P2SIF: Plug-N-Play Sensor Integration Framework

Sreevatsank Kadaveru and Pushkal Mishra, Faculty: Dinesh Bharadia; UC San Diego

I. Introduction

Large Language Models (LLMs), a key development in the field of Artificial Intelligence (AI), have demonstrated advanced reasoning capabilities and an understanding of complex logic [1], [2]. Many applications across a plethora of disciplines benefit from these advanced reasoning capabilities, and with the exponential growth in their size and performance, LLMs are poised to revolutionize numerous fields.

However, many of these applications often require to operate on specialized, domain-specific sensor data. Autonomous Driving (AD), for example, benefits from the use of LiDAR, Radar, GPS, and other modalities for robust perception and situational awareness [3], [4]. However, because the current generation of general-purpose LLMs are trained only on corpora of text and naturally occurring images from the internet, they lack the understanding necessary to comprehend domain-specific sensor data. We illustrate this problem in Figure 2a, where we prompt the state-of-the-art OpenAI o1 and Gemini 1.5 Pro models to interpret LiDAR and Radar data to identify any cars in the scene. Both models fail, identifying cars in a scene with no cars in the vicinity. In contrast, a human expert would easily interpret the radar signatures to understand that there were no cars. This example highlights the shallow domain expertise of LLM models in comprehending domain-specific sensor data. This limitation significantly constrains the applicability of LLMs to tasks that go beyond simple text and camera image prompting.

Current methods of augmenting LLM capabilities with domain knowledge include model retraining, fine-tuning, and adapter layers [5] or techniques such as LoRA [6]. These approaches are expensive, time-consuming, and result in an LLM model (and adapter weights) that is deeply coupled to the application, making them inflexible and non-extensible. Moreover, these techniques require access to the weights of the LLM model, which eliminates the ability to use state-of-the-art proprietary models that are typically closed source. This tight coupling between domain-specific knowledge and the LLM architecture prevents applications from readily benefiting from the rapid advancements in LLM technology, ultimately impeding rapid progress and broader adoption.

We propose a framework, illustrated in Figure 1, that offers a modular platform for sensor fusion. This platform enables the use of LLMs in applications that require complex domain knowledge and operate on domain-specific sensor data. Our architecture cleanly decouples the sensor data processing and application-specific domain expertise from the generalist reasoning processes of LLMs. We achieve this by employing independent specialist models to synthesize

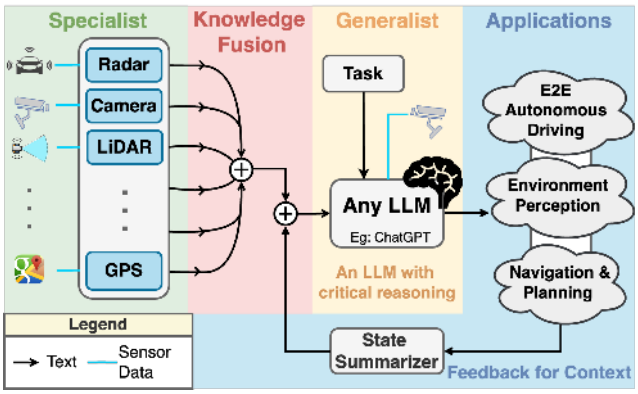


Fig. 1: Plug-N-Play Multi-Modal Sensor Fusion Platform

domain-specific information into comprehensible findings in the form of human-interpretable text. These sensor insights are fused together in the knowledge fusion stage before being passed to the Generalist LLM for decision-making, enabling downstream applications.

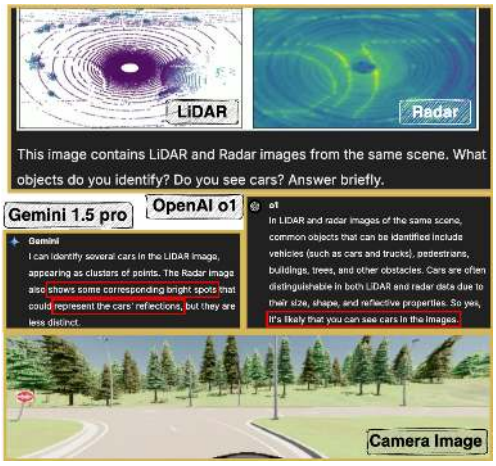
This plug-and-play architecture allows applications to readily benefit from the rapid advancements in LLM technology by upgrading the model without the need for expensive retraining. Additionally, the presence of multiple independent streams and modalities of sensor data enhances the overall system's redundancy and efficacy by guarding against erroneous sensor readings. Moreover, because the knowledge fusion is performed in the form of human-interpretable text, we gain unique insights into the functioning of the system, enhancing its interpretability.

We position this proposal with respect to the state-of-the-art in Section II and elaborate on the key innovations enabling the platform in Section III. In Section IV, we detail the evaluations that demonstrate the merits of this platform. We primarily focus our evaluation on AD applications; however, we contend that this architecture can extend to a diverse set of application domains beyond autonomous driving.

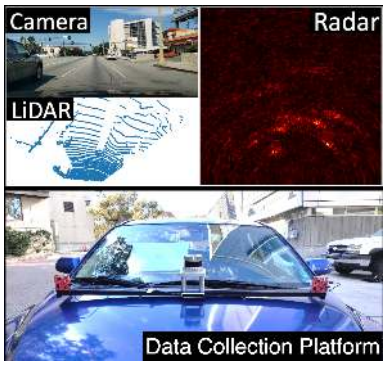
We believe that this novel construct has the potential for far-reaching implications, fundamentally shifting the paradigm of sensor fusion from combining inputs to combining insights, thereby revolutionizing the way we approach sensor integration in the era of LLMs. Our platform could have a broader impact that extends to a wide range of applications where sensor fusion is key!

II. Related Works

LLMs demonstrate remarkable logical reasoning, critical thinking, and decision-making capabilities, and an ability to



(a) LLMs failing to interpret LiDAR/Radar data



(b) Real-World Data collection Platform



(c) Real-World bad weather data from [7]

Fig. 2

transfer these skills to new and out-of-distribution scenarios, making them excellent generalists [8]–[10]. This is largely due to their training on broad and diverse corpora of internet-scale datasets of text and images. Furthermore, recent advancements such as Chain of Thought (CoT) reasoning [11] and self-reflection [12] techniques significantly enhance the quality of reasoning and decision-making abilities of LLMs.

Many applications could significantly benefit from the decision-making capabilities of LLMs. However, as alluded to earlier, the use of LLMs often requires access to weights or expensive training. Recent work has attempted to address this issue by feeding raw sensor data directly into an LLM in the form of arrays of structured text [13] or in the form of summarized trends of the data [14]. However, these methods are limited in the complexity and the amount of sensor data they can handle. Moreover, as shown in [15], default tokenizers of LLMs are not well-suited to handle raw data formatted as text. [16] and [17] attempt to better interface and integrate sensor data with an LLM by building a custom tokenizer that addresses the limitations of representing data in the form of text. However, these methods again require access to the weights of the LLM model. [18] highlights the various AD tasks that benefit from the use of LLMs. In particular, [19] and [20] demonstrate the benefits of LLMs and CoT reasoning for End-to-End (E2E) AD applications.

Our proposal contrasts with the paradigm of integrating sensor data directly into LLMs; rather, we outsource the sensor data processing to an external specialist model.

III. Key Innovations and Methodologies

A. Decoupling of Domain Expert from Generalist Reasoning

Our architecture distinctly separates the domain-specific expertise required for sensor data processing from the broader reasoning capabilities of LLMs. This separation enables a highly modular and extensible system, allowing new sensors to be integrated and the LLM to be upgraded with minimal effort. Additionally, this decoupling accommodates the use of both open-source and closed-source models, providing flexibility in model selection.

B. Specialist Models

The specialist models provide deep domain expertise lacking in general-purpose LLMs. They are trained to handle domain-specific sensor data and transform them into human-interpretable textual representations of insights, akin to an executive report.

A camera specialist can identify critical objects such as vehicles or pedestrians on the road. A LiDAR specialist can be particularly effective at identifying obstacles at night. A radar specialist can provide precise object position, velocity, and motion, even in poor weather. A GPS specialist can inform the model of navigational events such as a turn.

Note that we opt for a dual approach with the camera sensor, including a direct camera feed to the LLM model in addition to the camera specialist insights. This is because the camera data consists of dense semantic information, which in the case of AD includes drivable surfaces, lane markings, and other important road features. This information is critical for making driving decisions; however, this dense semantic information is hard to capture holistically in the form of text.

Generating Data & Training: Specialist models require captioned sensor data for training. However, a multi-modal captioned driving dataset does not exist; therefore, we need to generate one. We obtain captioned data in simulation using CARLA and in the real world using datasets like nuScenes [21].

In simulation, we obtain the sensor data from CARLA, and the captions are generated using the ground truth from the same, substituted into structured caption templates. In the real world, we use image captioning models such as mPLUG [22] and LidarCLIP [23] to generate the textual captions for camera and LiDAR sensors.

We train these specialist models using self-supervised cross-training techniques, such as CLIP loss [24].

C. Knowledge Fusion

The sensor insights from all the specialist models are collated together in the Knowledge Fusion stage, along with inputs from the state summarizer, which provides system state information for closed-loop operations.

A key innovation in our work is the use of text as a universal modality for knowledge representation. This architecture is inspired by hierarchical information flow structures in human organizations, where experts present findings in the form of reports (text), which are reviewed and synthesized by higher-level decision-makers to inform their choices.

IV. Evaluations

We evaluate our platform on three core AD tasks: E2E AD, Environment Perception, and Navigation & Planning, against state-of-the-art models from the CARLA leaderboard using metrics such as the Driving score, mAP scores and Predictive Driver Model score (PDM) respectively.

To showcase the unique benefits of our platform, we will evaluate its performance in the following scenarios:

- **Plug-n-Play Test**: We study the ease and efficacy of adding additional sensors to the system, and also the ability to readily swap out the LLM model.
- **Bad Weather Test**: Here, we evaluate the LLM's ability to integrate critical sensor knowledge fusion capabilities, specifically that of the radar sensor, under adverse driving conditions, such as dense fog and heavy rain (See Figure 2c for real-world bad weather data).
- **Erroneous Sensor Test**: We study the performance of the system in the presence of an erroneous sensor. This test attempts to demonstrate the increased reliability of the system brought about by the sensor fusion of the multiple independent sensor streams and modalities of information. We evaluate the robustness by intentionally injecting errors into the sensor data or sensor insights and assess the correctness of the final decision.
- **Failure Attribution Test**: We assess the interpretability of our platform by analyzing failure cases, such as collisions, focusing on the ease and accuracy of identifying the root cause. Failures may stem from sensor limitations, erroneous omissions by the specialist model, or flawed LLM reasoning. It is essential for system designers to pinpoint actual failure modes and implement targeted solutions. For instance, if a collision was a result of a limited range of the radar sensor, a better radar sensor can be used. Similarly, if the specialist model omitted a critical detection, efforts can be focused on improving the model (e.g., by data augmentation or using a larger model). If failure is attributed to the flawed LLM's reasoning, better prompting and models can be explored. To evaluate attribution, we generate a dataset by intentionally inducing failures by placing the system in controlled, impossible scenarios.

We evaluate our system in simulation using the CARLA simulator and in the real world using datasets collected for past work [7]. This dataset is a rich LiDAR-Radar-Camera dataset collected using a real-world data collection platform, which is a specially modified car fitted with camera, LiDAR, and radar sensors, and an onboard data acquisition and storage system to collect multimodal driving data, as shown in Figure 2b. It consists of 54K radar frames for five real-life traffic and driving scenes. Data was collected under different operating conditions, including day, night, and adverse weather (fog) conditions (See Figure 2c).

Additionally, we plan to leverage this real-world data collection platform to collect any additional data as required and to evaluate our system in the real world.

V. Timeline

The tasks for this project can be broadly categorized as captioned data generation, specialist training and prompt engineering, and real-world evaluations. *Captioned Data Generation* involves gathering the requisite data and textual captions for subsequent model training and requires building a CARLA-based framework for multi-modal sensor data collection and caption generation, in addition to a nuScenes-based captioning framework. *Specialist Training & Prompt Engineering* is where we study the efficacy of specialist model architectures and the effect of different prompting and captioning strategies. In *Real-World Evaluations*, we perform comprehensive evaluations of the models in the real world under normal and adverse weather conditions, which would involve using the Data Collection Platform. Refer to below for timeline.



VI. Team Strength

Sreevatsank Kadaveru and Pushkal Mishra are PhD students in the WCSNG lab at UC San Diego, supervised by Prof. Dinesh Bharadia. Sreevatsank has strong leadership experience delivering on several projects and has extensively worked with Kubernetes, particularly for MLOps on a large scale. Pushkal has worked with the CARLA simulator, training multiple end-to-end autonomous driving models, as demonstrated in his previous work [25], providing him with the necessary expertise in simulation. Both students possess a strong background in machine learning and have demonstrated significant research capacity through their prior work. Furthermore, the WCSNG lab has accumulated a substantial corpus of real-world data and platforms through previous research efforts [4], [7], [25], [26], and has access to a real-world data collection platform, as shown in Figure 2b. This enables them to test their solution in real-world scenarios. This combined skill set, experience, resources of the lab, and strong motivation of both students makes them the best candidates to successfully deliver on this proposal.

REFERENCES

[1] Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang, "Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning," 2023.

[2] Jie Huang and Kevin Chen-Chuan Chang, "Towards reasoning in large language models: A survey," 2023.

[3] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger, " TransFuser: Imitation With Transformer-Based Sensor Fusion for Autonomous Driving ," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 45, no. 11, pp. 12878–12895, Nov. 2023.

[4] Kshitiz Bansal, Keshav Rungta, and Dinesh Bharadia, "Radsegnet: A reliable approach to radar camera fusion," 2022.

[5] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly, "Parameter-efficient transfer learning for nlp," in *International conference on machine learning*. PMLR, 2019, pp. 2790–2799.

[6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[7] Kshitiz Bansal, Keshav Rungta, Siyuan Zhu, and Dinesh Bharadia, "Pointillism: accurate 3d bounding box estimation with multi-radars," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, New York, NY, USA, 2020, SenSys '20, p. 340–353, Association for Computing Machinery.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[9] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping, "Nv-embed: Improved techniques for training llms as generalist embedding models," 2024.

[10] Andrew Szot, Bogdan Mazoure, Omar Attia, Aleksei Timofeev, Harsh Agrawal, Devon Hjelm, Zhe Gan, Zsolt Kira, and Alexander Toshev, "From multimodal llms to generalist embodied agents: Methods and lessons," 2024.

[11] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al., "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022.

[12] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao, "Reflexion: Language agents with verbal reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[13] Sijie Ji, Xinzhe Zheng, and Chenshu Wu, "Hargpt: Are llms zero-shot human activity recognizers?," *arXiv preprint arXiv:2403.02727*, 2024.

[14] Xiaomin Ouyang and Mani Srivastava, "Llmsense: Harnessing llms for high-level reasoning over spatiotemporal sensor traces," *arXiv preprint arXiv:2403.19857*, 2024.

[15] Dimitris Spathis and Fahim Kawsar, "The first step is the hardest: Pitfalls of representing and tokenizing temporal data for large language models," *arXiv preprint arXiv:2309.06236*, 2023.

[16] Zechen Li, Shohreh Deldari, Linyao Chen, Hao Xue, and Flora D Salim, "Sensorllm: Aligning large language models with motion sensors for human activity recognition," *arXiv preprint arXiv:2410.10624*, 2024.

[17] Che Liu, Zhongwei Wan, Sibo Cheng, Mi Zhang, and Rossella Arcucci, "Etp: Learning transferable ecg representations via ecg-text pre-training," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 8230–8234.

[18] Zhenjie Yang, Xiaosong Jia, Hongyang Li, and Junchi Yan, "A survey of large language models for autonomous driving," *arXiv preprint arXiv:2311.01043*, 2023.

[19] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton, "Driving with llms: Fusing object-level vector modality for explainable autonomous driving," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 14093–14100.

[20] Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, et al., "Emma: End-to-end multimodal model for autonomous driving," *arXiv preprint arXiv:2410.23262*, 2024.

[21] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11621–11631.

[22] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si, "mplug: Effective and efficient vision-language learning by cross-modal skip-connections," 2022.

[23] Georg Hess, Adam Tonderski, Christoffer Petersson, Kalle Åström, and Lennart Svensson, "Lidarclip or: How i learned to talk to point clouds," 2023.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[25] Satyam Srivastava, Jerry Li, Pushkal Mishra, Kshitiz Bansal, and Dinesh Bharadia, "C-shenron: A realistic radar simulation framework for carla," https://wcsng.ucsd.edu/c-shenron/, 2024.

[26] Kshitiz Bansal, Gautham Reddy, and Dinesh Bharadia, "Shenron - scalable, high fidelity and efficient radar simulation," *IEEE Robotics and Automation Letters*, vol. 9, no. 2, pp. 1644–1651, 2024.