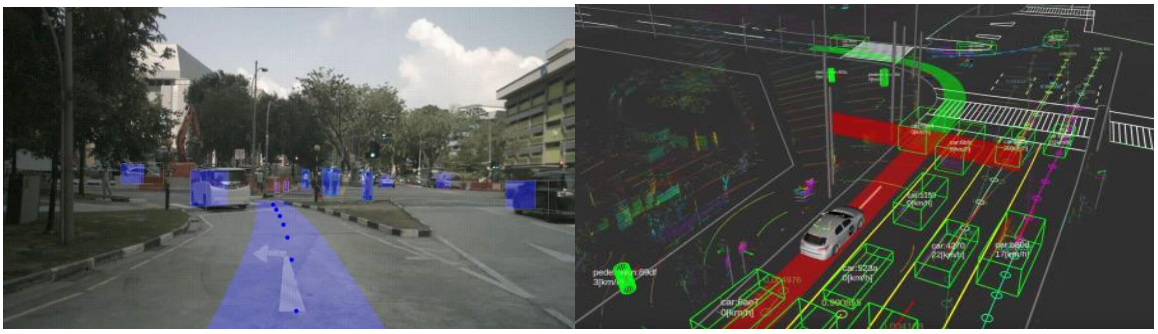
Bachelorarbeit: Technische Hochschule Augsburg / Fakultät für Elektrotechnik (Driverless Mobility Group) / TTZ Landsberg

Bridging Autoware and Vision-Language Models: Comparative Evaluation for Trajectory Planning in Autonomous Driving



Hintergrund und Motivation:

Autonomous driving today still relies on modular stacks like Autoware, where perception, prediction, and planning operate as separate components. This architecture is robust and well understood, but it can be limiting in situations that depend on broader contextual reasoning or high-level scene interpretation. Recent Vision-Language Models (VLMs), such as GPT-4o, GPT-5, InternVL-3.5, Qwen-3-VL, LLaVA, and Llama-3.2-Vision, are beginning to demonstrate a different kind of capability. They can process raw camera images, describe complex traffic scenes, reason about interactions, and even produce coarse trajectory suggestions. With frameworks like OpenEMMA, these models have become accessible, open-source, and easy to integrate.

Although some early studies draw comparisons between VLM based approaches and systems like Apollo, evaluation involving Autoware and modern Vision-Language Models in realistic, closed-loop settings remains very limited. Existing work, such as DriveMLM, is largely confined to simulation with earlier model generations, leaving open how today's state-of-the-art VLMs perform when examined alongside Autoware's perception and planning outputs, particularly under conditions that reflect real vehicle use.

This project fills that gap. It asks a simple but important question: How far can today's VLMs go when placed next to a mature autonomous driving stack, and where do their strengths or limits actually lie?

For students, this offers the rare chance to work at the intersection of robotics and foundation models, supported by a full research environment at Driverless Mobility Group/TTZ Landsberg including our autonomous research vehicles (drive-by-wire), the complete Autoware AD stack with sensors and compute systems, GPU infrastructure for neural network workflows, extensive recorded and open-source datasets, access to our workshop and restricted ADAC test area, and a highly supportive team.

Technische Hochschule Augsburg

Ziel der Arbeit:

The objective of this thesis is to **systematically evaluate modern Vision-Language Models as a redundant or comparative planning mechanism alongside Autoware**, using both recorded sensor data and closed-loop testing environments. The work aims to reveal how well state-of-the-art VLMs can:

- understand complex traffic scenes,
- reason about dynamic interactions, and
- produce meaningful trajectory suggestions

Teilaufgaben:

- A brief literature review of current Vision-Language Models (VLMs) used in autonomous driving.
- Review modular autonomous driving systems, with a focus on Autoware's perception and planning pipeline.
- Install and explore the OpenEMMA framework, including one or more supported VLMs.
- Run VLMs on selected driving scenes to generate scene descriptions and trajectory suggestions.
- Set up Autoware with provided recorded scenarios and extract its perception and planning outputs.
- Align VLM outputs and Autoware outputs for identical scenes to enable direct comparison.
- Compare both systems in terms of scene understanding, dynamic reasoning, and trajectory quality.
- Summarize findings with clear examples, visualizations, and short case studies.
- Conclude with recommendations for potential follow-up work, including extending the comparison toward closed-loop evaluation or preparing results suitable for publication.

Abschluss:

- A functioning evaluation framework comparing Autoware with VLM-based methods.
- Benchmarks on scene understanding and trajectory prediction.
- Initial closed-loop findings and recommendations.
- A high-quality written thesis and reproducible codebase.
- Optional: Publication-ready results for a workshop or conference.

Betreuung:

Prof. Dr.-Ing. Carsten Markgraf, Gautam Kumar Jain (Doktorand) (+4915259007167)