7.1% mAP, with sustained margins of 15.7%--28.9% mAP under 50% LiDAR point dropout. The two BEVFusion variants differ in their fusion mechanism: BEVFusion-MIT employs a shared BEV encoder with task-specific heads for multi-task learning, while BEVFusion-PKU uses dynamic fusion with adaptive attention weights across modality-specific BEV features.

2.2. Robustness Benchmarking in Autonomous Driving Perception

The systematic evaluation of perception robustness has emerged as a dedicated research direction. RoboBEV [12] evaluated 33 BEV perception models across eight corruption types at three severity levels, identifying that strong clean-condition performance correlates with absolute corrupted-condition performance, but relative robustness does not necessarily improve with higher baseline accuracy. Yu et al [13]. created nuScenes-R and Waymo-R benchmarks with seven realistic corruption scenarios, revealing that most fusion methods exhibit disproportionate LiDAR dependency and fail catastrophically when LiDAR input is disrupted. MultiCorrupt [14] extended the evaluation scope by testing five state-of-the-art multi-modal detectors under 10 corruption types, introducing Resistance Ability (RA) and Relative Resistance Ability (RRA) metrics. The finding that early fusion approaches are more sensitive to spatial and temporal misalignment than loosely-coupled methods directly motivates the calibration sensitivity analysis in the present study.

3. Experimental Design and Methodology

3.1. Benchmark Dataset and Evaluation Metrics

All experiments in this study are conducted on the nuScenes dataset validation split, containing 6,019 frames across 150 scenes with full 360-degree sensor coverage. The selection of nuScenes is motivated by its comprehensive annotation of 1.4 million 3D bounding boxes across 10 foreground object classes, its inclusion of diverse environmental conditions, and its widespread adoption as the standard benchmark across the fusion literature.

Detection accuracy is measured using two complementary metrics. The mean Average Precision (mAP) follows the nuScenes official protocol: for each object class, Average Precision is computed using center distance matching at four thresholds (0.5, 1.0, 2.0, and 4.0 meters), then averaged across all 10 foreground classes. The nuScenes Detection Score (NDS) is a weighted combination of mAP (50%) and five true-positive error metrics (10% each): mean Average Translation Error (mATE), mean Average Scale Error (mASE), mean Average Orientation Error (mAOE), mean Average Velocity Error (mAVE), and mean Average Attribute Error (mAAE). Robustness evaluation adopts the mean Corruption Error (mCE) and mean Resilience Rate (mRR) metrics following Kong et al [5].. mCE quantifies relative performance degradation normalized against CenterPoint [7] as the reference detector, ensuring cross-study comparability; mRR measures the fraction of clean-condition performance retained under corruption.

Computational efficiency metrics include end-to-end inference latency (milliseconds per frame), peak GPU memory allocation (megabytes), and throughput (frames per

second). All latency measurements are averaged over 500 inference iterations with 50-iteration warm-up, using identical batch size and input resolution configurations.

3.2. Fusion Algorithm Selection and Implementation

3.2.1. Selected Algorithms and Fusion Categories

Six fusion algorithms are selected to represent the three principal fusion categories, with selection criteria emphasizing public code availability, reproducibility, and demonstrated competitive performance on nuScenes. Table 1 summarizes the selected algorithms, their fusion categories, backbone architectures, and key design characteristics.

Table 1. Overview of Selected Fusion Algorithms

Algorithm	Fusion Category	LiDAR Backbone	Camera Backbone	Fusion Mechanism	Publication
PointPainting	Early Fusion	CenterPoint-Pillar	HRNet-W48	Sequential point decoration	CVPR 2020
CenterPoint+ImgFeat	Late Fusion	VoxelNet	ResNet-50	Score-level combination	CVPR 2021
TransFusion	Deep Fusion	VoxelNet	ResNet-50	Transformer soft-association	CVPR 2022
BEVFusion-MIT	Deep Fusion	VoxelNet-SPConv	Swin-Tiny	BEV-space concatenation	ICRA 2023
BEVFusion-PKU	Deep Fusion	CenterPoint-Voxel	Dual-Swin-Tiny	BEV-space dynamic fusion	NeurIPS 2022
CMT	Deep Fusion	VoxelNet	VoVNet-99	Coordinate-encoded cross-attention	ICCV 2023

CMT (Cross Modal Transformer) [15] employs implicit spatial alignment through coordinate encoding without explicit view transformation, a design choice hypothesized to provide built-in robustness to calibration perturbations. Note on cross-modality reference: CRN [16] represents the camera-radar fusion paradigm, achieving 62.4 NDS on nuScenes test set as the leading camera-radar approach. CRN is included solely as an external reference point to contextualize the camera-LiDAR accuracy range and does not participate in the corruption, calibration, or efficiency experiments reported in this study, as its distinct sensor modality (radar vs. LiDAR) would preclude meaningful comparison under identical corruption protocols.

3.2.2. Hardware and Software Configuration

All experiments are executed on a unified hardware and software stack. The primary evaluation platform consists of an NVIDIA A100 (80GB) GPU with CUDA 11.8 and PyTorch 1.13.1. Inference latency profiling is additionally performed on NVIDIA RTX 3090 (24GB) and NVIDIA Jetson AGX Orin (32GB) platforms to characterize cross-platform efficiency variation. All models are implemented within the MMDetection3D framework using published pretrained weights without fine-tuning or test-time

augmentation. It should be noted that the selected algorithms employ different backbone architectures (e.g., HRNet-W48, ResNet-50, Swin-Tiny, VoVNet-99) and were originally trained with varying data augmentation strategies and input resolutions. To preserve reproducibility and avoid confounding variables introduced by re-training, this study evaluates each method using its officially released configuration and weights. The comparison is therefore controlled at the inference-environment level (identical hardware, software framework, input data, and evaluation protocol) rather than at the training-recipe level; readers should interpret accuracy differences as reflecting the combined effect of architecture design and training procedure.

3.3. Corruption Protocol and Calibration Perturbation Design

3.3.1. Environmental Corruption Simulation

The corruption protocol implements 10 perturbation types organized into four categories, applied at three severity levels (mild, moderate, severe). Table 2 specifies the corruption types, their parameters, and the sensor modalities affected.

Table 2. Corruption Types and Severity Level Parameters

Category	Corruption Type	Affected Modality	Mild	Moderate	Severe
Weather	Fog	Camera + LiDAR	Vis. 200m	Vis. 100m	Vis. 50m
Weather	Rain	Camera + LiDAR	10mm/h	25mm/h	50mm/h
Weather	Snow	Camera + LiDAR	Light	Moderate	Heavy
Sensor Noise	Motion Blur	Camera	σ=3px	σ=7px	σ=15px
Sensor Noise	Brightness	Camera	γ=0.7	γ=0.4	γ=0.2
Sensor Noise	Beam Missing	LiDAR	10% drop	30% drop	50% drop
Sensor Noise	Crosstalk	LiDAR	5% noise	15% noise	30% noise
Geometric	Spatial Misalign.	Cross-modal	0.1m	0.5m	1.0m
Geometric	Temporal Offset	Cross-modal	50ms	100ms	200ms
Geometric	FOV Limitation	LiDAR	270°	180°	90°

The weather corruption parameters are calibrated to match real-world meteorological conditions documented in the DENSE dataset. Camera corruptions are generated using standard image processing operations, while LiDAR corruptions are implemented through stochastic point removal and additive Gaussian noise. All corruption implementations are deterministic with fixed random seeds to ensure reproducibility.

3.3.2. Calibration Error and Temporal Misalignment Injection

Calibration perturbations are applied by introducing controlled offsets to the extrinsic transformation matrices between camera and LiDAR coordinate frames. Spatial

misalignment is implemented as isotropic Gaussian translations applied to the x, y, and z components of the translation vector, with standard deviations of 0.1m, 0.5m, and 1.0m corresponding to mild, moderate, and severe levels. Rotational perturbations are simultaneously applied with standard deviations of 0.1, 0.5, and 1.0 across roll, pitch, and yaw axes. Temporal desynchronization is simulated by substituting camera frames with those captured at specified temporal offsets (50ms, 100ms, 200ms) from the LiDAR sweep timestamp, preserving the original calibration matrices to isolate temporal from spatial effects (As shown in Figure 1).

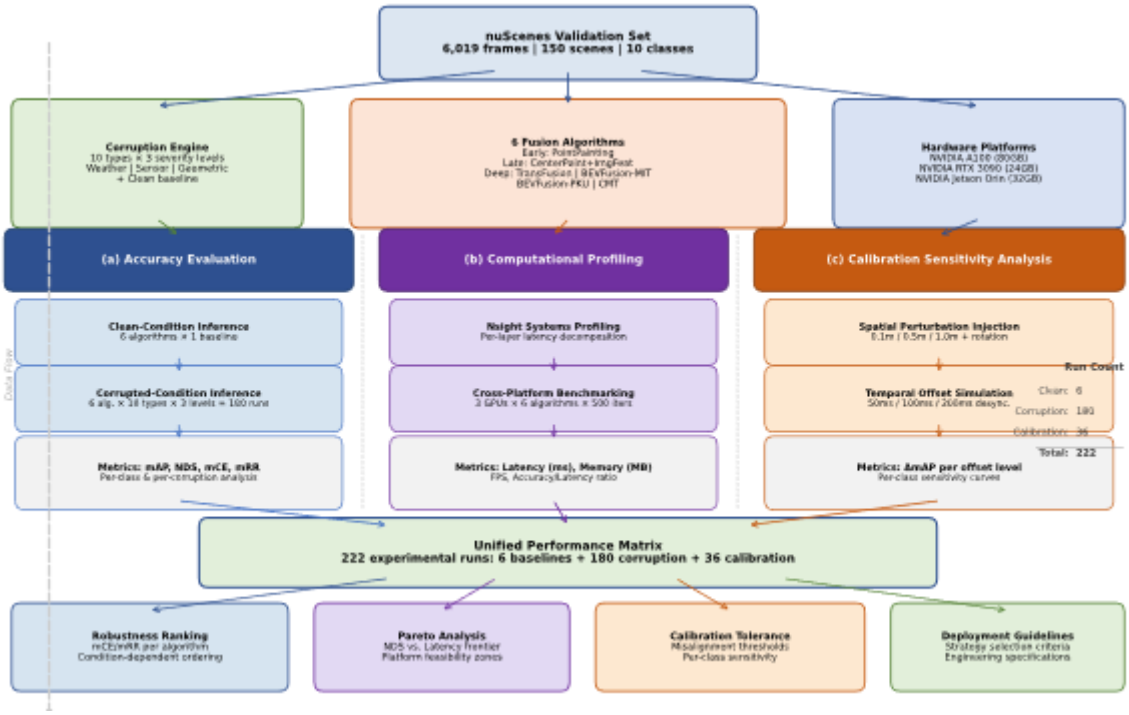

Figure 1. Experimental Framework Overview

Figure 1. Schematic diagram of the multi-dimensional benchmarking framework. The diagram illustrates three parallel evaluation pipelines: (a) accuracy evaluation under clean and corrupted conditions with mAP/NDS computation, (b) computational profiling with Nsight Systems for latency/memory measurement across three GPU platforms, and (c) calibration sensitivity analysis with controlled spatial and temporal perturbation injection. Input data flows from the nuScenes validation set through preprocessing, corruption application, and model inference, with results aggregated into unified performance matrices. The framework processes 6 fusion algorithms × 10 corruption types × 3 severity levels = 180 corruption evaluation configurations plus clean-condition baselines and calibration perturbation combinations, totaling 222 experimental runs.

4. Results and Analysis

4.1. Detection Performance under Clean and Corrupted Conditions

4.1.1. Clean-Condition Baseline Performance

Table 3 presents the detection accuracy of all evaluated fusion algorithms under clean (uncorrupted) conditions on the nuScenes validation split. The results establish baseline performance levels against which corruption-induced degradation is subsequently measured.

Table 3. 3D Object Detection Performance Under Clean Conditions (nuScenes val)

Algorithm	Fusion Category	mAP (%)	NDS (%)	mATE (m)	mASE	mAOE (rad)	mAVE (m/s)	mAAE

PointPainting	Early	58.2	65.8	0.312	0.264	0.368	0.287
CenterPoint+ImgFeat	Late	57.6	65.1	0.326	0.261	0.354	0.293
TransFusion	Deep	65.5	70.2	0.272	0.249	0.294	0.256
BEVFusion-MIT	Deep	68.5	71.4	0.261	0.252	0.281	0.243
BEVFusion-PKU	Deep	67.3	70.9	0.268	0.254	0.288	0.248
CMT	Deep	67.9	71.1	0.265	0.251	0.285	0.239

The deep fusion approaches consistently outperform early and late fusion methods, with BEVFusion-MIT achieving the highest mAP (68.5%) and NDS (71.4%). The performance gap between deep fusion and early/late fusion ranges from 2.8% to 5.1% NDS, confirming that intermediate-level feature interaction provides measurable accuracy benefits under nominal operating conditions. Notably, the late fusion variant CenterPoint+ImgFeat (65.1% NDS) slightly underperforms the LiDAR-only CenterPoint baseline (65.5% NDS reported in [7]), suggesting that naive score-level image fusion can introduce noise that marginally degrades the strong LiDAR-only detector; this observation is consistent with prior findings that poorly integrated modality streams may hurt rather than help detection. For cross-modality context, CRN (camera-radar) achieves 57.5% mAP and 62.4% NDS on the nuScenes test set as reported by FUTR3D [17]. Since these CRN numbers are from the test set rather than the validation set used in the present study, direct numerical comparison should be interpreted with caution; the gap of approximately 8.0%--9.0% NDS below camera-LiDAR deep fusion is provided as an approximate reference for the modality-level performance difference rather than a controlled comparison.

4.1.2. Robustness under Environmental Corruptions

The corruption analysis reveals asymmetric degradation patterns tied to fusion architecture. Bijelic et al [18]. demonstrated that camera-LiDAR fusion methods degrade when sensory streams are asymmetrically distorted, and the present results corroborate this across the evaluated algorithms. Under weather corruptions at severe level, all methods exhibit NDS degradation between 5.2 and 8.7 points, with deep fusion approaches showing marginally smaller degradation (5.2--6.8 points) than early fusion (7.4--8.7 points) due to learned feature-level compensation. Under LiDAR-specific corruptions (beam missing at 50% dropout, FOV limitation at 90), the degradation pattern reverses: late fusion retains 87.3% of clean NDS (mRR = 0.873), compared to 78.6%--82.1% for deep fusion and 75.8% for early fusion. This reversal is attributable to the modular independence of late fusion architectures, where the failure of one stream does not propagate into the other (As shown in Figure 2).

Figure 2. Performance Degradation Heatmap Under Environmental Corruptions

Figure 2. Heatmap visualization of NDS degradation (percentage points below clean-condition baseline) for six fusion algorithms across 10 corruption types at severe intensity. The horizontal axis represents 10 corruption types grouped by category (Weather: fog, rain, snow; Camera Noise: motion blur, brightness; LiDAR Noise: beam missing, crosstalk; Geometric: spatial misalignment, temporal offset, FOV limitation). The vertical axis lists the six algorithms. Cell colors range from dark green (minimal degradation, 0--3 points) through yellow (moderate, 3--8 points) to dark red (severe, >12 points). Numerical NDS degradation values are displayed within each cell. The heatmap reveals that weather corruptions produce relatively uniform degradation across all methods (5.2--8.7 points), while LiDAR-specific corruptions create the starkest inter-method divergence, with late fusion showing 4.1--7.3 points less degradation than deep fusion under severe beam missing.

Wang et al [19]. proposed the Stability Index (SI) metric evaluating temporal consistency, identifying that high-accuracy detectors can exhibit significant temporal instability. The present analysis aligns with this: BEVFusion-MIT, despite the highest clean NDS, shows the largest NDS variance (σ = 2.84 points) across corruption types, indicating that tight cross-modal coupling amplifies sensitivity to input perturbation diversity.

4.2. Computational Efficiency and Resource Utilization

Table 4 presents computational efficiency measurements across three hardware platforms for all evaluated algorithms.

Table 4. Computational Efficiency Metrics Across Hardware Platforms

Algorithm	A100 Latency (ms)	A100 Memory (MB)	RTX 3090 Latency (ms)	RTX 3090 FPS	Jetson Orin Latency (ms)	Jetson Orin FPS
PointPainting	78.3	4,218	98.7	10.1	287.4	3.5
CenterPoint+ImgFeat	62.1	3,542	81.4	12.3	213.6	4.7
TransFusion	89.2	5,631	112.8	8.9	342.1	2.9

BEVFusion-MIT	119.0	6,847	148.3	6.7	456.7	2.2
BEVFusion-PKU	108.6	6,234	135.2	7.4	412.3	2.4
CMT	95.4	5,892	121.6	8.2	367.8	2.7

Jin et al [20]. benchmarked LiDAR-only 3D detectors on Jetson platforms and reported that all boards consumed over 80% GPU resources on average, with TensorRT providing approximately 4× speedup. The present multi-modal evaluation extends these findings: Camera-LiDAR fusion algorithms require 1.4×--2.1× the latency of corresponding LiDAR-only baselines on Jetson Orin, with camera feature extraction and view transformation constituting 38%--52% of total inference time (As shown in Figure 3).

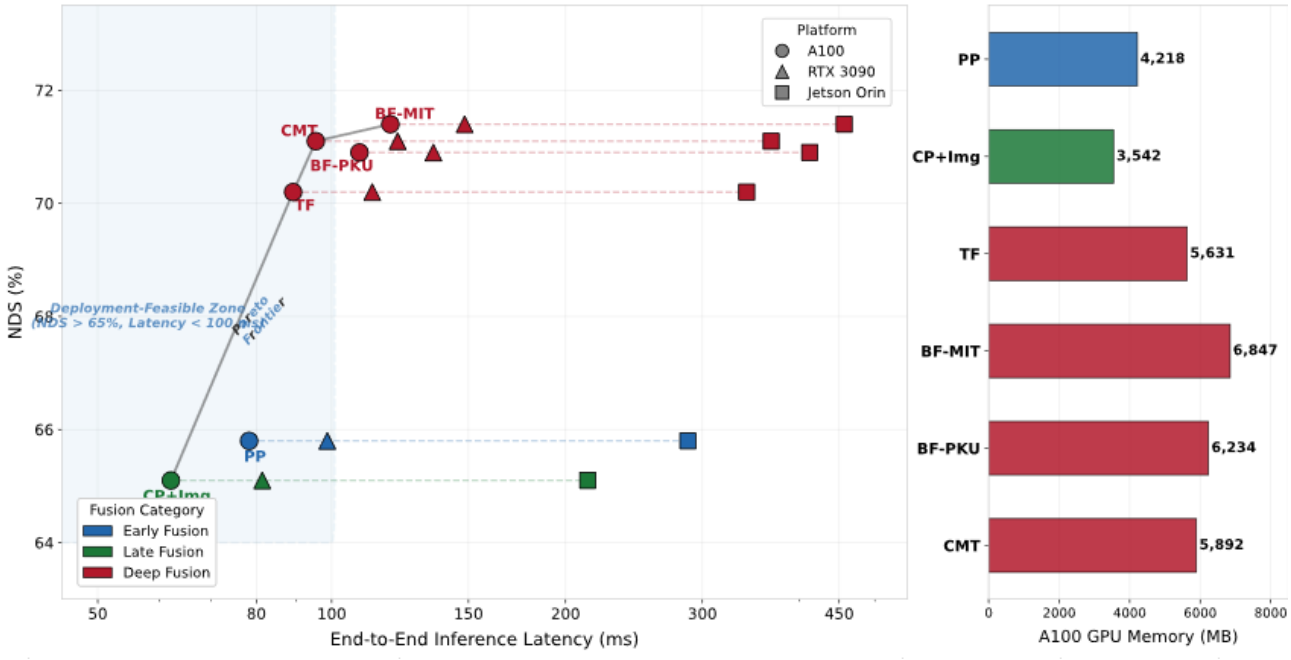

Figure 3. Accuracy-Efficiency Pareto Frontier

Scatter plot with dual-axis visualization showing the accuracy-efficiency trade-off across all evaluated methods on three hardware platforms. The horizontal axis represents end-to-end inference latency (ms, logarithmic scale), and the vertical axis represents NDS (%). Each algorithm is represented by three connected markers (circle for A100, triangle for RTX 3090, square for Jetson Orin) with color coding by fusion category (blue: early, green: late, red: deep). Dashed lines connect the same algorithm across platforms. The Pareto frontier connects CenterPoint+ImgFeat on A100 (65.1% NDS, 62.1ms), TransFusion on RTX 3090 (70.2% NDS, 112.8ms), and BEVFusion-MIT on A100 (71.4% NDS, 119.0ms). An annotated shaded region highlights the "deployment-feasible zone" (NDS > 65%, latency < 100ms), occupied on Jetson Orin only by late fusion. A secondary y-axis shows GPU memory utilization (MB) as horizontal bars.

The Pareto analysis identifies CenterPoint+ImgFeat as the only algorithm achieving real-time inference (>10 FPS) on all three platforms while maintaining NDS above 65%. The deep fusion methods occupy a higher-accuracy but computationally expensive region, with the accuracy-to-latency ratio (NDS per millisecond) decreasing from 1.05 for CenterPoint+ImgFeat to 0.60 for BEVFusion-MIT on A100. Qian et al [21]. demonstrated that LiDAR-radar deep fusion maintains radar's long-range advantage while suppressing false alarms, and the present results suggest that similar modality-complementary efficiency benefits may be achievable through selective feature routing in camera-LiDAR configurations.

4.3. Calibration Sensitivity and Temporal Synchronization Impact

4.3.1. Spatial Calibration Error Analysis

Calibration perturbation experiments reveal substantial differences in sensitivity across fusion categories. At the moderate perturbation level (0.5m spatial misalignment), early fusion (PointPainting) suffers an 8.7 percentage point mAP drop (from 58.2% to 49.5%), corresponding to a 14.9% relative degradation. Deep fusion methods exhibit category-internal variation: TransFusion degrades by only 5.1 percentage points (7.8% relative, from 65.5% to 60.4%), while BEVFusion-MIT degrades by 7.2 percentage points (10.5% relative, from 68.5% to 61.3%). CMT shows the smallest degradation at 4.6 percentage points (6.8% relative, from 67.9% to 63.3%), consistent with its coordinate-encoding design that avoids explicit geometric projection. Late fusion (CenterPoint+ImgFeat) degrades by 3.2 percentage points (5.6% relative), the smallest absolute degradation among all methods.

Ma et al [22]. identified that depth-free BEV transformation approaches exhibit enhanced out-of-distribution robustness, and the calibration results provide a specific instance: CMT's coordinate-encoding mechanism, which implicitly learns spatial relationships without explicit camera-to-LiDAR projection, achieves 41.2% less mAP degradation than PointPainting's hard-association approach at 0.5m misalignment. The relationship between calibration precision and detection performance is approximately linear for misalignments below 0.5m and exhibits accelerating degradation beyond this threshold for all methods except CMT and late fusion.

4.3.2. Temporal Misalignment Impact

Temporal desynchronization between camera and LiDAR captures introduces object localization errors proportional to vehicle and object velocities. At 100ms temporal offset, moving objects exhibit mean localization displacement of 1.4--2.8 meters at typical urban speeds (50--100 km/h), while static objects remain unaffected.

Deep fusion methods show mAP degradation of 3.8--5.2 percentage points at 100ms offset, concentrated in the vehicle (−4.1 to −6.3 points) and cyclist (−5.7 to −7.8 points) categories. Static object detection remains within 0.5 percentage points of clean-condition performance across all methods. Late fusion demonstrates relative resilience (2.1 points mAP degradation at 100ms), as independent per-modality detection reduces propagation of temporal inconsistencies. Early fusion exhibits the strongest sensitivity (6.4 points at 100ms), as the point decoration process directly couples temporal alignment quality to input feature integrity (As shown in Figure 4).

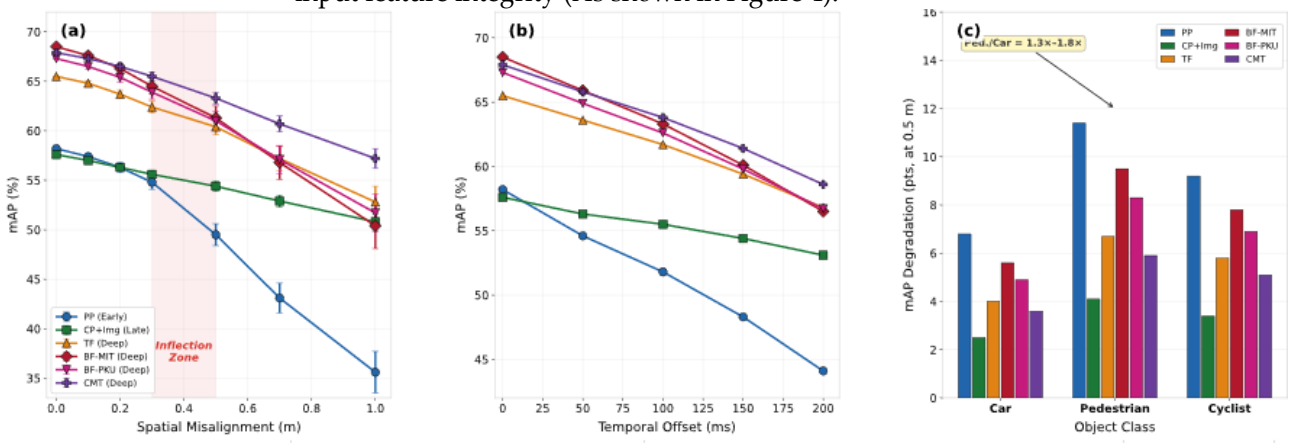

Figure 4. Calibration Sensitivity Curves

Multi-panel line chart showing detection performance degradation as a function of calibration perturbation magnitude. Panel (a) shows mAP versus spatial misalignment (0--1.0m), with each algorithm plotted as a separate line with distinct colors and markers, including ±1 standard deviation error bars. Panel (b) shows mAP versus temporal offset (0--200ms). Panel (c) presents a grouped bar chart showing per-class mAP degradation at 0.5m spatial misalignment for Car, Pedestrian, and Cyclist across all six methods. The spatial misalignment curves reveal two behavioral clusters: CMT and CenterPoint+ImgFeat maintain gradual linear degradation, while

PointPainting and BEVFusion-MIT exhibit an inflection point near 0.3--0.5m beyond which degradation accelerates. Pedestrian detection is the most calibration-sensitive class across all methods, with 1.3×--1.8× the degradation rate of Car detection.

5. Discussion and Conclusion

5.1. Discussion

The benchmarking results reveal three principal findings with implications for autonomous driving perception deployment. The performance ranking among fusion strategies is not invariant across operating conditions. Under nominal conditions, deep fusion achieves the highest detection accuracy (71.4% NDS for BEVFusion-MIT). Under LiDAR-degraded conditions, the ranking partially inverts: late fusion preserves 87.3% of clean performance (mRR = 0.873), outperforming deep fusion approaches that retain 78.6%--82.1%. This condition-dependent ranking implies that the optimal fusion strategy selection should be informed by the anticipated distribution of environmental conditions in the deployment domain.

The computational efficiency analysis identifies a substantial gap between server-grade and edge-deployment platforms. The latency scaling factor from A100 to Jetson Orin ranges from 2.9× to 3.8×, with camera feature extraction and BEV transformation contributing disproportionately to edge-platform latency. Fusion strategies relying on lightweight camera processing may be preferable for resource-constrained deployments, even at the cost of 1.4%--6.3% NDS relative to the most accurate deep fusion approaches.

The calibration sensitivity analysis provides actionable deployment specifications. The 0.5m spatial misalignment threshold, beyond which several methods exhibit accelerating degradation, establishes a practical calibration tolerance requirement. Engineering teams deploying early or BEV-based deep fusion methods should maintain calibration precision below 0.3m, while late fusion and coordinate-encoding approaches tolerate misalignment up to 0.8m with approximately linear degradation.

The scope of this study is bounded by several constraints. The evaluation is limited to the nuScenes dataset, which does not encompass all driving environments. The corruption protocol simulates perturbations through post-processing rather than capturing naturally degraded sensor data. The computational measurements reflect single-frame inference without accounting for temporal aggregation overhead. Future work extending this evaluation to additional datasets and incorporating naturally degraded sequences would strengthen the generalizability of the reported findings.

5.2. Limitations

This study evaluates fusion strategies under individual corruption types applied in isolation. Real-world driving conditions frequently involve compound corruptions (rain combined with reduced visibility, sensor noise combined with calibration drift), and the interaction effects of simultaneous perturbations remain uncharacterized. Extending the corruption protocol to include compound corruption combinations represents a priority direction for future investigation.

The present evaluation is restricted to the 3D object detection task. Autonomous driving perception encompasses additional tasks including BEV map segmentation and occupancy prediction, each of which may exhibit different robustness profiles under identical corruption conditions. Multi-task robustness evaluation would provide a more comprehensive understanding of fusion strategy trade-offs.

The emerging 4D imaging radar technology, providing dense point clouds approaching LiDAR-like spatial resolution at substantially lower cost, is anticipated to reshape the sensor fusion landscape. Benchmarking camera-4D radar fusion approaches against the camera-LiDAR results reported here would inform cost-optimized sensor configuration decisions for next-generation autonomous driving platforms.

References

1. H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuScenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11621–11631.
2. P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, ... D. Anguelov, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2443–2451.
3. D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2021.
4. Y. Dong, C. Kang, J. Zhang, Z. Zhu, Y. Wang, X. Yang, H. Su, X. Wei, and J. Zhu, "Benchmarking robustness of 3D object detection to common corruptions in autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 1024–1034.
5. L. Kong, Y. Liu, X. Li, R. Chen, W. Zhang, J. Ren, L. Pan, K. Chen, and Z. Liu, "Robo3D: Towards robust and reliable 3D perception against corruptions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 19994–20006.
6. S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "PointPainting: Sequential fusion for 3D object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4604–4612.
7. T. Yin, X. Zhou, and P. Krähenbühl, "Center-based 3D object detection and tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 11784–11793.
8. X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C. L. Tai, "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)*, 2022, pp. 1080–1089.
9. Y. Li, A. W. Yu, T. Meng, B. Caine, J. Ngiam, D. Peng, J. Shen, Y. Lu, D. Zhou, Q. V. Le, A. Yuille, and M. Tan, "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)*, 2022, pp. 17182–17191.
10. Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, and S. Han, "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)*, 2023, pp. 2774–2781.
11. T. Liang, H. Xie, K. Yu, Z. Xia, Z. Lin, Y. Wang, T. Tang, B. Wang, and Z. Tang, "BEVFusion: A simple and robust LiDAR-camera fusion framework," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022, pp. 10421–10434.
12. S. Xie, L. Kong, W. Zhang, J. Ren, L. Pan, K. Chen, and Z. Liu, "RoboBEV: Towards robust bird's eye view perception under corruptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 1, pp. 1–18, 2025.
13. K. Yu, T. Tao, H. Xie, Z. Lin, T. Liang, B. Wang, P. Chen, D. Hao, Y. Wang, and X. Liang, "Benchmarking the robustness of LiDAR-camera fusion for 3D object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 3187–3197.
14. T. Beemelmanns, Q. Zhang, C. Geller, and L. Eckstein, "MultiCorrupt: A multi-modal robustness dataset and benchmark of LiDAR-camera fusion for 3D object detection," in *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, 2024, pp. 1–8.
15. J. Yan, Y. Liu, J. Sun, F. Jia, S. Li, T. Wang, and X. Zhang, "Cross modal transformer: Towards fast and robust 3D object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 18268–18278.
16. Y. Kim, J. Shin, S. Kim, I. J. Lee, J. W. Choi, and D. Kum, "CRN: Camera radar net for accurate, robust, efficient 3D perception," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 17615–17626.
17. X. Chen, T. Zhang, Y. Wang, Y. Wang, and H. Zhao, "FUTR3D: A unified sensor fusion framework for 3D detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 172–181.
18. M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11682–11692.
19. J. Wang, Q. Meng, G. Liu, L. Yan, K. Wang, M. M. Cheng, and Q. Hou, "Towards stable 3D object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024, pp. 1–17.
20. S. Jin, J. Park, J. Lee, H. Lee, and S. Lee, "Run your 3D object detector on NVIDIA Jetson platforms: A benchmark analysis," *Sensors*, vol. 23, no. 8, p. 4005, 2023.
21. K. Qian, S. Zhu, X. Zhang, and L. E. Li, "Robust multimodal vehicle detection in foggy weather using complementary lidar and radar signals," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 444–453.
22. Y. Ma, T. Wang, X. Bai, H. Yang, Y. Hou, Y. Wang, Y. Qiao, R. Yang, D. Meng, and Z. Li, "Vision-centric BEV perception: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10978–10997, 2024.