SFF Rendering-Based Uncertainty Prediction using VisionLLM

Junyong Lee[1]*, Jeihee Cho[2], Ho Suk[2], Shiho Kim[2]

[1] BK21 Graduate Program in Intelligent Semiconductor Technology, Yonsei University, Korea
[2] Yonsei University, Korea
jjunilee@yonsei.ac.kr, jeiheec@gmail.com, sukho93@yonsei.ac.kr, shiho@yonsei.ac.kr

Abstract

In this work, we propose a novel framework for uncertainty prediction in autonomous driving using VisionLLM. Leveraging driving data collected from the CARLA simulator, we generate bird's-eye-view (BEV) images paired with next driving actions and uncertainty scores. To emulate real-world challenges, occlusion masks are introduced to the BEV images, representing regions of limited visibility due to sensor constraints. Our model predicts both the next driving action and uncertainty score, utilizing additional image inputs to enhance its reasoning capability under occlusion-rich conditions. By fine-tuning VisionLLM with Parameter-Efficient Fine-Tuning (PEFT) techniques such as LoRA, we demonstrate the efficacy of our approach in addressing occlusion-based uncertainty, paving the way for safer and more reliable decision-making in high-level driving automation systems.

Introduction

With the introduction of GPT (Radford 2018) and BERT (Devlin 2018) in 2018, large language models (LLMs) emerged as a transformative force in natural language processing research. Within a few years, the release of ChatGPT in 2022, based on GPT-3.5, showed remarkable achievements in performance. The widespread success of ChatGPT drew significant attention to the expansive potential of LLMs, leading to their active adoption in various applications, including automated agents. In the domain of autonomous vehicles, there have been growing efforts to integrate LLMs as key components of planning systems (Cui et al. 2024).

Vision language models (VLMs) are designed to combine the strengths of visual and textual modalities, enabling them to analyze and reason about complex, multimodal inputs. Recently, VisionLLM (Wang et al. 2024) has demonstrated promising performance across various vision-centric tasks by effectively integrating image and text understanding. Its performance, however, is susceptible to the type and quality of the visual inputs provided. The choice of input images significantly impacts its ability to make accurate predictions or perform effectively in downstream tasks.

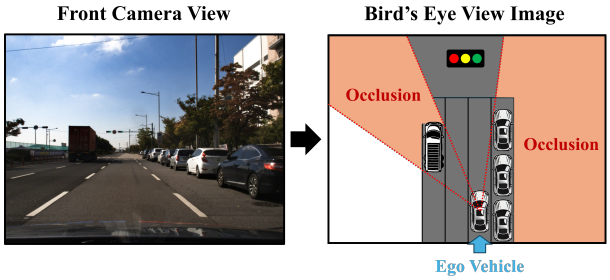

Figure 1: Illustration of the occlusion. The left image shows the front camera view, highlighting a partially visible road blocked by a truck. The right image represents the corresponding bird's-eye-view (BEV) with occlusion areas marked in red.

Several studies have explored generating text-based descriptions of driving environments to use as input for LLMs. However, describing complex driving scenarios using text alone remains a significant challenge. Consequently, autonomous driving applications increasingly adopt VLMs as a preferred solution. Nevertheless, these applications encounter limitations analogous to those of VLMs, highlighting the crucial importance of selecting suitable image representations—such as bird's-eye-view (BEV) maps, panoramic views, or sensor fusion outputs—to achieve robust performance in real-world scenarios.

In this work, we propose a novel framework that leverages VLM for autonomous driving applications by BEV images with corresponding driving actions (next state) and uncertainty scores as training data. The BEV images represent the spatial layout of the environment, including vehicle positions, past and future trajectories, and acceleration states. To simulate real-world challenges, we introduce occlusion masks on the BEV images to emulate uncertainty caused by limited visibility. A key contribution of our work is the model's ability to predict uncertainty scores not only from the BEV image itself but also by leveraging additional image inputs, enhancing its ability to reason under occlusion-rich conditions. By fine-tuning VLM on this dataset, we demonstrate its capability to predict both the next driving action and the associated uncertainty, addressing critical challenges

*These authors contributed equally.

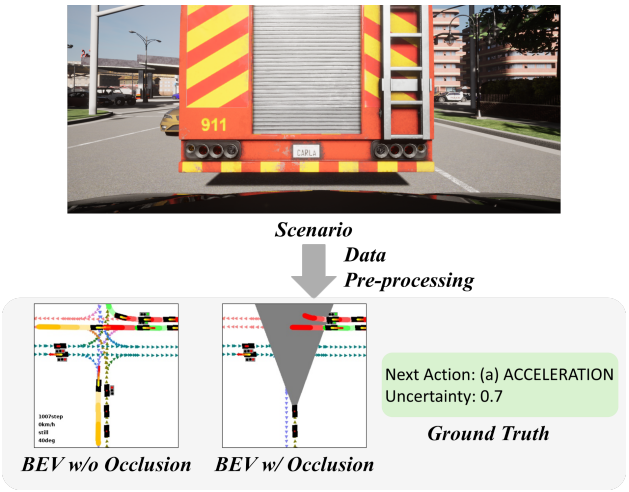

Figure 2: An illustration of data pre-processing with a given CARLA simulation scenario. The color scheme represents the planned trajectory over a 3-second interval, where green corresponds to acceleration, yellow indicates constant speed, and red signifies deceleration.

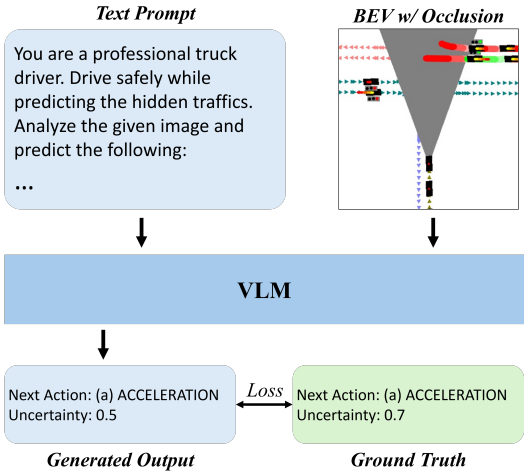

Figure 3: Overview of fine-tuning VLM using processed dataset. The VLM is trained by comparing with the ground truth and generated output.

in planning for autonomous vehicles. This work highlights the potential of integrating vision and language models to improve safety and reliability in high-level driving automation.

Related Works

Planning using LLM in Autonomous Driving Early studies attempting to adopt LLM for planning in autonomous driving used text input describing the driving environment. Mao et al. (2023) proposed GPT-Driver, which approaches the motion planning problem in driving by converting it into language modeling problem. Sha et al. (2023) proposed LanguageMPC, which is a method to determine low-level actions by converting the high-level decision made by LLM that received the text description of driving environment into mathematical parameter matrix. Wang et al. (2023) also proposed an approach similar to Sha et al. in that they use text description input and MPC. (Fu et al. 2024) used Llama-Adapter (Zhang et al. 2023) to convert image into text and then provide it as input to LLM. Chen et al. (2024) proposed DrivingQA, which embeds vector representations containing information about vehicles, pedestrians, and routes into LLM so that the policies can be decoded.

Planning using VLM in Autonomous Driving Beyond using only text descriptions of driving situations as input, studies have also been published that applied VLMs using image input. Wen et al. (2023) and Xu et al. (2024) presented an approach to utilize GPT4, which is provided with multimodal input, for applications in autonomous driving. However, previous studies only evaluate it in a simple highway environment (Wang et al. 2023; Fu et al. 2024) or use raw images of video as input (Wen et al. 2023; Xu et al. 2024).

Safety Force Field (SFF) Nistér et al. (2019b,a) proposed the Safety Force Field (SFF), which mathematically models that collisions do not occur if all vehicles on roads comply with safe control to ensure that the actor itself does not cause unsafe situations. In detail, actor can drive along a claimed set, which is a trajectory determined by a safety procedure, a family of control policies. According to the chain rule for safety potential, if all actors follow the safety procedure, it is guaranteed that the safety potential will not increase, and thus no unsafe situation will occur. The safety potential is the size of the intersection between the claimed sets of the actors, and the goal of the safety procedure is to minimize it.

Uncertainty due to Occlusion SFF simply addresses the visibility problem in autonomous driving by assuming that there may be invisible actors moving reasonably beyond occlusion or sight range 1. Similarly, Koschi and Althoff (2020) also consider the presence of phantom traffic participants in the occlusion caused by the field of view and line of sight of the ego vehicle. However, to perform safe planning even in long tail cases due to occlusion, more sophisticated processing is needed rather than simple assumptions.

Vision-language Models Models like LLaVA and Llama 3.2 have significantly influenced recent advances in VLMs (Liu et al. 2024; Dubey et al. 2024). LLaVA integrates a vision encoder with a LLM to enable comprehensive visual and linguistic understanding, reaching 85.1% of GPT-4's performance on multi-modal instruction-following tasks. Similarly, Llama 3.2 extends LLM capabilities by incorporating multilingual support, coding, reasoning, and tool usage, with its largest model comprising 405 billion parameters and a context window of up to 128,000 tokens. These developments underscore the potential for integrating visual and textual modalities within large-scale language models.

System prompt

You are a professional truck driver. Drive safely while predicting the hidden traffics.

User

[IMAGE]
Analyze the given image and predict the following:
1. Next state (action to take):
Choose one of [(1) ACCELERATION / (2) CONSTANT / (3) DECELERATION / (4) STILL].
2. Uncertainty
(confidence level of the prediction, range: 0 to 1):
Provide a numerical value.
Provide the prediction as:
- Next state: [your prediction]
- Uncertainty: [your prediction]

Assistant

- Next state: {action}
- Uncertainty: {uncertainty}

Figure 4: Illustration of the system prompt used for decision-making tasks in a simulated truck-driving scenario. The model analyzes an image to predict the next driving action (e.g., acceleration, constant speed, deceleration, or staying still) and provides an uncertainty score representing the confidence level of the prediction (range: 0 to 1).

Method

BEV Image Data Generation We collected driving data in the urban environment of the CARLA simulator Dosovitskiy et al. (2017) and generated BEV images based on this data. In the BEV image, the black rectangles represent the position and heading of vehicles, the lines behind the vehicles represent the paths the vehicles have traveled for 1 second, and the polygons in front of the vehicles represent the future trajectory for 3 seconds, corresponding to the claimed set defined in SFF. The color of the claimed set indicates the acceleration state of the vehicle. Green indicates acceleration, yellow indicates constant speed, and red indicates deceleration. A claimed set representing a 3-second future trajectory is drawn in 3 parts separated by 1 second intervals. Each lane on the road is represented by an virtual wedge-shaped center line that vehicles follow. To distinguish between roads when they intersect, roads are painted in different colors depending on direction.

Occlusion Mask Generation We define uncertainty as the inability to perceive a part of the driving environment because the partial field of view is blocked due to the sensor's line of sight issue. We generate an occlusion mask centered on the ego vehicle at random orientation and angle. The occlusion is represented by a gray mask, which obscures everything, including road structures. If the center of the vehicle is inside the mask, the vehicle is judged to be occluded and is not expressed, and the claimed set of the vehicle occluded by the mask is also not expressed at all.

Dataset Pre-processing For fine-tuning, we utilized a dataset collected from 125 vehicles driving in Carla Town 03, generating a total of 600,000 samples amounting to 33GB of data. Each sample in the dataset consists of (1) a BEV image representing the spatial layout of the environment, (2) the next driving action to be taken, and (3) a confidence score ranging from 0 to 1, indicating the level of occlusion in the corresponding scene. Due to the large dataset size, we performed random sampling to select 10,000 samples for training and evaluation, ensuring a diverse representation of the driving scenarios within the environment. This preprocessing step enabled efficient model fine-tuning without compromising the dataset's integrity.

Model Architecture The proposed model leverages both Llama 3.2 (11B Vision-Instruct) and LLaVa v1.6 (Mistral 7B Vision-Instruct) models to achieve a robust architecture capable of processing visual and textual inputs (AI 2024; Liu 2024). These models are built upon a base language model that has been fine-tuned using instruction-following data, enabling it to perform effectively in a conversational, chat-based format. The model takes as input a BEV image with occlusions and a simple text prompt describing the scenario. The Vision Encoder processes the BEV image, extracting high-level visual features, which are combined with the text prompt and passed to the LLM. The LLM generates two outputs: (1) the next driving action, which is one of four predefined options—ACCELERATION, CONSTANT, DECELERATION, or STILL, and (2) the confidence score (uncertainty) associated with the prediction. This architecture effectively integrates multimodal inputs to deliver context-aware driving actions, demonstrating the synergy between large-scale vision-language models.

Fine-tuning Details For fine-tuning, we utilized Parameter-Efficient Fine-Tuning (PEFT) with the LoRA (Low-Rank Adaptation) method to adapt two large pre-trained models (Hu et al. 2021; Mangrulkar et al. 2022): Llama 3.2 (11B Vision-Instruct) and LLaVa v1.6 (Mistral 7B Vision-Instruct). Instead of full fine-tuning, LoRA enabled to efficient training of task-specific parameters with reduced computational cost.

The fine-tuning was conducted on a single NVIDIA A100 GPU (80GB) with a learning rate of 1e-5 over 3 epochs. The training data consisted of BEV images paired with corresponding next driving actions and uncertainty scores.

Figure 4 shows the entire prompt. Validation was performed during training to monitor performance, and a batching strategy of padding was applied to handle variable input sizes. The training used a batch size of 2, while for LLaVa, a batch size of 1 was employed, combined with gradient accumulation steps of 2 to optimize memory usage. Checkpoints for both models were saved in their respective directories, allowing efficient evaluation and further analysis.

This setup demonstrates the feasibility of fine-tuning large-scale vision-language models on a single GPU, leveraging PEFT techniques like LoRA to achieve task-specific optimization with minimal hardware resources.

Table 1: Hyperparameter Settings for Fine-Tuning

Hyperparam	**Llama 3.2**	**LLaVa v1.6**	Qwen2 VL
Model size	11B	7B	7B
Learning rate		1e-5	
Epochs		3	
Batch size		2	
Optimizer		AdamW	
Scheduler		Linear decay	
PEFT		LoRA	

Experimental Results

We evaluate the fine-tuned models on two key metrics: action prediction accuracy and the uncertainty gap, which represents the absolute difference between the predicted uncertainty and the ground truth. As shown in Table 2, the LLaMA 3.2 model achieved the best overall performance, demonstrating its ability to effectively learn both action prediction and uncertainty simultaneously. The results indicate that LLaMA 3.2 is more effective in both predicting the next action and aligning its uncertainty prediction.

The results indicate that VLMs successfully learned to predict both the next driving action and uncertainty simultaneously. This demonstrates the potential of utilizing VLMs for driving uncertainty prediction, even when employing a simplified model and a streamlined dataset. These findings highlight the feasibility of this approach under constrained settings and validate the effectiveness of fine-tuning. Furthermore, they underscore the importance of model architecture in achieving robust performance across multiple tasks within complex, real-world scenarios.

Table 2: Accuracy of next action and uncertainty and the absolute gap between predicted uncertainty and ground truth.

Model	**Action**	**Uncertainty Gap**
Llama 3.2	63.8%	0.03
Qwen2 VL	8.7%	0.03
LLaVa v1.6	17.4%	0.18

Future Works

Future Works This study generates occlusion with arbitrary orientations and angles to simulate real-world driving scenarios. However, several enhancements could further improve the modeling of occlusion and the representation of the driving environment:

- *Model-Specific Prompt Engineering*:We designed dataset prompts specifically for the LLaMA 3.2 model to optimize its performance in action prediction and uncertainty estimation. To ensure fair comparisons with other models, such as LLAVA v1.6, it is necessary to develop tailored prompts that align with the strengths and architecture of each model.
- *Sophisticated Occlusion Modeling*: Current occlusion masks are simplified and do not account for the extent to which surrounding objects interfere with the sensor's line of sight. More precise modeling of occlusions, based on the geometry and positions of objects relative to the ego vehicle, could more accurately emulate real-world scenarios and better address uncertainty issues caused by occlusions.
- *Enhanced BEV Image Representation*: The current BEV images abstract the driving environment with lanes represented only as virtual center lines. To enrich the representation, additional features such as the edges of road structures could be included, providing a more comprehensive view of the environment.
- *Adaptive BEV Image Shapes*: The current square-shaped BEV images could be adapted to better align with sensor sight distances. For example, circular BEV images could be used to reflect the radial visibility range of common automotive sensors more naturally.

These improvements would contribute to a more realistic and detailed simulation of driving environments, enhancing the ability of large language models to reason about uncertainty and make robust prediction.

Discussions and Conclusion

This study explores the integration of VLMs into autonomous driving systems for uncertainty prediction. By leveraging BEV images encoded with task-specific information alongside carefully crafted textual prompts, we show an effective framework for enabling VLMs to comprehend complex driving scenarios. These findings underscore the critical role of designing visual inputs that encapsulate relevant environmental information and pairing them with appropriately constructed textual prompts to enhance the model's reasoning capabilities.

Our results reveal that VLMs exhibits strong inherent capabilities. However, its effectiveness in uncertainty prediction is significantly enhanced when trained to infer uncertainty through environmental understanding. Relying on pre-defined or inherent uncertainty values alone proves less effective. This approach better aligns with the dynamic and context-rich challenges of real-world driving. Factors such as occlusion and sensor limitations critically impact decision-making in these scenarios.

We successfully fine tuning VLMs to predict both the next driving actions and associated uncertainties with high accuracy under occlusion-rich conditions. This work highlights the potential of VLMs as a core component in planning systems for autonomous driving, offering robust performance in safety-critical applications.

Future research directions include refining visual input representations, such as incorporating road-edge structures or adopting dynamic BEV shapes tailored to sensor-specific

sight distances. Furthermore, sophisticated occlusion modeling techniques that more accurately emulate real-world obstructions could further enhance the model's predictive performance. These advancements pave the way for more reliable and safer autonomous driving systems.

Acknowledgments

This research was supported by Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT(MSIT, Korea)&Gwangju Metropolitan City.

References

AI, M. 2024. LLaMA 3.2 11B Vision-Instruct Model. https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct. Accessed: 2024-11-28.

Chen, L.; Sinavski, O.; Hünermann, J.; Karnsund, A.; Willmott, A. J.; Birch, D.; Maund, D.; and Shotton, J. 2024. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 14093–14100. IEEE.

Cui, C.; Ma, Y.; Cao, X.; Ye, W.; Zhou, Y.; Liang, K.; Chen, J.; Lu, J.; Yang, Z.; Liao, K.-D.; et al. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 958–979.

Devlin, J. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*, 1–16. PMLR.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Fu, D.; Li, X.; Wen, L.; Dou, M.; Cai, P.; Shi, B.; and Qiao, Y. 2024. Drive like a human: Rethinking autonomous driving with large language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 910–919.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Koschi, M.; and Althoff, M. 2020. Set-based prediction of traffic participants considering occlusions and traffic rules. *IEEE Transactions on Intelligent Vehicles*, 6(2): 249–265.

Liu, H. 2024. LLAVA v1.6 Mistral 7B Model. https://huggingface.co/liuhaotian/llava-v1.6-mistral-7b. Accessed: 2024-11-28.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Mangrulkar, S.; Gugger, S.; Debut, L.; Belkada, Y.; Paul, S.; and Bossan, B. 2022. PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods. https://github.com/huggingface/peft.

Mao, J.; Qian, Y.; Ye, J.; Zhao, H.; and Wang, Y. 2023. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*.

Nistér, D.; Lee, H.-L.; Ng, J.; and Wang, Y. 2019a. An introduction to the safety force field. *NVIDIA White Paper*.

Nistér, D.; Lee, H.-L.; Ng, J.; and Wang, Y. 2019b. The safety force field. *NVIDIA White Paper*, 15.

Radford, A. 2018. Improving language understanding by generative pre-training.

Sha, H.; Mu, Y.; Jiang, Y.; Chen, L.; Xu, C.; Luo, P.; Li, S. E.; Tomizuka, M.; Zhan, W.; and Ding, M. 2023. Languagempc: Large language models as decision makers for autonomous driving. *arXiv preprint arXiv:2310.03026*.

Wang, W.; Chen, Z.; Chen, X.; Wu, J.; Zhu, X.; Zeng, G.; Luo, P.; Lu, T.; Zhou, J.; Qiao, Y.; et al. 2024. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36.

Wang, Y.; Jiao, R.; Zhan, S. S.; Lang, C.; Huang, C.; Wang, Z.; Yang, Z.; and Zhu, Q. 2023. Empowering autonomous driving with large language models: A safety perspective. *arXiv preprint arXiv:2312.00812*.

Wen, L.; Yang, X.; Fu, D.; Wang, X.; Cai, P.; Li, X.; Ma, T.; Li, Y.; Xu, L.; Shang, D.; et al. 2023. On the road with gpt-4v (ision): Early explorations of visual-language model on autonomous driving. *arXiv preprint arXiv:2311.05332*.

Xu, Z.; Zhang, Y.; Xie, E.; Zhao, Z.; Guo, Y.; Wong, K.-Y. K.; Li, Z.; and Zhao, H. 2024. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*.

Zhang, R.; Han, J.; Liu, C.; Gao, P.; Zhou, A.; Hu, X.; Yan, S.; Lu, P.; Li, H.; and Qiao, Y. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.