ACADEMIA International Journal for Social Sciences
Volume 5, Issue 3(s11), 2026 ISSN-L (Online): 3006-6638

From CNNs To VLMs: A Complete Review And Experimental Comparison of AI Models on Autonomous Vehicles

Asma Jabeen
asma.jabeen@must.edu.pk
Department of CS & IT, MUST

Hasan Asif
hassanasif2003@gmail.com
Department of Software Engineering, MUST

Faisal Riaz
faisal.riaz@must.edu.pk
Department of CS & IT, MUST

Corresponding Author: Asma Jabeen asma.jabeen@must.edu.pk

Received: 03-02-2026 **Revised:** 27-02-2026 **Accepted**: 05-03-2026 **Published:** 21-03-2026

ABSTRACT

This paper presents a critical analysis and experimental comparison of three AI paradigms in autonomous driving: convolutional neural networks (CNNs), large language models (LLMs), and vision-language models (VLMs). We reviewed more than thirty published papers spanning 2015- 2024 and conducted a practical benchmark using ResNet-50 (CNN), Flan-T5-base (LLM), and BLIP-2 OPT-2.7B (VLM) on 10 real-world driving scenes across 7 assessment categories, yielding 70 question-answer pairs. Our experiments show that VLMS achieve an overall accuracy of 84.3% (59/70), substantially outperforming CNN (61.4%, 43/70) and LLM (60.0%, 42/70) in visual object detection and natural-language scene understanding. The results indicate that VLMs represent the most promising next-generation paradigm for autonomous driving, unifying perception, reasoning, and explainability within a single model.

Keywords: *Autonomous Vehicles, CNN, LLM, Vision Language Models, BLIP-2, Scene Understanding, Explainable AI.*

INTRODUCTION

Autonomous driving requires AI systems capable of perceiving road conditions, reasoning about complex traffic scenarios, and making safe real-time decisions [1]. Over the past decade, three generations of paradigms have emerged: convolutional neural networks (CNNs) for visual perception [2][3], large language models (LLMs) for reasoning [6][8], and vision-language models (VLMs) that integrate both modalities into a unified architecture [10][11]. Knowing where each paradigm succeeds and where it falls short is essential for guiding future research and development.

CNNs such as YOLO [2], Faster R-CNN [3], and SSD [4] enabled real-time object detection and lane recognition, but inherently lack reasoning ability and interpretability [5]. These networks extract hierarchical features from pixel data and match them to learned patterns, effective for detecting individual objects yet insufficient for comprehending the broader context of a driving scene or making high-level navigation decisions.

LLMS such as GPT-Driver [8] and LanguageMPC [7] introduced chain-of-thought planning but cannot process raw visual input and rely on text descriptions, creating an information bottleneck [11]. A separate perception module must first convert the visual scene to text, inevitably losing spatial details and appearance nuances.

VLMs such as DriveVLM [11], DriveLM [12], and BLIP-2 [10] couple visual encoders with language decoders, enabling simultaneous perception, reasoning, and natural-language explanation of raw images. This multimodal approach eliminates the need for a separate scene-to-text conversion step.

This paper makes three contributions: (1) a targeted literature review of over 30 publications across three AI paradigms; (2) an experimental comparison using ResNet- 50, Flan-T5-base, and BLIP-2 on 10 real-world driving images with 70 QA pairs; and (3) quantitative evidence that VLMS achieve 84.3% accuracy and offer multimodal capabilities that CNNs and LLMs cannot match independently.

RELATED WORK

CNNs in Self-Driving

The foundation of autonomous driving perception rests on CNN architectures. SSD [4] and YOLO [2] represent real-time single-stage detection, while Faster R-CNN [3] achieves higher precision through region proposals. SCNN [16] applies spatial message passing for lane detection, and the lightweight variant of Ultra-Fast Lane Detection [17] exceeds 300 frames per second. Semantic segmentation networks such as PSPNet [18] and DeepLab [19] enable pixel-level scene understanding. Despite these advances, CNNs function as pattern recognizers without reasoning capacity, generalize poorly to unseen scenarios, and cannot explain their predictions to human operators or regulators [5]. This inability to provide explanations has been recognized as a key barrier to regulatory acceptance.

LLMs in Autonomous Driving

LLMs have been adapted for text-based reasoning in driving tasks. GPT-Driver [8] frames motion planning as language modelling. DriveGPT4 [9] offers interpretable driving decisions by leveraging LLM multimodal capabilities. LanguageMPC [7] uses GPT-3.5 to generate control policies for complex scenarios, while Talk2BEV [14] enables natural-language queries over bird's-eye-view representations. The fundamental limitation is that LLMs operate on textual scene representations, creating an information bottleneck in which vital spatial relationships and appearance details are irretrievably lost [11][12]. A text description can indicate that vehicles are ahead but cannot capture exact spatial configurations, subtle pedestrian gestures, or degraded road-surface conditions.

VLMs in Autonomous Driving

VLMs encode and decode visual and textual representations within a co-located architecture. DriveVLM [11] pairs chain-of-thought reasoning with visual grounding and reports leading planning results on the nuScenes benchmark [20]. DriveLM [12] models driving as a graph-structured VQA task, receiving an oral presentation at ECCV 2024. EM-VLM4AD [13] shows that a lightweight VLM can need at least 10× less memory and fewer floating-point operations than earlier approaches while still scoring higher on the DriveLM benchmark.

Collectively, these studies position VLMs as the optimal general approach for scene understanding, explanation, and generalization in autonomous driving, capable of detecting objects, analysing spatial relations, reasoning about driving manoeuvres, and generating plain-text explanations.

EXPERIMENTAL METHODOLOGY

Dataset and Benchmark Design

To ensure a fair comparison across all three paradigms under identical conditions, we assembled a controlled benchmark using ten real-world driving scene photographs: highway driving, urban intersection, residential road, rainy weather, night driving, construction zone, rural area, pedestrian

crossing, parking area, and foggy highway. This selection spans the diversity of scenarios an autonomous vehicle encounters in practice.

Seven questions were formulated for each image: scene classification, weather detection, time-of-day detection, vehicle presence estimation, pedestrian detection, traffic-signal recognition, and driving-decision recommendation, yielding 70 question-answer pairs with manually verified ground-truth answers.

Models Under Evaluation

Three models were evaluated. (1) CNN: ResNet-50 [15] trained on ImageNet with heuristic pixel-based scene understanding. ResNet-50 is among the most widely used CNNs in autonomous driving perception modules and serves as a reasonable baseline for a purely visual feature extractor without language reasoning.

LLM: Flan-T5-base [21] receiving text input from a noisy perception simulator that generates realistic detection errors, emulating LLM pipelines such as GPT-Driver [8]. The noisy perception unit introduced misclassifications, false negatives, and inaccurate counts to test the LLM paradigm under realistic conditions.
VLM: BLIP-2 [10] with an OPT-2.7B language backbone [22] receiving raw images and text queries, representative of the multimodal paradigm. All experiments ran on an NVIDIA A100-SXM4-40GB GPU. Fuzzy matching and semantic aliasing were used in scoring to enable fair comparison across heterogeneous output formats.

EXPERIMENTAL RESULTS

Overall Performance

Results are summarised in Table I. VLM (BLIP-2) achieved 84.3% overall accuracy (59/70), clearly outperforming CNN (61.4%, 43/70) and LLM (60.0%, 42/70). BLIP-2 attained 100% accuracy in five of seven categories: weather analysis, time-of-day, vehicle presence, pedestrian detection, and traffic-signal recognition, demonstrating the power of its multimodal architecture.

Category-Level Analysis

The CNN achieved perfect accuracy in time-of-day prediction (100%) through brightness thresholding, but struggled with scene classification (30%) and driving decisions (50%), reflecting fundamental limits of pixel-level features for semantic comprehension. The LLM demonstrated the highest proficiency in scene classification (80%), yet obtained only 10% on driving decisions—unacceptable for practical deployment.

Detection and Visual Understanding

Figures 3 and 4 show the differences between the three approaches in detection performance. The CNN (Faster R-CNN) successfully produced bounding boxes and confidence scores for individual objects but could not interpret broader scene semantics or infer driving implications. LLM-based models generated responses from text input without accessing visual elements. VLMs integrated object detection with natural-language scene descriptions and decision recommendations, illustrating the advantage of joint visual-linguistic processing.

TABLE I. Experimental Results on 10 Real Driving Scenes (70 QA Pairs)

Model	Acc.	Correct	Latency	Input	Scene	Weather	Time	Vehicle	Ped.	Signal	Decision
CNN	**61.4**	43/7	0.0	Image	30%	60%	100	70%	70%	50%	50%

(ResNet-50)	**%**	0	39 s	only			%				
LLM (Flan-T5)	**60.0 %**	42/70	0.68 7 s	Text only	80%	40%	100 %	80%	80%	30%	10%
VLM (BLIP-2)	**84.3 %**	**59/70**	4.0 68 s	Img + Text	30%	**100%**	**100 %**	**100%**	**100 %**	**100%**	60%

Fig. 1. Overall accuracy comparison. VLM (84.3%) outperforms CNN (61.4%) and LLM (60.0%) across all 70 QA pairs.

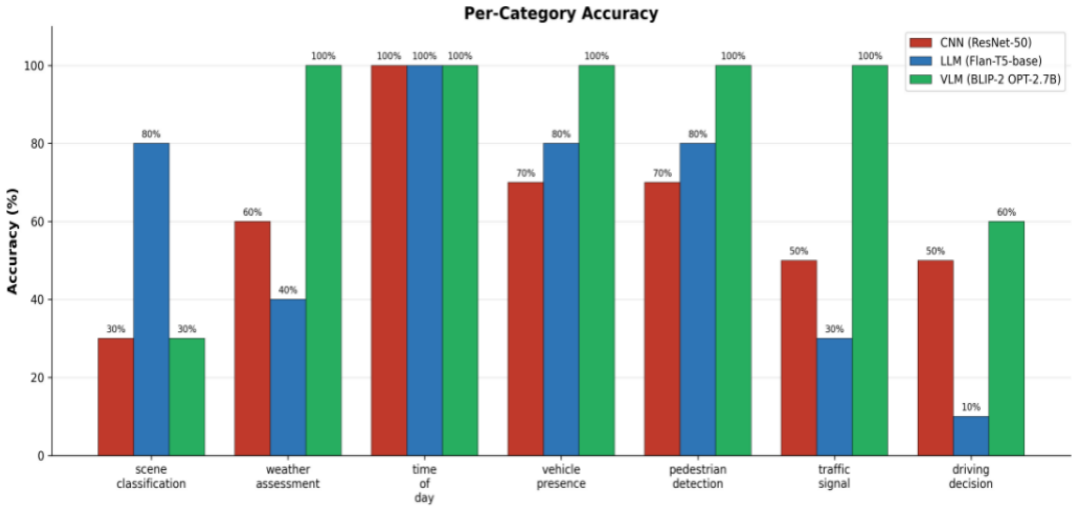


Fig. 2. Per-category accuracy breakdown across the seven evaluation categories. VLM achieves 100% in five of seven categories.

Fig. 3. Highway scene: CNN detects objects with bounding boxes (left). LLM has no visual access and relies on text descriptions (centre). VLM detects objects and generates descriptive scene understanding (right).

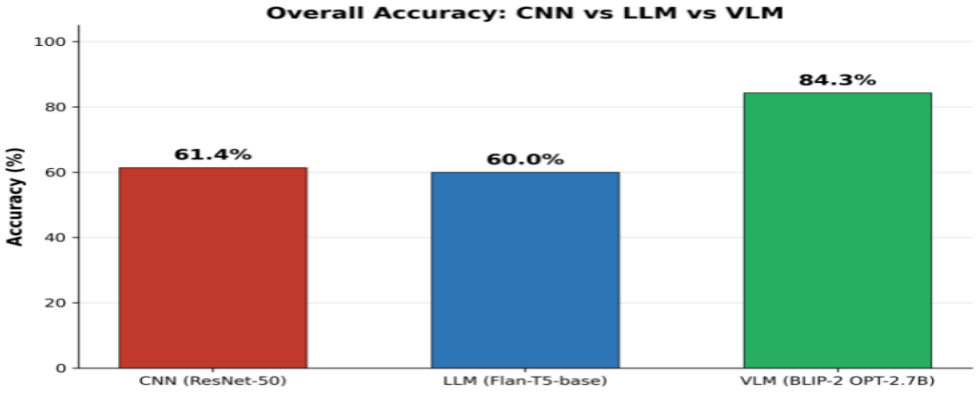
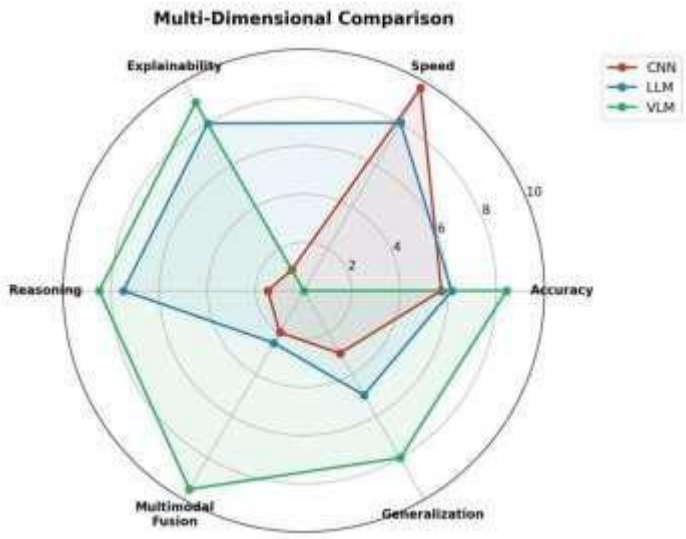


Fig. 4. Urban intersection: CNN detects objects (left). LLM receives text input only (centre). VLM provides full scene understanding with contextual reasoning (right).

Fig. 5. Multi-dimensional radar chart comparing CNN, LLM, and VLM across accuracy, explainability, reasoning, and multimodal fusion. VLM dominates across all evaluated dimensions.



DISCUSSION

Five significant observations validate the theoretical advantages of VLMs discussed in the literature review.

First, VLMs demonstrate genuine multimodal integration. The 84.3% accuracy confirms that analysing raw images with natural-language queries substantially enhances scene perception compared with pixel-level analysis (CNN: 61.4%) or perception-description-based reasoning (LLM: 60.0%). The VLM achieves perfect accuracy in weather prediction, vehicle and pedestrian counting, and traffic-signal recognition by reasoning directly from visual evidence.

Second, VLMs provide intrinsic explainability. As shown in Figs. 3 and 4, the VLM generates explanations of what it observes and why it makes recommendations, enabling autonomous vehicles to communicate their understanding in human-friendly terms. This directly addresses a key barrier to regulatory acceptance of CNN-based systems [5].

Third, the LLM text bottleneck remains a critical limitation. Despite receiving full perceptual descriptions, the LLM scored only 10% on driving decisions because textual representations cannot encode the visual information required for reliable context-dependent choices.

Fourth, CNNs offer the fastest inference speed (0.039 s/image on A100 GPU) but are limited to visual pattern matching. Their 30% scene-classification accuracy and 50% driving-decision accuracy confirm that CNNs are insufficient as stand-alone solutions, though they remain valuable as perception components within larger architectures.

Fifth, the computational cost of VLMs (4.068 s/image) is being addressed through rapid advances in model design and hardware. EM-VLM4AD [13] shows that lightweight architectures can cut memory and computation by 10× or more. Quantisation, pruning, and knowledge distillation are expected to bring VLM inference to sub-second levels. The radar chart in Fig. 5 confirms VLMs provide the most balanced capability profile across all evaluated dimensions.

FUTURE DIRECTIONS

Several research directions follow from our results. First, lightweight VLM architectures tuned for low-latency inference on vehicular edge hardware are urgently needed. BLIP-2, for example, needed about 4 s per image in our experiments—far exceeding the millisecond budgets of real-time autonomous driving.

Second, integrating VLMs with 3D scene representations—LiDAR point clouds, bird's-eye-view maps, and neural radiance fields—would extend their capabilities beyond the 2D RGB camera input considered here.

Third, adversarial robustness of VLMs must be rigorously evaluated before deployment. Understanding model responses to adversarial manipulation of both visual and text inputs, and designing robust defences, are critical safety requirements. Standardised multimodal benchmarks for assessing driving-related VLM capabilities should also be developed.

CONCLUSION

This paper reviewed over 30 publications and conducted an experimental comparison of CNN, LLM, and VLM paradigms for autonomous driving scene understanding. Using ResNet-50, Flan-T5-base, and BLIP-2 OPT-2.7B on10 real-world driving scenes with 70 QA pairs, BLIP-2 (VLM) achieved

84.3% accuracy, outperforming ResNet-50 (61.4%) and Flan-T5 (60.0%), while attaining 100% accuracy in five of seven evaluation categories and providing natural-language scene explanations.

Three conclusions follow: (1) CNNs are highly efficient for perception but lack reasoning capability; (2) LLMs offer strong reasoning but suffer from a visual information bottleneck; (3) VLMs unite the advantages of both, enabling next-generation autonomous driving systems that combine visual perception, reasoning, and explainable natural-language output. The path forward lies in lightweight, adversarially robust VLMs integrated with 3D sensing modalities, V2X communications, and standardised multimodal driving benchmarks.

REFERENCES

E. Yurtsever et al., "A survey of autonomous driving," IEEE Access, vol. 8, 2020. https://ieeexplore.ieee.org/document/9046805

J. Redmon et al., "You only look once: Unified real-time object detection," CVPR, 2016. https://arxiv.org/abs/1506.02640

S. Ren et al., "Faster R-CNN: Region proposal networks," IEEE TPAMI, vol. 39, 2017. https://arxiv.org/abs/1506.01497W. Liu et al., "SSD: Single shot multibox detector," ECCV, 2016. https://arxiv.org/abs/1512.02325

S. Grigorescu et al., "Deep learning techniques for autonomous driving," J. Field Robotics, vol. 37, 2020. https://doi.org/10.1002/rob.21918

A. Vaswani et al., "Attention is all you need," NeurIPS, 2017. https://arxiv.org/abs/1706.03762

H. Sha et al., "LanguageMPC: LLMs as decision makers for driving," arXiv, 2023. https://arxiv.org/abs/2310.03026

J. Mao et al., "GPT-Driver: Learning to drive with GPT," NeurIPS Workshop, 2023. https://arxiv.org/abs/2310.01415

Z. Xu et al., "DriveGPT4: Interpretable end-to-end driving via LLM," IEEE RA-L, vol. 9, 2024. https://arxiv.org/abs/2310.01412

J. Li et al., "BLIP-2: Language-image pre-training with frozen encoders," ICML, 2023. https://arxiv.org/abs/2301.12597

X. Tian et al., "DriveVLM: Autonomous driving and large VLMs," arXiv, 2024. https://arxiv.org/abs/2402.12289

C. Sima et al., "DriveLM: Graph visual question answering,"

ECCV, 2024. https://arxiv.org/abs/2312.14150

A. Gopalkrishnan et al., "Efficient VLMs for autonomous driving QA," CVPR Workshop, 2024. https://arxiv.org/abs/2403.19838

T. Choudhary et al., "Talk2BEV: Language-enhanced bird's-eye view maps for autonomous driving," ICRA, 2024. https://arxiv.org/abs/2310.02251

K. He et al., "Deep residual learning for image recognition,"

CVPR, 2016. https://arxiv.org/abs/1512.03385

X. Pan et al., "Spatial as deep: Spatial CNN for traffic scene understanding," AAAI, 2018. https://arxiv.org/abs/1712.06080

Z. Qin et al., "Ultra fast structure-aware deep lane detection,"

ECCV, 2020. https://arxiv.org/abs/2004.11757

H. Zhao et al., "Pyramid scene parsing network," CVPR, 2017. https://arxiv.org/abs/1612.01105

L. -C. Chen et al., "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," IEEE TPAMI, vol. 40, 2018.https://arxiv.org/abs/1606.00915

H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," CVPR, 2020. https://arxiv.org/abs/1903.11027

H. W. Chung et al., "Scaling instruction-finetuned language models," JMLR, vol. 25, 2024. https://arxiv.org/abs/2210.11416

S. Zhang et al., "OPT: Open pre-trained transformer language models," arXiv, 2022. https://arxiv.org/abs/2205.01068remive