engineering proceedings

MDPI

Proceeding Paper

Interactive In-Cabin Perception for Electric Vehicles: Vertical Domain Large Model and Label Design †

Hong Ming, Kuan Sun and Qiuchen Zhu *

Dongfeng Motor Institute of Research and Development, Wuhan 430000, China
* Correspondence: qiuchenzhu@hotmail.com
† Presented at the 2026 International Conference on Energy, Infrastructure and Environmental Research (EIER 2026), Ho Chi Minh City, Vietnam, 15–17 July 2026.

Abstract

Human–Machine Interaction (HMI) in Electric Vehicles (EVs) demands a paradigm shift from passive monitoring to intelligent, interactive cabin perception. Current systems are often constrained by rigid architectures and limited semantic understanding, failing to meet the dynamic requirements of personalized user experiences. This paper presents interactive in-cabin perception, a novel framework that leverages a Vertical-Domain Large Model to achieve a holistic understanding of the in-cabin environment. We introduce a specialized label design strategy that captures complex spatial relationships and multimodal human states, enabling the model to perform generative reasoning for open-vocabulary queries. By unifying perception and cognition into an end-to-end trainable architecture, our approach eliminates the error accumulation found in traditional modular pipelines. We validate the framework through extensive experiments, demonstrating superior performance in both object localization and attribute recognition within challenging EV cabin scenarios.

Keywords: in-cabin perception; human–machine interaction; multimodal perception

1. Introduction

In the context of Electric Vehicles (EVs), their evolution drives the need for advanced interactive in-cabin perception, in which the vehicle must master the internal shared physical space to ensure safe and adaptive Human–Machine Interaction (HMI) [1]. As EVs transform from a transportation tool into interactive personal spaces, the objective of HMI shifts toward a comprehensive understanding of the occupants. In this scenario, accurate, real-time spatial and semantic perception methods have been developed to provide the critical grounding for proactive safety interventions and personalized services [2]. The full perception of the occupants aims to resolve their locations and activities within a constrained environment. Unlike traditional DMS, which focuses solely on the frontal-view driver [3], holistic perception handles various multi-occupant configurations across varying fisheye viewpoints and seating layouts [4]. Existing approaches include hardware-driven sensing systems [2] and vision-based perception models [5]. Early solutions utilize dedicated sensors such as infrared modules [6], radar [7], or ultrasonic devices [8] for occupancy detection. Although they contribute positively to robustness, these systems provide limited semantic understanding, increase cost, and typically focus on occupancy rather than fine-grained localization and attribute recognition.

To address these limitations, we propose a Multimodal Large Language Model (MLLM) framework for unified in-cabin perception (Figure 1), reformulating occupant understanding as a holistic reasoning task over OMS images. Our end-to-end architecture jointly models

Check for updates

Academic Editor: Fengling Han

Published: 13 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

spatial configurations, context, and semantics through vision–language representations. For real-world robustness, we introduce structured data augmentation to simulate viewpoint variations, illumination shifts, and occlusions. Furthermore, a hybrid training paradigm integrates single-turn supervised learning with multi-turn conversational fine-tuning to encourage self-refinement and enhance contextual awareness. This approach shifts in-cabin perception from modular pipelines toward semantically grounded reasoning, enabling flexible frameworks for intelligent interaction.

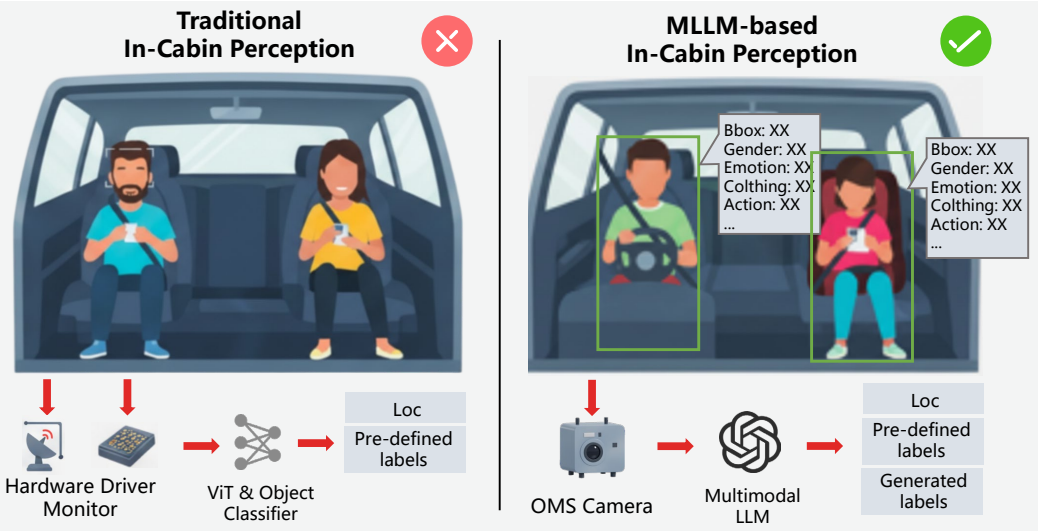


Figure 1. Comparison between traditional multi-sensor modular occupant monitoring pipelines and the proposed MLLM-based end-to-end semantic understanding framework.

The main contributions of this work are as follows:

- Occupant localization and attribute recognition are reformulated as a unified spatial-semantic task, grounded in interactive in-cabin environments, thereby establishing a generative basis for holistic environmental understanding.
- A Multimodal Large Language Model framework is introduced to jointly model spatial grounding and semantic reasoning. To enhance robustness in fine-grained scenarios, a hybrid training paradigm is proposed that combines single-turn structured prediction with multi-turn decomposition.

2. Related Work

The integration of MLLMs into the smart cabin marks a shift toward semantic-level understanding of the occupants' behavior. To standardize evaluation, IntelliCockpitBench [9] and SLU benchmark [10] assess multimodal interaction and command responsiveness. Performance is further enhanced by merging RGB, Infrared Spectroscopy (IR), and depth data for robust action recognition [11]. To enable real-time deployment, models are increasingly optimized for edge hardware, facilitating the on-device monitoring of fatigue and distraction. Safety is extended through holistic, risk-aware frameworks that combine physiological signals with MLLM-based perception [12]. Finally, the convergence of MLLMs with embodied intelligence enables spatial reasoning via active exploration [13], while LLM-based reasoning ensures human-centric, adaptable in-cabin responses.

https://doi.org/10.3390/engproc2026164013

3. Materials and Methods

3.1. Preliminaries

To achieve fine-grained occupant behavior within the embodied environment of a smart cockpit, we define the In-Cabin Occupant Perception (ICOP) task. Given a monocular image $\mathcal{I} \in \mathcal{R}^{H \times W}$ captured by an Occupant Monitoring System (OMS) camera, the goal is to simultaneously perform spatial localization and Multi-Attribute Reasoning for all occupants.

Formally, for an image $\mathcal{I}$, the system must output a set of occupant descriptors $O = \{o_1, o_2, \ldots, o_n\}$, where n is the number of occupants. Each occupant o_i is defined as a tuple:

$$o_i = \langle P_i, A_i \rangle$$

Spatial Occupancy (P_i): $P_i \in \{L_1, L_2, \ldots, L_5\}$ represents the categorical seat position. Multi-Dimensional Attribute Set (A_i): A_i is a high-dimensional feature vector encompassing semantic labels in the following sections:

- Biometric & Identity: Personalized features related to cabin customization, such as gender and age.
- Expression & State: Facial expression of the occupants for emotional interaction.
- Action & Safety: Behavioral features such as pose and gesture, including some specific behavior closely related to driving safety.
- Clothing & Accessories: Apparel details to improve the customization of the dialog in the interaction.

The objective is to learn a mapping function $f : \mathcal{I} \rightarrow O$ that takes advantage of the zero-shot reasoning capabilities of a MLLM to interpret these complex and inter-related attributes under varying lighting and occlusion conditions typical of automotive environments.

3.2. Supervised Fine-Tuning

Although the performance of fundamental MLLMs is appealing for generic AIGC tasks, they are barely used for in-cabin perception due to the lack of expert knowledge in this vertical domain. To bridge the gap between raw in-cabin data and the generative nature of MLLM, we developed an instruction-tuning pipeline that transforms multimodal annotations into high-quality training pairs. The model is optimized using a standard Supervised Fine-Tuning (SFT) paradigm. Given the multimodal input $\mathcal{I}$ and the corresponding ground-truth sequence $\mathcal{Y}$, we minimize the cross-entropy loss $\mathcal{L}_{SFT}$ over the generated tokens:

$$\mathcal{L}_{SFT} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, \mathcal{I}) \quad (1)$$

where y_t represents the t-th token of the occupant descriptors O. This objective ensures that the MLLM accurately aligns the spatial coordinates of the occupants with their respective attributes A_i in a unified generative framework.

4. Experiment

4.1. Dataset Preparation

4.1.1. Multimodal Data Acquisition

We constructed a comprehensive in-cabin multimodal dataset comprising 20,000 high-fidelity images captured via the OMS system. The data encompasses diverse lighting conditions and seating configurations. All samples were collected in real vehicle environments with informed consent, featuring annotations across the dimensions of "Human-

https://doi.org/10.3390/engproc2026164013

Vehicle-Space", including seating positions and other semantic features, as introduced in Section 3.1.

4.1.2. Hierarchical Dataset Architecture

To ensure robustness in complex scenarios, we implemented a "Comprehensive & Task-Specific" hierarchical structure, including two sources in data processing as below:

- Complexity-Prioritized Selection: 5000 samples were selected using occupant-density stratified sampling, prioritizing complex multi-passenger scenarios to enhance perceptual stability under full-load conditions.
- Attribute-Specific Subsets: To mitigate long-tail imbalances, we integrated 5000 specialized emotion samples. Additional subsets for age-specific recognition and gesture-based interaction were included to strengthen fine-grained reasoning.

4.1.3. SFT Training

We utilized a high-capacity Qwen MLLM to perform semantic redundancy on activity descriptions, ensuring that the final instruction set is both linguistically concise and pedagogically effective for the MLLM training.

4.2. Experimental Setup and Evaluation Metrics

The framework is implemented in PyTorch V2.3.0. Given the unique fisheye perspective and the lack of standard MLLMs for in-cabin scenarios, we systematically compared Qwen-series models. We selected Qwen2.5-VL-3B [14] as the primary backbone to balance reasoning ability with the real-time inference constraints of edge devices in cars. The model was optimized using the Adam algorithm with a learning rate of 5×10^{-5} and a total batch size of 8 on 8 NVIDIA A100 GPUs. For visual inputs, we employed a dynamic resolution strategy, rescaling OMS images to approximately one million pixels to preserve fine-grained details like expressions and seatbelt buckles.

To comprehensively assess the performance of the large model in occupant perception, we define three distinct metrics based on the nature of the tasks:

Spatial Localization Accuracy (Acc_{loc}). For the occupant positioning task, the model predicts the coordinates of the bounding box $B = (x_{min}, y_{min}, x_{max}, y_{max})$. In this case, the quality of the localization is evaluated using the Intersection over Union (IoU) metric. It is considered a true positive (TP) sample if the IoU between the predicted box B_{pred} and the ground-truth box B_{gt} exceeds a threshold of 0.5, following the Pascal VOC standard:

$$\mathrm{IoU}(B_{pred}, B_{gt}) = \frac{\mathrm{Area}(B_{pred} \cap B_{gt})}{\mathrm{Area}(B_{pred} \cup B_{gt})} > 0.5 \tag{2}$$

And spatial localization accuracy Acc_{loc} is thus the proportion of samples that satisfy this criterion.

Categorical Attribute Accuracy (Acc_{cat}). For closed-vocabulary attributes, we measure the alignment between the predicted label y_{cat} and the ground truth g_{cat}. The accuracy is defined as the ratio of correct predictions to the total number of categorical instances N_{cat} across all TP detections:

$$Acc_{cat} = \frac{1}{N_{cat}} \sum_{i=1}^{N_{cat}} \mathcal{I}(y_{cat,i} = g_{cat,i}) \tag{3}$$

where $\mathcal{I}(\cdot)$ is the indicator function.

Generative Semantic Similarity (Acc_{gen}). For open-ended attributes such as Activity and Clothing Accessories, where exact string matching is overly restrictive, we utilize

https://doi.org/10.3390/engproc2026164013

a Sentence-BERT (SBERT)-based [15] semantic metric. Let $\sqsubseteq_{gen}$ and $\sqsubseteq_{gt}$ be the high-dimensional embeddings generated by a pre-trained SBERT model for the predicted and ground-truth descriptions, respectively. A generative prediction is deemed accurate if the Cosine Similarity S_{cos} exceeds a threshold $\tau = 0.5$ for balanced performance.

$$S_{cos}(\sqsubseteq_{gen}, \sqsubseteq_{gt}) = \frac{\sqsubseteq_{gen} \cdot \sqsubseteq_{gt}}{|\sqsubseteq_{gen}||\sqsubseteq_{gt}|} > 0.5 \quad (4)$$

The semantic generative similarity Acc_{gen} is the ratio of samples that meet this criterion.

4.3. Main Results

Table 1 summarizes the evaluation of various MLLMs on the in-cabin perception task, yielding the following key insights.

Table 1. Comparison of performance across different MLLM backbones before and after SFT. The metrics include spatial localization accuracy for occupant positions and attribute recognition scores across four semantic dimensions.

Models	Occupant Position	Biometric & Identity	Expression & State	Action & Safety	Clothing & Accessories
Qwen3-vl-4B	99.08	77.37	75.20	51.58	67.48
+ training	99.55	86.75	82.77	87.63	94.64
Qwen3-vl-8B	99.23	77.49	77.41	60.05	69.87
+ training	99.64	89.33	87.77	89.97	96.05
Qwen2.5-vl-3B	99.15	58.01	78.26	52.72	66.67
+ training	99.52	82.95	81.68	86.28	90.68

The experiment shows that the fine-tuned vertical model reaches saturation in spatial localization. All pre-trained models exhibit near-perfect accuracy in positioning (>99%) even before fine-tuning, indicating strong inherent spatial grounding capabilities within the vertical domain. Our SFT process refines this to 99.64%.

The results also demonstrate significant SFT gains in semantic reasoning. Zero-shot performance in Action & Safety and Clothing & Accessories is initially poor, respectively at 51.58% and 69.87%. Following our training paradigm, substantial leaps are observed, with a relative gain of 63.6% for Qwen2.5-vl-3B in Action & Safety, demonstrating that SFT is essential for teaching the model "what" the occupants are doing. Although Qwen3-vl-8B achieves the highest accuracy, particularly in Expression & State at 87.77%, the performance gap with the 3B model is less than 5%. This suggests that a high-quality, task-aligned dataset compensates for smaller parameter counts. For real-time in-vehicle deployment, Qwen2.5-vl-3B provides the best trade-off, maintaining high accuracy in safety-critical categories, especially at 81% for Expression, and above 86% for Action.

5. Conclusions

This paper presents a unified, end-to-end MLLM framework tailored for interactive in-cabin EV perception. By formulating perception as a generative reasoning task, our method effectively addresses the inherent limitations of conventional modular pipelines. The core contributions—Spatial Alignment Mirroring, Hybrid Dialogue Formatting, and Attribute-Balanced Sampling—specifically target critical challenges within the vertical domain, namely, spatial bias and data imbalance in complex cabin environments. Experimental results demonstrate that the proposed framework, particularly when utilizing the Qwen2.5-VL-3B backbone, achieves high accuracy while satisfying real-time inference constraints for in-vehicle deployment. This work thus establishes a robust perceptual foundation, enabling

a new generation of intelligent cockpits capable of proactive, embodied interaction. Nevertheless, it lacks validation under extreme cabin conditions and cannot adapt to dynamic data distribution shifts. Future research will enhance model robustness against complex driving scenes, while expanding perception tasks to build a more comprehensive in-cabin intelligent interaction system.

Author Contributions: Conceptualization, Q.Z.; methodology, H.M. and Q.Z.; software, K.S.; validation, H.M.; formal analysis, H.M.; investigation, H.M.; resources, K.S. and Q.Z.; data curation, H.M.; writing—original draft preparation, H.M. and Q.Z.; writing—review and editing, Q.Z.; visualization, H.M.; supervision, Q.Z.; project administration, Q.Z.; funding acquisition, Q.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: During the preparation of this manuscript/study, the authors used Gemini V3.0 for the purposes of figure design (Figure 1) and paragraphing for grammar correction. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: Authors Hong Ming, Kuan Sun and Qiuchen Zhu were employed by Dongfeng Motor Institute of Research and Development. All authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Chen, M.; Wang, C.; He, X.; Zhu, F.; Wang, L.; Vasilakos, A. Embodied Artificial Intelligence-Enabled Internet of Vehicles: Challenges and Solutions. *IEEE Veh. Technol. Mag.* **2025**, *20*, 63–70. [CrossRef]
2. Xiao, Z.; Ye, K.; Cui, G. PointNet-Transformer Fusion Network for In-Cabin Occupancy Monitoring with mm-Wave Radar. *IEEE Sens. J.* **2024**, *24*, 5370–5382. [CrossRef]
3. He, B.; Yu, N.; Han, H. Driver Intention Prediction Using Text Prompts with In-Cabin and Out-Cabin Cameras. In Proceedings of the 2024 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), Guangzhou, China, 24–26 October 2024; pp. 40–43.
4. Sato, K.; Wandale, S.; Ichige, K.; Kimura, K.; Sugiura, R. Low-Cost Vehicle In-Cabin Occupancy Detection Using an Approximate Model of XGBoost. In Proceedings of the 2023 IEEE International Radar Conference (RADAR), Sydney, Australia, 6–9 November 2023; pp. 1–6.
5. Ma, Y.; Ye, W.; Cao, X.; Abdelraouf, A.; Han, K.; Gupta, R.; Wang, Z. Cemformer: Learning to Predict Driver Intentions from In-Cabin and External Cameras via Spatial-Temporal Transformers. In Proceedings of the 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), Bilbao, Spain, 24–28 September 2023; pp. 4960–4966.
6. Lyu, J.; Yang, Y.; Shi, Y.; Lian, Z. Application-Driven Development of a Thermal Imaging-Based Cabin Occupant Thermal Sensation Assessment Model and Its Validation. *Build. Simul.* **2024**, *17*, 1401–1417. [CrossRef]
7. Farsaei, A.; Meyer, B.; Sheikh, A.; Soussi, M.E.; Zhang, P.; Ramachandra, G.K.; Govers, J. An IEEE 802.15.4z-Compliant IR-UWB Radar System for In-Cabin Monitoring. In Proceedings of the 2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Toronto, ON, Canada, 5–8 September 2023; pp. 1–5.
8. Wang, S.; Zeng, Y.; Jain, V.; Pathak, P. Umusic: In-Car Occupancy Sensing via High-Resolution UWB Power Delay Profile. In Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems (SenSys), Irvine, CA, USA, 6–9 May 2025; pp. 116–129.
9. Lin, L.; Chai, S.; Wu, J.; Hu, H.; Gu, X.; Hu, H.; Zhang, F.; Wang, W.; Zhang, D. IntelliCockpitBench: A Comprehensive Benchmark to Evaluate VLMs for Intelligent Cockpit. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025, Vienna, Austria, 27 July–1 August 2025; pp. 15453–15475.
10. Peng, Y.; Cai, C.; Liu, Z.; Fan, S.; Jiang, S.; Xu, H.; Liu, Y.; Chen, Q.; Xu, K.; Li, Y. MAC-SLU: Multi-Intent Automotive Cabin Spoken Language Understanding Benchmark. *arXiv* **2025**, arXiv:2512.01603.

11. Wang, R.; Wang, W.; Gao, J.; Lin, D.; Yap, K.I.; Li, B. MultiFuser: Multimodal Fusion Transformer for Enhanced Driver Action Recognition. In Proceedings of the 2024 IEEE 26th International Workshop on Multimedia Signal Processing (MMSP), West Lafayette, IN, USA, 2–4 October 2024; pp. 1–6.
12. Tang, W.; You, J.; Liu, J.; Wang, Z.; Gan, R.; Huang, Z.; Wei, F.; Ran, B. HERMES: A Holistic End-to-End Risk-Aware Multimodal Embodied System with Vision-Language Models for Long-Tail Autonomous Driving. *arXiv* **2026**, arXiv:2602.00993.
13. Zhang, P.; Huang, Z.; Wang, Y.; Zhang, J.; Xue, L.; Wang, Z.; Wang, Q.; Chandrasegaran, K.; Zhang, R.; Choi, Y.; et al. Theory of Space: Can Foundation Models Construct Spatial Beliefs through Active Exploration? In Proceedings of the Fourteenth International Conference on Learning Representations (ICLR), Rio de Janeiro, Brazil, 23–27 April 2026.
14. Qwen Team. Qwen2.5-VL Technical Report. *arXiv* **2025**, arXiv:2502.00001.
15. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3982–3992. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.