Hard Cases Detection in Motion Prediction by Vision-Language Foundation Models

Yi Yang[1,2], Qingwen Zhang[1], Kei Ikemura[1], Nazre Batool[2], John Folkesson[1]

[1] KTH Royal Institute of Technology, Stockholm, 114 28, Sweden
[2] Scania AB, Södertälje, 151 87, Sweden

Abstract. Addressing hard cases in autonomous driving, such as anomalous road users, extreme weather conditions, and complex traffic interactions, presents significant challenges. To ensure safety, it is crucial to detect and manage these scenarios effectively for autonomous driving systems. However, the rarity and high-risk nature of these cases demand extensive, diverse datasets for training robust models. Vision-Language Foundation Models (VLMs) have shown remarkable zero-shot capabilities as being trained on extensive datasets. This work explores the potential of VLMs in detecting hard cases in autonomous driving. We demonstrate the capability of VLMs such as GPT-4v in detecting hard cases in traffic participant motion prediction on both agent and scenario levels. We introduce a feasible pipeline where VLMs, fed with sequential image frames with designed prompts, effectively identify challenging agents or scenarios, which are verified by existing prediction models. Moreover, by taking advantage of this detection of hard cases by VLMs, we further improve the training efficiency of the existing motion prediction pipeline by performing data selection for the training samples suggested by GPT. We show the effectiveness and feasibility of our pipeline incorporating VLMs with state-of-the-art methods on NuScenes datasets.

1 Introduction

Recent advancements in deep learning have driven rapid progress in the field of autonomous driving. One of the challenges remains in addressing complex and unpredictable scenarios, such as dealing with unusual road user behaviors, navigating extreme weather conditions, responding to emergencies, and managing intricate interactions. These challenging situations pose substantial safety concerns due to the sparsity in the whole dataset and high variability [23].

Existing approaches to address this include collecting more real-world data [2, 8, 17, 21] or generating synthetic data conditioned on specific needs using generative models [12, 16, 26], or reconstructing 3D environments [10, 18, 25]. However, these methods can be expensive and require substantial human intervention. Incremental learning has limitations in interpretability and sample identification [13, 22]. This raises the question: is there a more **explainable and independent** method available?

The surprising zero-shot capabilities of large language models (LLMs) and vision-language foundation models (VLMs) [1,3,14,15,19,20] have sparked interest in leveraging these models for autonomous driving [5,6,11,24,27]. Rather than replacing the existing pipeline, an intriguing question is how current methods can benefit from integrating these powerful models.

Therefore, in this paper, **we explore leveraging VLMs for detecting hard cases at both the agent-level and scene-level, focusing on motion prediction.** At the agent-level, the goal is to identify road users with unexpected behaviors, which often cause current algorithms to fail resulting in large prediction displacement errors. At the scene-level, it is useful to pinpoint challenging scenarios, such as unusual traffic patterns, emergencies, extreme weather conditions, etc. These scenarios often pose difficulties for existing motion prediction networks. Therefore, having a pipeline that can recognize and be aware of potential failure is crucial.

We summarize our key contributions: (a) We introduce a feasible pipeline to leverage VLM to detect hard cases in autonomous driving contexts; (b) We verify the detection capability of VLM using existing prediction networks; (c) We demonstrate that by detecting hard cases, VLM can facilitate more efficient network training via proper data selection.

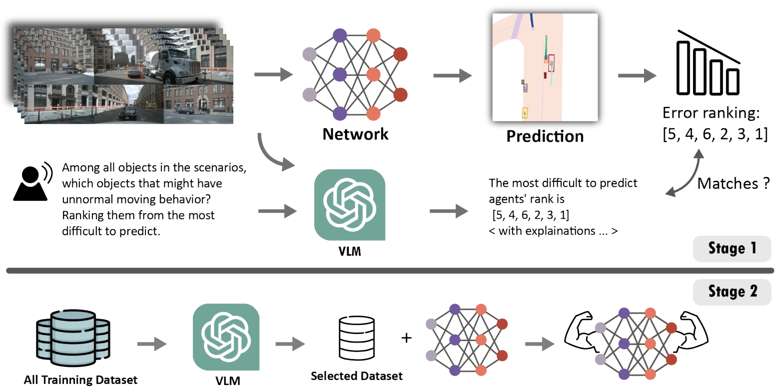

Fig. 1: Two-stages of evaluations. *Stage 1:* Verify the ability of VLM to detect hard cases, using existing motion prediction results as ground truth. We examine if the VLM's prediction of the most difficult-to-predict agents matches the order based on the highest displacement error in existing motion prediction networks. *Stage 2:* Improve training efficiency by training the network with a smaller subset of data selected by VLM.

2 Methodology and Experiments

We evaluate our hypothesis in two stages (Fig. 1).

The first stage verifies the capability of vision-language models (VLMs) to detect hard cases in motion prediction, using existing prediction networks as a benchmark. Our approach focuses on prediction networks that use raw sensor data, particularly camera images, which provide richer visual information for

VLM prompts design with test images.	
a) System	You are a mature driver behind the wheel. You will see consecutive frames. Each frame has six images captured and put together by the surround view camera in your vehicle. The field of view (FOV) of each surround-view camera is approximately 120°. The view angle is written in the images. On the right, it is the bird-eye-view image with labeled agents, centered on your vehicle. The HD map is also plotted in the bird-eye-view image. You need to answer the following questions: Question 1: Among all objects, which objects might have abnormal moving behavior that you might need to pay more attention to? Rank it from the most difficult to predict to the least difficult. Question 2: Score the difficulty of prediction from 1 to 10, where 1 is the easiest and 10 is the most difficult. If the traffic is very dense and at the crossroads or it is bad weather, you might give it a high score. If it is a straight road with few vehicles, it is easy to predict, you might give it a low score. If there is an anomalous behavior or situation, you might give it a high score. Present your answers in this format: 1. The most difficult to predict agents' rank is [A1, A2, A3, A4, A5, A6, A7, A8]. <explanation> Replace A1-A8 with the agent ID number, and make sure the output includes all agents. 2. Overall, the prediction difficulty is <score>. From 1 to 10. <explanation>
b) Assistant	[*A few shots are provided to instruct GPT-4V.*]
c) User	[*Evaluate on test images.*] There are 5 different IDs of agents in these images. From ID 1 to ID 5, they are pedestrian, pedestrian, car, motorcycle, and pedestrian. 
GPT-4v [14]	Output: 1: The most difficult to predict agents' rank is [4, 1, 2, 5, 3]. Explanation: Motorcycle 4 is in motion and could change speed or direction suddenly, making it unpredictable. Pedestrian 1 is close to the crosswalk and might decide to cross the street, which requires attention. Pedestrian 2 is also near the crosswalk and could potentially enter the street. Pedestrian 5 is walking along the sidewalk and seems less likely to enter the street suddenly. Car 3 is parked and not showing any signs of movement, making it the least difficult to predict. 2: Overall, the prediction difficulty is 5. Explanation: The road is not very busy, and most agents are behaving predictably. However, the presence of pedestrians near the crosswalk and a moving motorcycle increases the level of caution required.

Table 1: Designed prompts for our approach. a) GPT-4v [14] is instructed to answer two questions with certain format. b) Few-shot learning: two examples are provided to VLM to learn. c) Given inputs that include 3 frames of consecutive camera images from six different views and their description, GPT-4v outputs the answer in a required format along with explanations. Here we show a real output from GPT-4v, where it correctly associates most agents with corresponding marks and types, and reasons about their respective states. In addition, descriptive texts are given to reason the generated ranking of agents and difficulty score of the scene, showing superior explainability.

hard case detection. For each image sequence, the prediction models generate agent trajectories, which are ranked by average displacement error (ADE) from ground truth. Concurrently, the same images are fed into GPT-4v [14] VLM alongside prompts to rank agents by prediction difficulty, score scenario difficulty, and provide explanations as demonstrated in Tab. 1. Few-shot learning [4] is used to enhance VLM accuracy. We assess VLM's accuracy in mirroring the difficulty rankings of the existing prediction model using various ranking metrics. The results in Fig. 2, show that GPT-4v demonstrates robust performance across all evaluation metrics, outperforming over 90% of random rankings.

In the second stage, the study demonstrates the utility of VLM-based hard case detection for efficient data selection. A subset of difficult scenes identified by the VLM (Fig. 3) is used to train prediction models, and their performance is compared to models trained on the full dataset. Detailed quantitative results are provided in the original paper. The approach tests VLMs' potential to improve training efficiency by creating a smaller, yet representative portion of the training dataset.

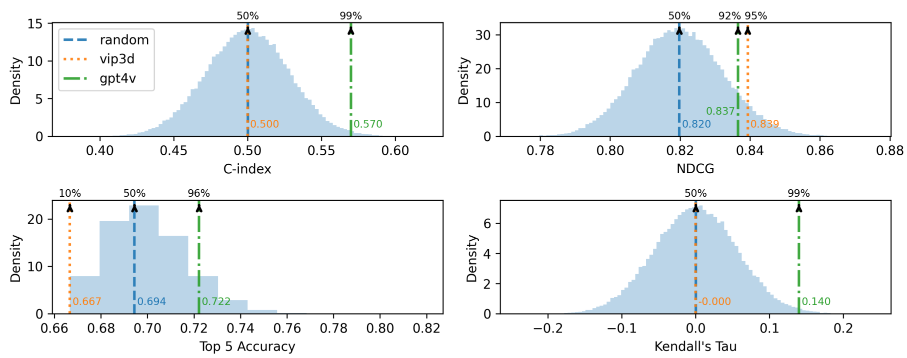

Fig. 2: Result of agents ranking according to higher prediction error / difficulty. Using the UniAD [9] ranking as ground truth, we compare it with random order, order from ViP3D [7], and GPT-4v. The evaluation is conducted using four metrics: C-index, NDCG, top-5 accuracy, and Kendall's Tau, where larger values indicate a higher correlation with UniAD order. The x-axis is the metric value. Note that for the random ordering, we conducted 10,000 trials, and the distribution of the results is shown in the blue histogram, with the y-axis representing the probability density / frequency; note that the metric value of random is the mean from all trials. The percentage values above the graph indicate the percentage of random trials that are surpassed by this value (cumulative probability).

(a) GPT-4v: Prediction difficulty is 9. Nighttime and wet conditions reduce visibility and alter vehicle behavior, with light reflections complicating movement prediction of other road users. The scene's complexity is heightened by multiple vehicles and a pedestrian crossing.

(b) GPT-4v: Prediction difficulty is 7. The intersection setting includes traffic lights and crosswalks. A large truck could block views and impede movement, alongside multiple vehicles and pedestrians, increasing prediction difficulty due to potential blind spots and varied road users.

Fig. 3: Two real examples scored by GPT-4v. Higher scores denote greater difficulty.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)
2. Alibeigi, M., Ljungbergh, W., Tonderski, A., Hess, G., Lilja, A., Lindstrom, C., Motorniuk, D., Fu, J., Widahl, J., Petersson, C.: Zenseact open dataset: A large-scale and diverse multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
5. Cui, C., Ma, Y., Cao, X., Ye, W., Zhou, Y., Liang, K., Chen, J., Lu, J., Yang, Z., Liao, K.D., et al.: A survey on multimodal large language models for autonomous driving. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 958–979 (2024)
6. Ding, X., Han, J., Xu, H., Zhang, W., Li, X.: Hilm-d: Towards high-resolution understanding in multimodal large language models for autonomous driving. arXiv preprint arXiv:2309.05186 (2023)
7. Gu, J., Hu, C., Zhang, T., Chen, X., Wang, Y., Wang, Y., Zhao, H.: Vip3d: End-to-end visual trajectory prediction via 3d agent queries. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5496–5506 (2023)
8. H. Caesar, J. Kabzan, K.T.e.a.: Nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. In: CVPR ADP3 workshop (2021)
9. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., et al.: Planning-oriented autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17853–17862 (2023)
10. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023), `https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/`
11. Keysan, A., Look, A., Kosman, E., Gürsun, G., Wagner, J., Yu, Y., Rakitsch, B.: Can you text what is happening? integrating pre-trained language encoders into trajectory prediction models for autonomous driving. arXiv preprint arXiv:2309.05282 (2023)
12. Marathe, A., Ramanan, D., Walambe, R., Kotecha, K.: Wedge: A multi-weather autonomous driving dataset built from generative vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3317–3326 (2023)
13. Mirza, M.J., Masana, M., Possegger, H., Bischof, H.: An efficient domain-incremental learning approach to drive in all weather conditions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3001–3011 (2022)
14. OpenAI, :, Achiam, J., et al., S.A.: Gpt-4 technical report (2023)

15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
16. Song, Z., He, Z., Li, X., Ma, Q., Ming, R., Mao, Z., Pei, H., Peng, L., Hu, J., Yao, D., et al.: Synthetic datasets for autonomous driving: A survey. arXiv preprint arXiv:2304.12205 (2023)
17. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
18. Tonderski, A., Lindström, C., Hess, G., Ljungbergh, W., Svensson, L., Petersson, C.: Neurad: Neural rendering for autonomous driving. arXiv preprint arXiv:2311.15260 (2023)
19. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
20. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
21. Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J.K., Ramanan, D., Carr, P., Hays, J.: Argoverse 2: Next generation datasets for self-driving perception and forecasting. In: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021) (2021)
22. Yang, B., Fan, F., Ni, R., Li, J., Kiong, L., Liu, X.: Continual learning-based trajectory prediction with memory augmented networks. Knowledge-Based Systems **258**, 110022 (2022)
23. Yang, K., Li, B., Shao, W., Tang, X., Liu, X., Wang, H.: Prediction failure risk-aware decision-making for autonomous vehicles on signalized intersections. IEEE Transactions on Intelligent Transportation Systems (2023)
24. Yang, Y., Zhang, Q., Li, C., Marta, D.S., Batool, N., Folkesson, J.: Human-centric autonomous systems with llms for user command reasoning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 988–994 (2024)
25. Yang, Z., Chen, Y., Wang, J., Manivasagam, S., Ma, W.C., Yang, A.J., Urtasun, R.: Unisim: A neural closed-loop sensor simulator. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1389–1399 (2023)
26. Yang, Z., Chai, Y., Anguelov, D., Zhou, Y., Sun, P., Erhan, D., Rafferty, S., Kretzschmar, H.: Surfelgan: Synthesizing realistic sensor data for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11118–11127 (2020)
27. Zhou, X., Liu, M., Zagar, B.L., Yurtsever, E., Knoll, A.C.: Vision language models in autonomous driving and intelligent transportation systems. arXiv preprint arXiv:2310.14414 (2023)