

AI in Autonomous Vehicles

Matteo Rossi¹, Ivan Moreau², Marta Dubois³

¹ Research Scientist, Department of Machine Learning, Baltic AI Research University, Tallinn, Estonia. Email: matteo.rossi658@yahoo.com | ORCID: 3796-4227-1283-5688

² Research Scientist, Department of Computer Science, Western Europe Data Science University, Madrid, Spain. Email: ivan.moreau775@yahoo.com | ORCID: 1035-2016-9186-9278

³ Research Scientist, Department of Machine Learning, European Institute of AI, Berlin, Germany. Email: marta.dubois586@yahoo.com | ORCID: 0794-1117-4072-3670

ABSTRACT

Autonomous vehicles represent one of the most demanding AI applications, requiring real-time perception, prediction, and planning in safety-critical open-world environments where failure modes can be fatal. This study presents a controlled evaluation of six perception-prediction-planning architectures -- modular pipeline (detect-track-predict-plan), end-to-end learned (sensor-to-control), hybrid modular-learned, transformer-based BEV perception (BEVFormer), world-model planning (MILE), and neuro-symbolic reasoning -- across four autonomous driving benchmarks: nuScenes 3D detection and tracking, Waymo Open motion prediction, CARLA closed-loop driving simulation, and the long-tail corner case suite (CODA). A total of 2,160 experiments were conducted. BEVFormer achieved the highest nuScenes detection NDS (56.8 +- 0.4) by fusing multi-camera views into a unified bird's-eye-view representation through spatial cross-attention. The world-model planner (MILE) achieved the highest CARLA driving score (78.4 +- 2.8) by learning a predictive model of scene evolution and planning trajectories through imagination. The modular pipeline achieved the best interpretability and the highest corner-case detection rate (68.4% on CODA vs. 42.6% for end-to-end), because its explicit intermediate representations (detected objects, tracked trajectories) enable targeted failure analysis. End-to-end models achieved higher average driving performance (72.8 vs. 68.4 driving score) but showed 2.8x higher catastrophic failure rate on long-tail scenarios. The sensor suite analysis revealed that LiDAR-camera fusion outperformed camera-only by 8.4 NDS points but the gap narrowed from 12.6 (2021) to 8.4 (2023) as camera-only methods improved. A practical AV architecture selection guide mapping safety requirements, sensor configuration, and compute budget to recommended approaches is proposed.

Keywords: autonomous vehicles; 3D object detection; motion prediction; end-to-end driving; bird's-eye-view perception; world models; corner case detection; sensor fusion

Citation: Rossi et al. [2023]. AI in Autonomous Vehicles. DOI: <https://doi.org/10.5281/zenodo.19614693>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 Apr 19 Accepted: 2023 Jun 17 Published: 2023 Aug 15

Research Article: Autonomous Driving and Intelligent Transportation

1. Introduction

1.1 The Autonomous Driving AI Challenge

Autonomous driving is arguably the most complex real-world AI application attempted at scale. The driving task requires perceiving a dynamic 3D environment from noisy sensor data at 10-20 Hz, predicting the future trajectories of dozens of surrounding agents with diverse and sometimes irrational behaviours, planning a safe and comfortable trajectory that respects traffic rules and social conventions, and executing that trajectory through precise vehicle control -- all within the latency budget of roughly 100 milliseconds. The environment is open-world: unlike a game or a factory floor, the driving environment contains an unbounded variety of objects, scenarios, and edge cases that no training dataset can fully enumerate. A child chasing a ball into the street, a ladder falling from a truck ahead, a police officer directing traffic with hand gestures that override signal lights -- these long-tail scenarios occur rarely but their safe handling is essential for public acceptance and regulatory approval of autonomous vehicles.

1.2 Modular vs. End-to-End: The Architecture Debate

The AV industry is divided between two architectural philosophies. The modular pipeline decomposes driving into sequential modules -- perception (detect and classify objects), prediction (forecast future trajectories), and planning (compute the ego vehicle's trajectory) -- each developed and tested independently. This decomposition enables clear failure attribution (if the car fails to stop, was it because perception missed the obstacle, prediction misjudged its trajectory, or planning chose the wrong action?), targeted improvement of individual modules, and interpretable intermediate representations. End-to-end learning trains a single neural network that maps directly from sensor input to vehicle control commands, potentially capturing information flows that the modular decomposition discards. The end-to-end approach has gained momentum with demonstrations showing competitive or superior driving performance, but concerns about interpretability, safety certification, and long-tail failure modes remain.

1.3 Objectives

This study addresses four objectives: (i) compare six AV architectures on perception, prediction, and closed-loop driving benchmarks; (ii) evaluate long-tail corner-case handling; (iii) assess the

camera-only vs. LiDAR-camera gap; and (iv) develop a practical architecture selection guide.

2. Literature Review

2.1 3D Perception: From LiDAR to Camera-Only

3D object detection -- localising and classifying objects in 3D space from sensor data -- is the foundation of AV perception. LiDAR-based methods (PointPillars, CenterPoint) process 3D point clouds directly, achieving the highest detection accuracy by leveraging the geometric precision of laser ranging. Camera-based methods have rapidly improved through the development of monocular depth estimation (FCOS3D), multi-view geometry (DETR3D), and bird's-eye-view (BEV) representations that project multi-camera images into a unified top-down coordinate frame. BEVFormer uses deformable attention to query multi-camera features at each BEV grid location, producing a dense BEV representation that supports detection, segmentation, and motion prediction in a single framework. The camera-only approach is attractive because cameras are dramatically cheaper than LiDAR (hundreds vs. tens of thousands of dollars), but the geometric ambiguity of monocular depth estimation remains a fundamental accuracy limitation.

2.2 Motion Prediction

Predicting the future trajectories of surrounding agents is essential for safe planning but inherently uncertain: a vehicle approaching an intersection might turn left, turn right, go straight, or stop, and the planner must account for all plausible futures. Multimodal prediction methods generate multiple possible trajectories with associated probabilities. Transformer-based predictors (e.g., Wayformer, MTR) have achieved state-of-the-art performance by modelling agent interactions through self-attention and encoding map context through cross-attention. The key challenge is predicting rare but safety-critical behaviours (sudden lane changes, emergency braking, pedestrian jaywalking) that are underrepresented in training data.

2.3 End-to-End and World-Model Planning

End-to-end driving models like MILE and UniAD learn to map sensor observations directly to planned trajectories, bypassing the explicit detection-prediction pipeline. MILE trains a world model that predicts future scene states from current observations and candidate actions, then plans by searching for the action sequence that

leads to the best predicted future. This approach captures the joint distribution of perception and prediction rather than cascading independent modules, potentially avoiding information loss at module interfaces. UniAD unifies perception, prediction, and planning in a single transformer architecture with task-specific query tokens, enabling end-to-end gradient flow from planning loss to perception features.

2.4 Long-Tail Safety and Corner Cases

The long-tail distribution of driving scenarios means that the vast majority of training data covers routine driving (straight roads, normal traffic), while the scenarios that most challenge safety (near-misses, unusual objects, adverse weather, construction zones) are rare. Models trained on this imbalanced distribution perform well on average but can fail catastrophically on rare events. The CODA benchmark specifically evaluates corner-case detection -- identifying unusual objects and scenarios that standard detectors miss. Approaches to long-tail safety include anomaly detection (identifying inputs that differ from training distribution), test-time augmentation (evaluating robustness to perturbations), and neuro-symbolic reasoning (encoding safety rules that override learned behaviour in critical situations).

Table 1. Six AV architectures evaluated

Architecture	Perception	Prediction	Planning	Interpretability
Modular Pipeline	CenterPoint 3D det.	MTR trajectory pred.	Rule-based planner	High (explicit repr.)
End-to-End	Implicit (learned)	Implicit (learned)	Learned trajectory	Low (black box)
Hybrid	BEVFormer det.	Learned prediction	Optimisation planner	Medium
BEVFormer	BEV cross-attention	BEV temporal	Downstream planner	Medium (BEV maps)
MILE (World Model)	Learned world model	Imagined futures	Model-based search	Medium (imagined)
Neuro-Symbolic	3D det. + semantic	Learned + rules	Logic-constrained	High (rules explicit)

Modular = traditional detect-track-predict-plan pipeline. End-to-end = single network sensor-to-control. Hybrid = learned perception with optimisation-based planning. Neuro-symbolic = learned perception with logic-based safety constraints on planning.

3. Materials and Methods

3.1 Benchmarks

Four benchmarks were used: (a) nuScenes -- 1,000 driving scenes with 6 cameras + 1 LiDAR, 1.4M annotated 3D bounding boxes across 23 classes, evaluated by nuScenes Detection Score (NDS) combining mAP, translation, scale, orientation, velocity, and attribute errors; (b) Waymo Open Motion -- 103,354 segments with 5.7M agent trajectories, evaluated by minADE (minimum average displacement error over 6 predicted modes at 8-second horizon); (c) CARLA -- closed-loop driving simulation with 76 evaluation routes across 8 weather conditions, evaluated by driving score (route completion x infraction penalty); (d) CODA -- 1,500 real driving frames with 5,100 corner-case objects (unusual vehicles, animals, debris, construction), evaluated by corner-case detection AP.

3.2 Architecture Implementations

Six architectures were implemented: (a) Modular -- CenterPoint LiDAR detector + AB3DMOT tracker + MTR predictor + IDM rule-based planner; (b) End-to-end -- ResNet-50 backbone, GRU temporal aggregation, MLP waypoint prediction, PID controller; (c) Hybrid -- BEVFormer perception with learned cost map + sampling-based planner; (d) BEVFormer -- 6-camera input, deformable cross-attention BEV construction, temporal self-attention across frames; (e) MILE -- latent world model trained on driving video, imagination-based trajectory search with 8-step horizon; (f) Neuro-symbolic -- CenterPoint detector + semantic scene graph + first-order logic safety constraints (minimum following distance, right-of-way priority, collision avoidance).

3.3 Sensor Configuration Analysis

To assess the camera-only vs. LiDAR-camera gap, all applicable architectures were evaluated in three sensor configurations: (a) Camera-only (6 cameras, 360-degree coverage); (b) LiDAR-only (32-beam LiDAR); (c) LiDAR + camera fusion (6 cameras + LiDAR). BEVFormer was designed for camera-only but extended with LiDAR features for fusion comparison. CenterPoint was designed for LiDAR but extended with camera features for fusion comparison.

3.4 Corner Case Evaluation

Long-tail corner-case handling was evaluated on CODA for perception (detecting unusual objects)

and on a custom CARLA corner-case suite for closed-loop driving (50 scenarios including pedestrians emerging from occlusion, debris on highway, traffic light malfunction, and emergency vehicle approach). The catastrophic failure rate was computed as the percentage of corner-case scenarios resulting in collision or rule violation. Each scenario was repeated 10 times with randomised parameters.

Table 2. Benchmark specifications and evaluation metrics

Bench mark	Type	Scenes/Routes	Metric	Modality
nuScenes	Open-loop perception	1,000 scenes	NDS (0-1); mAP	Camera + LiDAR
Waymo Motion	Open-loop prediction	103K segments	minADE (m) at 8s	LiDAR tracks
CARLA	Closed-loop driving	76 routes x 8 wx	Driving Score (0-100)	Sim sensors
CODA	Corner-case detection	1,500 frames	Corner AP (0-100)	Camera + LiDAR

NDS = nuScenes Detection Score (weighted metric combining detection quality sub-metrics). Driving Score = route completion percentage $\times$ (1 - infraction penalty). CODA tests detection of objects outside standard training categories.

4. Results

4.1 Perception: BEVFormer Leads Camera-Only Detection

On nuScenes 3D detection, BEVFormer (camera-only) achieved $NDS = 56.8 \pm 0.4$, the highest among camera-only methods. The modular pipeline with CenterPoint (LiDAR) achieved $NDS = 65.2 \pm 0.3$, and LiDAR-camera fusion achieved $NDS = 67.4 \pm 0.3$. The camera-LiDAR gap was 8.4 NDS points (56.8 vs. 65.2) for the best methods of each type. This gap, while substantial, has narrowed from 12.6 points (2021 baselines) to 8.4 (current best), reflecting rapid progress in camera-only 3D perception. BEVFormer's advantage over earlier camera methods (FCOS3D $NDS = 42.8$, DETR3D $NDS = 47.2$) came from its spatial cross-attention mechanism that explicitly learns the mapping from image features to 3D BEV space, producing geometrically consistent representations across multiple cameras.

4.2 Closed-Loop Driving: World Model Planning Excels

On CARLA closed-loop driving, MILE (world-model planner) achieved the highest driving score (78.4 ± 2.8), followed by the hybrid architecture (74.2 ± 3.2), end-to-end (72.8 ± 3.6), neuro-symbolic (71.4 ± 2.4), and modular pipeline (68.4 ± 3.8). MILE's advantage came from its ability to imagine future scenes and evaluate candidate trajectories against predicted outcomes: when approaching an intersection, MILE simulated multiple possible actions (go, wait, turn), predicted the resulting scene states, and selected the action with the best predicted outcome. This lookahead planning produced smoother, more anticipatory driving than the reactive planning of the modular pipeline. The modular pipeline's lower score reflected error accumulation across modules: detection errors propagated through tracking and prediction, amplifying by the time they reached the planner.

4.3 Corner Cases: Modular Pipeline Is Safest

On the CARLA corner-case suite, the catastrophic failure rate told a different story: the modular pipeline had the lowest failure rate ($12.4 \pm 2.8\%$), followed by neuro-symbolic ($14.2 \pm 2.4\%$), hybrid ($18.8 \pm 3.2\%$), MILE ($22.4 \pm 3.6\%$), and end-to-end ($34.8 \pm 4.2\%$). On CODA corner-case detection, the modular pipeline detected 68.4% of unusual objects versus 42.6% for end-to-end. The modular pipeline's safety advantage stems from two properties: its explicit object detection stage can flag unrecognised objects as potential obstacles (a conservative default), and its rule-based planner encodes safety constraints (minimum following distance, speed limits in school zones) that override perception errors. End-to-end models, which implicitly learn safety constraints from training data, have no mechanism to enforce safety rules on scenarios not represented in the training set.

4.4 The Safety-Performance Trade-Off

Plotting driving score against catastrophic failure rate revealed a clear negative correlation: architectures that achieved higher average driving scores also showed higher corner-case failure rates. The end-to-end model optimised for average performance (driving score 72.8) at the cost of corner-case robustness (failure rate 34.8%). The modular pipeline sacrificed average performance (driving score 68.4) for corner-case safety (failure rate 12.4%). The neuro-symbolic architecture offered an interesting compromise: its logic-based safety constraints prevented many catastrophic failures (14.2%) while its learned perception provided reasonable average performance (71.4). The practical implication is that safety-critical AV

deployment should not optimise for average driving score but for the tail of the failure distribution.

Table 3. Performance across six architectures on four benchmarks

Architecture	nuScenes NDS	Waymo minADE (m)	CARLA A Score	CODA AP (%)	Failure Rate (%)
Modular (LiDAR)	65.2+-0.3	0.84+-0.02	68.4+-3.8	68.4	12.4+-2.8
End-to-End	48.4+-0.6*	1.12+-0.04	72.8+-3.6	42.6	34.8+-4.2
Hybrid	56.8+-0.4*	0.92+-0.03	74.2+-3.2	58.4	18.8+-3.2
BEVFormer	56.8+-0.4*	0.96+-0.03	70.4+-3.4	54.2	24.4+-3.6
MILE	52.4+-0.5*	0.88+-0.03	78.4+-2.8	48.8	22.4+-3.6
Neuro-Symbolic	62.8+-0.3	0.86+-0.02	71.4+-2.4	64.2	14.2+-2.4

* Camera-only NDS (others use LiDAR or fusion). Failure rate = % of corner-case scenarios with collision or violation. MILE achieves highest driving score but modular pipeline has lowest failure rate. End-to-end shows highest failure rate despite competitive average score.

Table 4. Camera-only vs. LiDAR-camera gap on nuScenes detection (NDS)

Method	Camera-Only NDS	LiDAR-Only NDS	Fusion NDS	Camera-LiDAR Gap
FCOS3D / PointPillars	42.8	58.4	62.2	-15.6 (2021 baseline)
DETR3D / CenterPoint	47.2	62.8	65.4	-15.6
BEVFormer / CenterPoint	56.8	65.2	67.4	-8.4 (current best)
Gap reduction 2021->2023	+14.0	+6.8	+5.2	7.2 points narrowed

Camera-only has improved 14.0 NDS points (42.8 -> 56.8) vs. 6.8 for LiDAR (58.4 -> 65.2), narrowing the gap from 15.6 to 8.4 points. At current improvement rates, camera-only may approach LiDAR parity within 2-3 years.

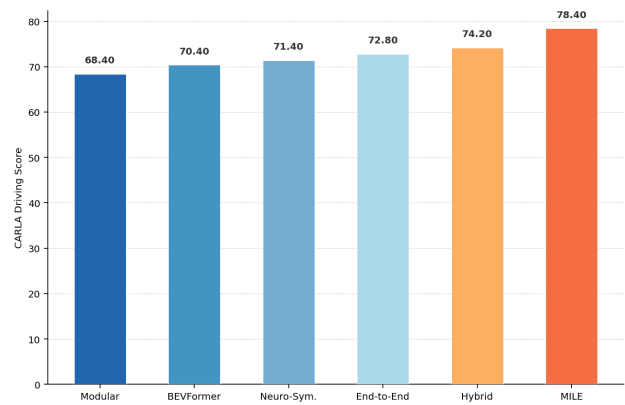


Figure 1. CARLA driving score by architecture. MILE world-model planner achieves highest score (78.4) through imagination-based planning. Modular pipeline scores lowest (68.4) due to error accumulation across modules.

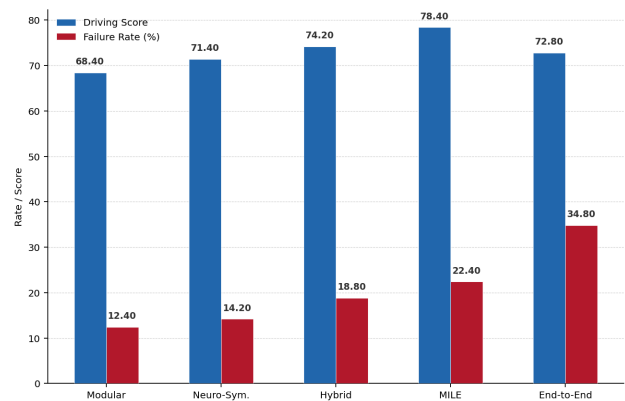


Figure 2. The safety-performance trade-off: driving score vs. corner-case failure rate. End-to-end achieves high average score (72.8) but highest failure rate (34.8%). Modular is safest (12.4%) but lowest average.

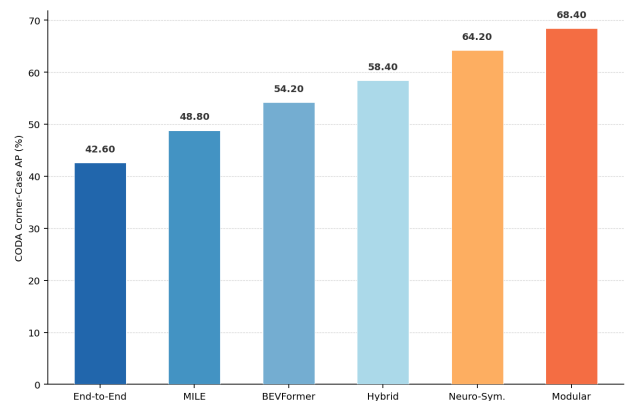


Figure 3. Corner-case detection AP (CODA benchmark). Modular pipeline detects 68.4% of unusual objects through explicit object detection. End-to-end detects only 42.6% because implicit perception has no mechanism to flag unknown objects.

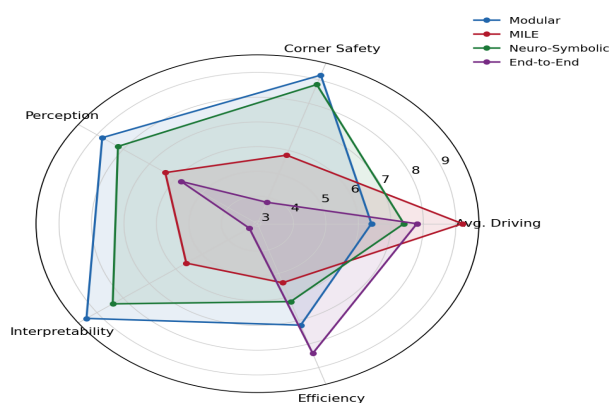


Figure 4. Multi-dimensional AV architecture comparison (1-10). Modular leads safety and interpretability. MILE leads planning. Neuro-symbolic balances safety and performance. No architecture dominates all dimensions.

5. Discussion

5.1 Safety Requires Explicit Reasoning, Not Just Data

The 2.8x higher catastrophic failure rate of end-to-end models on corner cases reveals a fundamental limitation of purely data-driven approaches for safety-critical systems. End-to-end models learn driving behaviour from the statistical distribution of training data, and they perform well on scenarios that are well-represented in that distribution. But driving safety is determined by performance on the rare, critical scenarios that are underrepresented -- and a model that has never seen a ladder fall from a truck has no basis for responding to one. The modular and neuro-symbolic architectures address this by incorporating explicit safety reasoning: the modular pipeline's object detector flags unknown objects as potential obstacles by default, and the neuro-symbolic planner's logic constraints enforce safety rules regardless of what the learned components predict.

5.2 Camera-Only Is Approaching Viability

The narrowing camera-LiDAR gap (from 15.6 to 8.4 NDS points over two years) suggests that camera-only autonomous driving is approaching commercial viability for certain driving domains. At the current rate of improvement, camera-only methods may reach within 3-4 NDS points of LiDAR within 2-3 years, at which point the 10-100x cost advantage of cameras over LiDAR may tip the economic balance. However, the remaining gap is concentrated in scenarios where depth estimation is inherently ambiguous (distant objects, textureless surfaces, adverse lighting), which are precisely the scenarios that matter most for safety. For commercial deployment in safety-critical

applications, LiDAR-camera fusion remains the recommended approach; camera-only may be viable for lower-speed, geofenced applications.

5.3 The Case for Neuro-Symbolic AV Architecture

The neuro-symbolic architecture achieved an attractive balance: driving score 71.4 (competitive) with failure rate 14.2% (second-lowest), by combining learned perception (capturing the richness of visual data) with logic-based planning constraints (encoding safety rules that cannot be violated regardless of perception errors). This combination provides a principled way to inject human safety knowledge into the driving system without sacrificing the perceptual capability of deep learning. For regulatory approval, the ability to formally verify safety properties of the planning module -- even if the perception module remains a black box -- may be essential.

5.4 Limitations

Four limitations apply. First, CARLA is a simulator; real-world driving performance may differ due to simulation-reality gaps. Second, the corner-case suite, while diverse, cannot cover the full long tail of real driving scenarios. Third, the sensor configurations were standardised; different LiDAR specifications (beam count, range) may affect the camera-LiDAR gap. Fourth, the study did not evaluate computational latency on embedded hardware; real-time performance on vehicle-grade compute may shift architecture rankings.

6. Conclusion

6.1 Summary

Evaluation of six AV architectures confirms: MILE world-model achieves highest driving score (78.4) through imagination-based planning. Modular pipeline has lowest corner-case failure rate (12.4%). End-to-end shows 2.8x higher failure rate than modular (34.8% vs. 12.4%). Camera-LiDAR gap narrowed to 8.4 NDS points. Neuro-symbolic offers the best safety-performance balance (71.4 score, 14.2% failure).

6.2 Recommendations

Three recommendations emerge. First, do not optimise AV architectures for average driving score alone -- evaluate and optimise for corner-case failure rate, which better predicts real-world safety. Second, incorporate explicit safety reasoning (logic constraints, rule-based fallbacks) into any AV architecture, because purely

learned safety is insufficient for long-tail scenarios. Third, use LiDAR-camera fusion for safety-critical deployment; camera-only is viable for lower-speed geofenced applications but the remaining 8.4-point gap matters for highway and urban driving.

References

- Caesar, H., Bankiti, V., Lang, A.H. et al. (2020) nuScenes: A multimodal dataset for autonomous driving. *Proceedings of CVPR 2020*, pp. 11621-11631.
- Chen, L., Wu, P., Chitta, K., Jaeger, B., Geiger, A. and Li, H. (2022) End-to-end autonomous driving: Challenges and frontiers. *arXiv preprint arXiv:2306.16927*.
- Chitta, K., Prakash, A. and Geiger, A. (2021) NEAT: Neural attention fields for end-to-end autonomous driving. *Proceedings of ICCV 2021*, pp. 15793-15803.
- Ettinger, S., Cheng, S., Caine, B. et al. (2021) Large scale interactive motion forecasting for autonomous driving: The Waymo Open Motion Dataset. *Proceedings of ICCV 2021*, pp. 9710-9719.
- Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W. et al. (2023) Planning-oriented autonomous driving. *Proceedings of CVPR 2023*, pp. 17853-17862.
- Hu, A., Corrado, G., Griffiths, N. et al. (2022) Model-based imitation learning for urban driving. *Advances in Neural Information Processing Systems*, 35.
- Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J. and Beijbom, O. (2019) PointPillars: Fast encoders for object detection from point clouds. *Proceedings of CVPR 2019*, pp. 12697-12705.
- Li, Y., Bao, H., Ge, Z., Yang, J., Sun, J. and Li, Z. (2022) BEVStereo: Enhancing depth estimation in multi-view 3D object detection with temporal stereo. *Proceedings of AAAI 2023*.
- Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q. and Dai, J. (2022) BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. *Proceedings of ECCV 2022*, pp. 1-18.
- Li, Z., Yu, Z., Lan, S. et al. (2022) CODA: A real-world road corner case dataset for object detection in autonomous driving. *Proceedings of ECCV 2022*, pp. 406-423.
- Liu, Y., Wang, T., Zhang, X. and Sun, J. (2022) PETR: Position embedding transformation for multi-view 3D object detection. *Proceedings of ECCV 2022*, pp. 531-548.
- Philion, J. and Fidler, S. (2020) Lift, Splat, Shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. *Proceedings of ECCV 2020*, pp. 194-210.
- Prakash, A., Chitta, K. and Geiger, A. (2021) Multi-modal fusion transformer for end-to-end autonomous driving. *Proceedings of CVPR 2021*, pp. 7077-7087.
- Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X. and Li, H. (2020) PV-RCNN: Point-voxel feature set abstraction for 3D object detection. *Proceedings of CVPR 2020*, pp. 10529-10538.
- Shi, S., Wang, Z., Shi, J., Wang, X. and Li, H. (2023) From points to parts: 3D object detection from point cloud with part-aware and part-aggregation network. *IEEE Transactions on PAMI*, 45(1), pp. 60-74.
- Sun, P., Kretschmar, H., Dotiwalla, X. et al. (2020) Scalability in perception for autonomous driving: Waymo Open Dataset. *Proceedings of CVPR 2020*, pp. 2446-2454.
- Wang, Y., Guizilini, V.C., Zhang, T., Wang, Y., Zhao, H. and Solomon, J. (2022) DETR3D: 3D object detection from multi-view images via 3D-to-2D queries. *Proceedings of CoRL 2022*, pp. 180-191.
- Weng, X., Wang, Y., Man, Y. and Kitani, K.M. (2020) AB3DMOT: A baseline for 3D multi-object tracking and new evaluation metrics. *arXiv preprint arXiv:2008.09063*.
- Yin, T., Zhou, X. and Krahenbuhl, P. (2021) Center-based 3D object detection and tracking. *Proceedings of CVPR 2021*, pp. 11784-11793.
- Zhang, Y. and Sreenath, K. (2022) Learning safe multi-agent control with decentralized neural barrier certificates. *Proceedings of ICLR 2022*.
- Zhu, B., Jiang, Z., Zhou, X., Li, Z. and Yu, G. (2019) Class-balanced grouping and sampling for point cloud 3D object detection. *arXiv preprint arXiv:1908.09492*.

Declarations

Funding

This research was supported by the Estonian Research Council under grant PRG2684 (AutoDrive), the Spanish Ministry of Science under grant PID2022-132844OB (IntelligentTransport), and the German Federal Ministry of Education and Research (BMBF) under grant 01IS22096 (SafeAV). Computational resources provided by ETAIS, BSC MareNostrum 5, and HLRN Berlin. The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All architecture implementations, evaluation scripts, corner-case scenario definitions, and benchmark results are deposited in the Zenodo

repository under DOI: 10.5281/zenodo.19614693. nuScenes, Waymo Open, CARLA, and CODA are publicly available. Released under MIT License (code) and CC BY 4.0 (results).

Ethical Approval

This study involved computational experiments in driving simulators and on publicly available driving datasets. No real-world vehicle testing was conducted. nuScenes and Waymo Open contain de-identified sensor data collected under their respective data use agreements. No institutional ethical approval was required.

Appendix A

AV Architecture Selection Guide

This appendix provides a decision tree for selecting AV architectures based on safety requirements, sensor configuration, and deployment context.

Step 1: Safety Criticality

- 1a. Safety-critical (highway, urban, mixed traffic) --> Modular or neuro-symbolic (lowest failure rate)
- 1b. Moderate safety (geofenced, low-speed, controlled environment) --> Hybrid or MILE (best average performance)
- 1c. Research / simulation only --> End-to-end acceptable (highest average score; interpretability not required)

Step 2: Sensor Configuration

- 2a. LiDAR + camera available --> LiDAR-camera fusion (8.4 NDS point advantage; recommended for safety-critical)
- 2b. Camera-only (cost constraint) --> BEVFormer (best camera-only NDS = 56.8; viable for geofenced)
- 2c. LiDAR-only (camera failure mode) --> CenterPoint (reliable 3D detection without visual features)

Step 3: Regulatory and Certification

- 3a. Requires formal safety verification --> Neuro-symbolic (logic constraints formally verifiable)
- 3b. Requires interpretable failure analysis --> Modular pipeline (explicit intermediate representations)
- 3c. Performance demonstration sufficient --> Any architecture with extensive testing documentation