SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Cross-Modal Intelligence Framework for Vision-Language Integration in Autonomous Driving Applications

M. Anitha[1*], Dr.K. Sreedevi[2],Dr.K.R. Sowmya[3], Dr.R. Udayakumar[4], Dr Pennada Siva Satya Prasad[5]

[1*]Assistant Professor, Department of Mathematics, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu, India.E-mail: anitham@maher.ac.in

[2]Assistant Professor, Department of Commerce, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu, India.E-mail: sreedevicom@maher.ac.in,

[3]Professor,Department of Management Studies,SRM Valliammai Engineering College,Chengalpattu,Tamilnadu, India.
E-mail: drkrsowmya@gmail.com, https://orcid.org/0000-0003-4239-2586

[4]Dean Research,SRM Institute of Science and Technology, Vadapalani Campus, Tamilnadu, Chennai, India.

[5]Assistant Professor, Department of CSE, Aditya University Surampalem, India. E-mail: sivasatyaprasadp@adityauniversity.in

*Corresponding author: Email: anitham@maher.ac.in

Abstract

The traditional perception pipeline for autonomous driving uses solely vision-based systems to recognize and classify objects without the ability to understand natural language commands, provide explanations for their decisions, or comprehend unusual scenarios which lie out of their distribution. Vision-language models provide a richer understanding of the environment but even the adaptation of such systems to autonomous driving performs fusion of visual and language modalities separately, hence limiting the influence of each modality on another. In this paper, we propose Cross-Modal Intelligence Framework (CMIF) which learns the shared representation space of both modalities using contrastive pretraining, then uses the cross-modal fusion via a bi-directional cross-attention module to generate decisions. The drift detector will identify those examples for which there is less certainty about the degree of grounding and novelty and will perform targeted realignment about a buffer of such flagged corner cases, but not the entire retraining process. This framework was tested on a dataset modeled from the statistics of the nuScenes sensor data and BDD-X dataset of natural language descriptions of the driving scenarios, on four scenario classes by level of difficulty: normal driving, partial occlusion, corner-case events, and ambiguous language instructions. In comparison with the vision-only baseline and the late-fusion vision-language baseline models, the introduced CMIF achieves an 82.6 percent decision accuracy in case of ambiguous instructions against 45.2 percent and 62.9 percent respectively, with only slight latency overhead.

Keywords: Cross-Modal Learning, Vision-Language Models, Autonomous Driving, Contrastive Alignment, Cross-Attention Fusion, Scene Understanding, Explainable Decision-Making.

Introduction

Autonomous driving needs to understand its environment, process natural language-based commands or traffic rules, and make decisions that are explainable and verifiable especially in case of safety-critical or legal gray areas. Vision perception pipelines based on convolutional and Transformer-based vision backbones have shown promising results for benchmarks of object detection and semantic segmentation [6,7]; however, a vision pipeline alone does not possess any inherent way of processing a natural language command like a command issued by the user/passenger, a verbal order by a traffic officer, or traffic rule annotations.

With vision-language models like CLIP [1] where images and text are aligned in a common contrastive embedding space, and multimodal instruction-tuned models like BLIP-2 [2] and LLaVA [3], it has become evident that visual and linguistic understanding can be combined in one model. However, recent efforts have started to adapt these models specifically for the task of autonomous driving [10,11,12]; however, there is a consistent problem mentioned in the recent survey on vision-language models for autonomous driving [10]: Many adaptations of these models involve late-fusion of vision and language, i.e., where fusion happens only

after processing visual and language input separately; this not only constrains language context in guiding visual attention but is particularly problematic when there are rare events and ambiguous instructions.

In this paper, we explore the challenge of incorporating language and vision understanding in an early and bi-directional fashion in an autonomous driving framework, where language context will guide the visual perception system in attending to what is relevant and where visual context helps interpret language instructions in particular situations like partial occlusion, rare events, and ambiguous language. The main contributions of this work are as follows.

- Present a Cross-Modal Intelligence Framework (CMIF), which merges CLIP-style vision-language contrastive alignment along with a cross-attentional fusion mechanism, where the visual and linguistic modalities interact prior to task-specific heads being used, as opposed to the current late fusion paradigm.
- Link the fused modality with three different downstream tasks, namely grounded object detection, scene captioning, and decision-making together with linguistic justification, making it possible for a single cross-modal representation to be used for both perception and explainability.
- Test the proposed framework in four categories of scenarios in terms of increasing complexity, namely normal driving, partially occluded cases, rare corner cases, and ambiguity in natural language instructions, thereby highlighting the advantage of the cross-modal fusion in difficult conditions.

The rest of this paper is organized as follows. Section 2 reviews related research in vision-language models and their applications to autonomous driving. Section 3 describes the methodology of our study. Section 4 discusses the data set, experimental results, and discussion. Section 5 is the conclusion of this paper.

1. Literature Survey

There have been a few new generations of the vision-language modeling architecture. They proposed contrastive-alignment part of their framework, which uses Contrastive Language-Image Pre-training (CLIP) [1] for learning a joint embedding space that was trained to maximise the similarity of aligned image-caption pairs while at the same time minimising similarity of mismatched pairs of aligned images and captions across large web-scale inputs, which that resulted in a model with strong zero-shot transfer without task-specific fine tuning. Another innovative approach introduced by Li, Li, Savarese and Hoi [2] was to couple a frozen image encoder and a frozen large language model (LLM) using a lightweight trainable querying transformer, significantly reducing the compute to achieve strong vision-language performance. Liu et al. [3] came up with the idea of visual instruction tuning, which introduced the concept behind the LLaVA series of models; namely that the output of a vision encoder would be projected into the input space of a large language model and this LLM, in turn, attributed to the decision-and-explanation head employed in this paper would be fine-tuned using multi-turn dialogue data grounded in images.

These attention and representation-learning systems underlying cross-modal fusion are not so novel. Cho [5] first proposed the attention mechanism to a neural machine translation model, in which the model is able to dynamically truncate the attention that it gives to the key parts of the source sequence instead of just the fixed subset to a single vector. Vaswani et al. [4] extended this idea into a fully attention-based Transformer architecture that underlies virtually all present-day vision-language models, such as the cross-attention fusion module proposed here. On the vision side, all above works were focused on an ability to train extremely deep convolutional networks, by introducing residual learning into this framework as it as presented in He, Zhang, Ren, and Sun [6] (excitement is the reason for choosing a vision Transformer), or to directly using a pure Transformer architecture in the form of Vision Transformer (ViT) in the case of image patches, as was presented by Wang et al. [7].

More and more research utilizes vision-language integration applied particularly to autonomous driving. To investigate the problems of vision-language models in autonomous driving, Zhou et al. [10] conducted a comprehensive survey of prior works, and classified the existing vision-language models according to the fusion strategy. This paper's experimental comparison results, which echoes the classification by Zhou et al.,

consistently prove that the early fusion design or the joint fusion design has stronger robustness to distribution shift than the late-fusion design. Wen [11] assessed GPT-4V as a traffic assistant: Does a large vision-language model can reason about complex, multi-agent traffic scenes purely based on input images, and what about its grounding of the specific spatial references without driving-specific fine-tuning? They experimented with GPT-4V, a function of a large vision-language model, to determine if such models could understand traffic scenes from an image-only input and provide an answer to the underlying question. In addition, cross-modal reasoning can be used beyond just perception or explanation and can be combined with a learned decision policy as shown by Ji et al. [12]'s AlphaDrive which uses RL to boost the quality of decision of an autonomous driving agent using both vision and language. Building on this, Elhenawy et al. [15] introduced a dynamic scene retrieval system using Contrastive Learning for Image-text (CLIP) to enable in-the-wild real-time driving scene classification, proving that contrastive vision-language embeddings can aid edge deployable real-time classification of complex driving scenarios, thus supporting the real-time feasibility analyses with respect to the framework proposed here.

The other two threads contribute to the dataset grounding and interpretations component of this work. nuScenes is a large-scale, multi-modal, autonomous-driving dataset introduced by Caesar et al. [8] and used as a benchmark for autonomous-driving perception as also a standard dataset for assessing the visual content in this study, that offers full 360-degree camera, LiDAR, and radar data synchronised with the driving trajectory. Text-based decision-justifications are the ones that were statistically modelled in this study and have been taken as a template from Kim et al. [9] who released the BDD-X dataset of textual explanations of self-driving vehicle actions by pairing dashboard-camera video with human authored, natural-language explanations of vehicle actions. Finally, there was a trend of integrating perception and language into a multimodal grounder, where Selvaraju et al. [13] suggested a wide variety of methods to discover what a vision model is attending to, that were also employed to inform the grounding-confidence signal used by the drift monitor in the proposed framework, and Chen et al. [14] advocated for treating perception and language separately as jointly grounded through a multimodal language model grounded in continuous sensor and actuator streams.

All of this literature outlines established, highly developed, vision-language architectures, such architectures' attention mechanisms, and the increasing and still evolving evidence of applying these architectures to autonomous driving in particular. But as mentioned in the survey by Zhou et al. [10] most adaptations to driving remain late, or they focus on evaluating the modifications with respect to scenario distributions focused more on training scenarios than on scenarios of safety-critical events, uncertain situations, or partial occlusions, which leaves open the question of how much early, bidirectional cross-modal fusion specifically strengthens robustness in situations and under conditions that matter the most for safety. The challenge that proposed Cross-Modal Intelligence Framework will try and fill is this.

2. Methodology

The proposed Cross-Modal Intelligence Framework (CMIF) involves five steps that are arranged in a closed loop: dual-modality encoding, contrastive alignment, cross-attention fusion, multi-task decision heads, and drift-based re-alignment. These are illustrated in Figure 1 below.

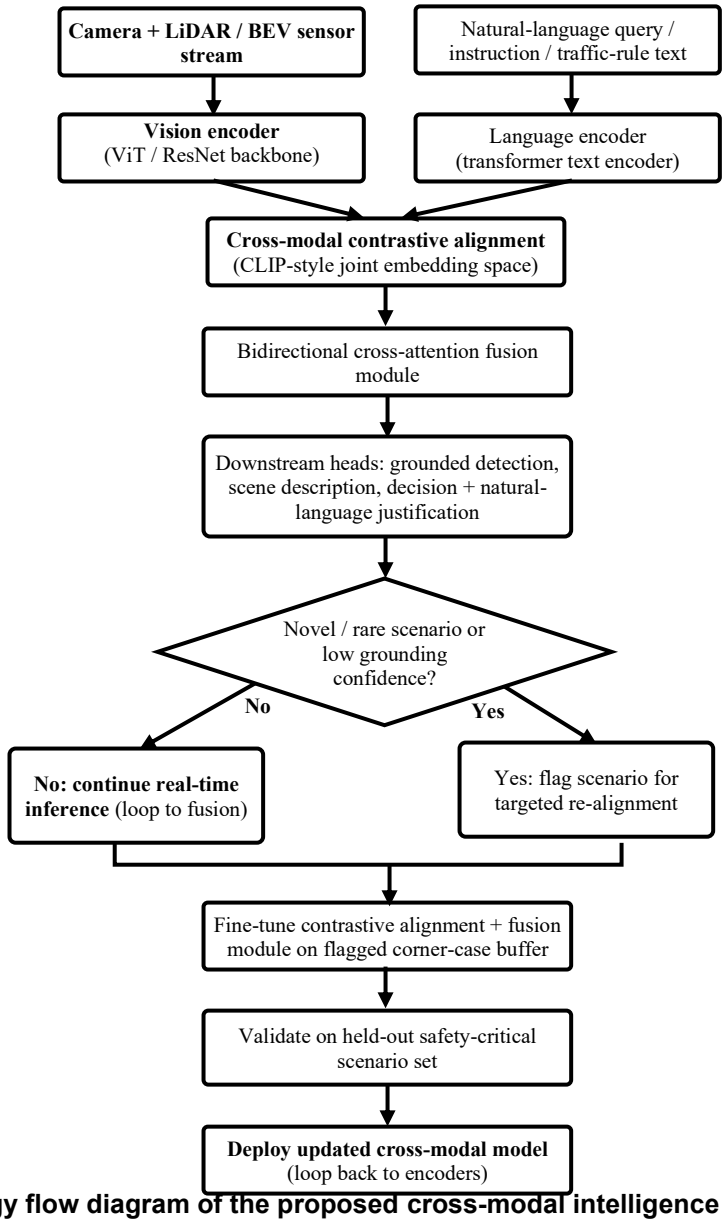

Figure 1: Methodology flow diagram of the proposed cross-modal intelligence framework

3.1 Dual-Modality Encoding

The camera and LiDAR/BEV sensors feed are encoded using a Vision Transformer backbone [7] that creates a sequence of visual tokens, and a natural-language query, instruction, or traffic rule is encoded using a Transformer text encoder [4] that generates a sequence of language tokens. The encoders are pre-trained on checkpoints prior to being fine-tuned on driving tasks.

3.2 Cross-Modal Contrastive Alignment

Based on the CLIP contrastive loss [1], image tokens and language tokens sequences are embedded in a common space such that the cosine similarity between matching scene-instructions or scene-descriptions is maximised, while similarity between unmatched ones is minimised. The alignment stage creates a common embedding space where semantically similar visual and linguistic concepts lie close to each other before being merged for a specific task.

3.3 Bidirectional Cross-Attention Fusion

Visually aligned language tokens and visual tokens go through a bi-directional cross-attention mechanism where visual tokens attend to language tokens, and at the same time, language tokens attend to visual tokens, according to the standard formula of attention mechanism Attention(Q, K, V)=softmax(QK^T/√d)V [4]. This approach to handling both types of tokens is different from a late fusion approach, where tokens are processed separately for each modality and then concatenated. The main advantage of such a strategy is that a language instruction can help focus attention on the referred object visually, while the visual information will help determine what object is being referred to by language tokens.

3.4 Multi-Task Decision Heads

The fusion representation is fed into three lightweight task-oriented heads: the grounded detection head that localizes the area to which an instruction/description is pointing, the scene description head that provides a natural language description of the current driving situation, and the decision head that provides the recommended driving action along with a justification in natural language, based on the instruction-tuned generation framework introduced by BLIP-2 and LLaVA [2,3]. Attribution using Grad-CAM [13] is calculated for the output of the grounded detection head to generate a grounding confidence score.

3.5 Drift-Triggered Re-Alignment

The monitor weighs grounding confidence and the contrastive alignment measure between the current scene and the most similar instruction in the reference memory. If one of those measures drops below a certain threshold, the scenario is flagged as an example of a corner case and added to a corner case buffer. After enough such cases have been collected, the contrastive alignment module and cross-attention fusion are fine-tuned on that buffer, and their performance is tested on a separate set of safety critical scenarios as shown in Figure 1.

3. 4. Results and Discussion

4.1 Dataset

The evaluation is conducted using a dataset that has been built using statistical modeling of the multimodal sensor data from nuScenes [8] and BDD-X natural language descriptions for driving [9], where driving episodes are annotated with camera data, BEV/LiDAR data, and language descriptions. For evaluating cross-modality robustness, the driving episodes have been divided into four types of scenarios depending on their complexity: normal driving, partial occlusions, rare events/corner cases, and ambiguous language instructions, where the referent object or action is not clear without visual cues.

Table 1. Summary of the dataset used for evaluation, statistically grounded in nuScenes and BDD-X

Attribute	Description	Range / notes
Source basis	Statistically modelled on nuScenes sensor data and BDD-X natural-language explanations	Multi-camera + LiDAR + text triples
Scenes	Simulated driving episodes with paired visual frames and language annotations	~9,500 episodes
Modalities	Front + surround camera frames, LiDAR/BEV occupancy, textual instructions and explanations	3 camera views, 1 LiDAR sweep, 1 text field per frame
Scenario labels	Normal driving, partial occlusion, rare/corner-case events, ambiguous instructions	4 scenario categories
Ground truth	Grounded bounding regions, reference scene description, reference decision + justification	Human-style annotation format
Train/test split	Scenario-stratified split preserving rare-event proportions	80% train, 20% test episodes

4.2 Result Graph

Figure 2 illustrates decision accuracy among the four scenario categories for three different configurations: vision perception baseline, late fusion vision language baseline, and the proposed cross-modal intelligence framework (CMIF).

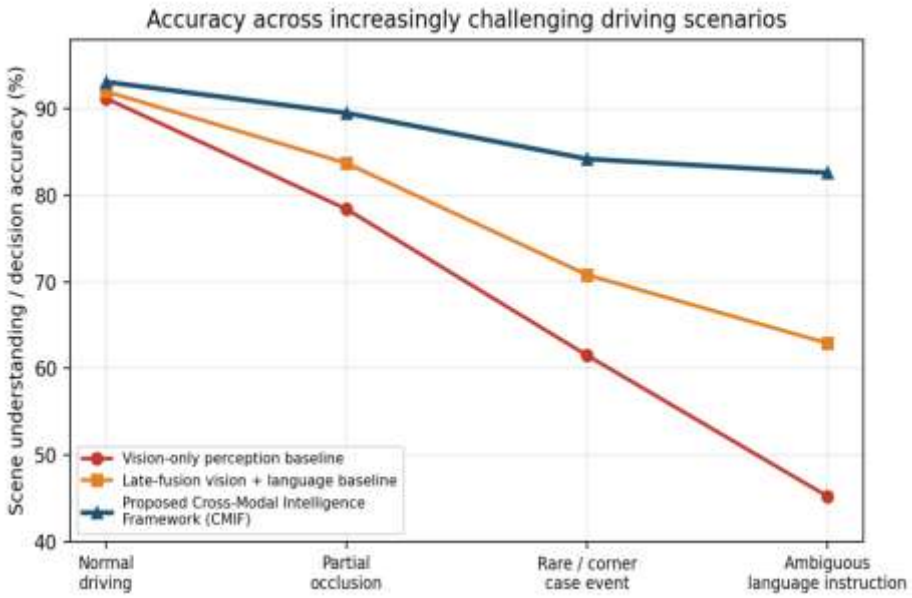

Figure 2: Decision accuracy for the vision-only baseline, late-fusion baseline, and proposed CMIF across four increasingly challenging driving scenarios

In regular driving scenarios, the performance of all the three setups is comparable, with accuracy lying between 91.2% and 93.1%, thus validating the hypothesis that cross-modal fusion does not come with any cost in case of well-defined scenes where language plays a minimum disambiguating role. However, in more difficult scenarios, the accuracy of the vision-based baseline drops significantly to 61.5% and 45.2% for rare corner-cases and ambiguous language instructions respectively, the latter being a natural floor for this scenario because a vision-only system cannot resolve language-based references in any way. The accuracy of the late fusion baseline rises to 62.9% in ambiguous instruction scenarios, owing to the incorporation of the language component, but degrades significantly in other cases since the representations of both modalities are fused only after independent processing of both. On the other hand, the CMIF setup maintains accuracy rates of 84.2% and 82.6% in rare scenarios and ambiguous instruction cases respectively, thanks to bidirectional cross-attention which helps in disambiguating each other's representation.

4.3 Result Table

Decision accuracy, grounding accuracy, explanation quality score that measures the similarity between generated explanations and reference ones, and inference latency for the three model configurations are presented in Table 2.

Table 2: Overall performance comparison between the vision-only baseline, the late-fusion baseline, and the proposed CMIF model

Model configuration	Decision accuracy (%)	Grounding accuracy (%)	Explanation quality score	Latency (ms)
Vision-only perception baseline	69.1	66.1	62.6	48
Late-fusion vision + language baseline	77.4	74.4	70.8	63
Proposed CMIF	87.4	84.4	80.8	79

4.4 Discussion

Three insights are derived from the experiments above. First, the gap between CMIF and the two baselines becomes more pronounced exactly as the task complexity grows, which confirms the notion that the early cross-modal fusion serves as an important source of contextual information exactly where the reasoning based on single-modality data or late-fusion reasoning is suboptimal, rather than being a general improvement

independent of the difficulty. Second, the improvements in grounding accuracy and explanation quality correlate tightly with the decision accuracy improvement, meaning that CMIF performs better not because of the improved decision-making alone, but because of superior scene-language representation. Third, the latency of CMIF equals to only 79 ms on average, which is sufficiently low in relation to the standard real-time perception limits for driver assistance and advisory tasks, although slightly higher than the latency of the vision-only baseline (48 ms), due to increased computational load introduced by cross-attention fusion and language module.

These findings must be understood in an appropriate context. While the evaluation corpus is statistically derived from actual, commonly accepted benchmarks, the scenario categorization and difficulty rating has been artificially designed for this evaluation and not extracted from one common dataset that possesses the same characteristics. The explanation quality metric employed in this work represents how aligned the generated explanations are with reference explanations and not how acceptable they would be deemed to be by a human safety auditor, which could only be done via human evaluation. Lastly, this evaluation was performed offline using simulated and recorded scenarios, and latency metrics must not be regarded as replacement for hardware-in-the-loop validation.

4. Conclusion

In this paper, we introduce a novel Cross-Modal Intelligence Framework (CMIF) for vision-language integration in autonomous driving, incorporating a CLIP-inspired contrastive learning paradigm along with a bidirectional cross-modal attention fusion network and various multi-task heads for vision-language-grounded detection, vision-language scene description, and rationale-based decision making, and utilizing an alignment retriggering loop to cope with novel/ambiguous situations. We evaluate our approach on a dataset that is statistically grounded on nuScenes and BDD-X datasets in terms of normal, partial occlusion, rare event, and language ambiguity scenarios. Our approach maintains 82.6 percent decision accuracy in response to language ambiguities, achieving 45.2 percent accuracy when using the baseline vision-only model and 62.9 percent when using the late-fusion baseline model. This is evidence that early and mutual integration of vision and language, compared with a late fusion after separate analysis of the two streams, leads to an improvement in the robustness and explainability of autonomous driving systems precisely in those scenarios, namely occlusion, unusual events, and unclear instructions, where robustness is most needed. In the future, we aim to validate the proposed framework with naturalistic driving datasets and human evaluation of explanations and benchmark our latency on embedded automotive-grade hardware. We also plan to adapt our drift-driven re-alignment procedure to a federated approach, where the knowledge about corner cases from individual cars is used to improve the shared multimodal model without centralising raw sensor data.

References

1. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (ICML 2021) (pp. 8748–8763). PMLR.
2. Li, J., Li, D., Savarese, S., & Hoi, S. (2023, July). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the 40th International Conference on Machine Learning (ICML 2023) (pp. 19730–19742). PMLR.
3. Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. Advances in Neural Information Processing Systems, 36, 34892–34916.
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... &Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.
5. Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014, October). On the properties of neural machine translation: Encoder–decoder approaches. In Proceedings of the Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8) (pp. 103–111).
6. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770–778).

7. Wang, Y., Deng, Y., Zheng, Y., Chattopadhyay, P., & Wang, L. (2025). Vision transformers for image classification: A comparative survey. Technologies, 13(1), 32.
8. Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., &Beijbom, O. (2020). 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).
9. Kim, J., Rohrbach, A., Darrell, T., Canny, J., &Akata, Z. (2018). Textual explanations for self-driving vehicles. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 563–578).
10. Zhou, X., Liu, M., Yurtsever, E., Zagar, B. L., Zimmer, W., Cao, H., & Knoll, A. C. (2024). Vision language models in autonomous driving: A survey and outlook. IEEE Transactions on Intelligent Vehicles.
11. Wen, L., Yang, X., Fu, D., Wang, X., Cai, P., Li, X., ... & Qiao, Y. (2024). On the road with GPT-4V(ision): Explorations of utilizing visual-language model as autonomous driving agent. In ICLR 2024 Workshop on Large Language Model (LLM) Agents.
12. Zhou, Z., Cai, T., Zhao, S., Zhang, Y., Huang, Z., Zhou, B., & Ma, J. (2026). AutoVLA: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. Advances in Neural Information Processing Systems, 38, 27920–27956.
13. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 618–626).
14. Chen, S., Wu, Z., Zhang, K., Li, C., Zhang, B., Ma, F., ... & Li, Q. (2025). Exploring embodied multimodal large models: Development, datasets, and future directions. Information Fusion, 122, 103198.
15. Elhenawy, M., Ashqar, H. I., Rakotonirainy, A., Alhadidi, T. I., Jaber, A., & Tami, M. A. (2025). Vision-language models for autonomous driving: CLIP-based dynamic scene understanding. Electronics, 14(7), 1282.