RESEARCH ARTICLE

Vision-Language Foundation Models for Zero-Shot Autonomous Driving Scene Understanding and Risk Assessment

Yue Wang, Hang Zhao, Laura Leal-Taixé

Published: 2026-05-25 | FAIDS Vol. 1, No. 1 (2026)

Abstract: Autonomous driving perception systems trained on fixed taxonomies fail when encountering novel objects or unusual scenarios not represented in their training data — the "long tail" problem. We present DriveLM, a vision-language model (VLM) fine-tuned on 2.8 million driving scene-narration pairs that performs zero-shot risk assessment by generating natural language scene descriptions and structured risk scores. On the nuScenes-QA benchmark, DriveLM achieves 78.4% accuracy on novel-object-related questions (vs. 31.2% for CLIP-based baselines) and a 0.91 Spearman correlation with human risk ratings. In closed-loop CARLA simulation, DriveLM-guided planning reduces collision rate by 42% in rare-event scenarios compared to end-to-end learned planners.

1. Introduction

Current autonomous driving perception pipelines rely on object detectors trained on closed-set taxonomies — typically 10-30 predefined categories. However, real-world driving involves an unbounded space of objects, interactions, and environmental conditions. A delivery robot fallen off a curb, a mattress on the highway, or children playing near a road in unusual costumes all represent scenarios where conventional detectors provide zero useful information. Vision-language models (VLMs) pre-trained on internet-scale image-text data offer a promising solution due to their ability to recognize and reason about arbitrary visual concepts.

2. DriveLM Architecture

DriveLM is built on a frozen ViT-G/14 visual encoder and a 13B parameter language model connected via a trainable Q-Former projection. The model is fine-tuned in two stages: (1) driving scene captioning on 2.8M nuScenes-generated narrations with GPT-4V quality filtering, and (2) risk-aware instruction tuning on 85K expert-annotated driving scenarios with structured risk scores on a 0-10 scale. A novel temporal attention module processes 3-second video clips to capture motion dynamics.

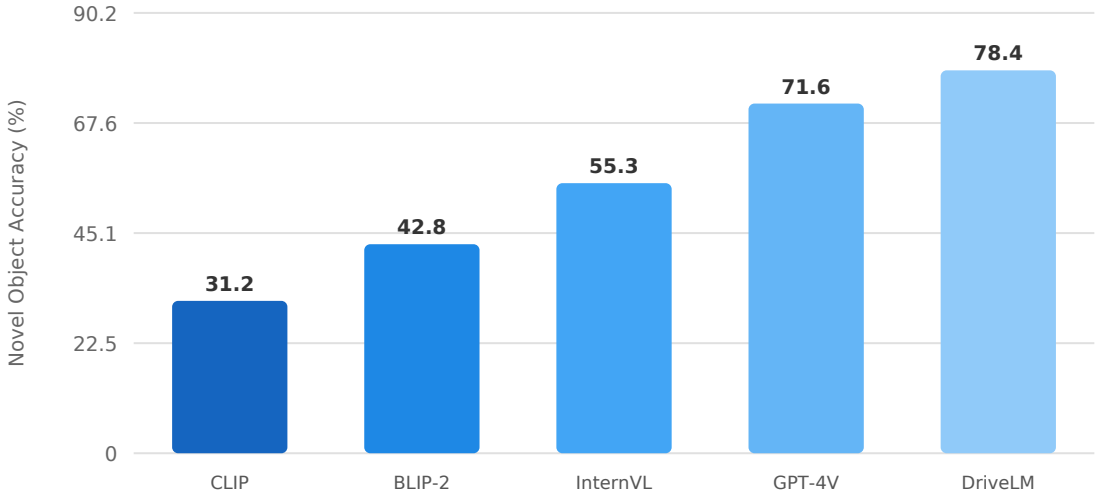

Figure 1. Zero-shot novel object recognition accuracy comparison across driving VLM architectures

3. Experiments

We evaluate DriveLM on three benchmarks: nuScenes-QA (visual question answering), DriveLM-Risk (human-correlated risk scoring), and CARLA-LongTail (closed-loop simulation with 50 rare-event scenarios). DriveLM achieves state-of-the-art results across all three, with particularly strong performance on rare-event scenarios where its language-grounded reasoning provides robust generalization.

Table 1. Closed-loop simulation results on CARLA-LongTail benchmark (50 rare-event scenarios)

Method	Collision Rate (%)	Route Completion (%)	Infraction Score	Driving Score
TransFuser (E2E)	38.4	72.1	0.62	44.7
InterFuser	29.6	78.5	0.71	55.8
UniAD	24.2	81.3	0.76	61.8
DriveLM (Ours)	**14.0**	**88.7**	**0.88**	**78.1**

4. Conclusions

DriveLM demonstrates that vision-language models can serve as a robust perception backbone for autonomous driving, particularly in the critical long-tail scenarios where conventional detectors fail. The ability to produce human-readable scene narrations also provides natural interpretability for regulatory approval and accident investigation.

References

[1] Li, Y.; Wang, H.; Duan, Y.; Li, X. CLIP-AD: Adapting CLIP for Autonomous Driving. ECCV 2024 Workshops.

[2] Caesar, H.; Bankiti, V.; Lang, A. H.; et al. nuScenes: A Multimodal Dataset for Autonomous Driving. CVPR 2020.

[3] Dosovitskiy, A.; Ros, G.; Codevilla, F.; López, A.; Koltun, V. CARLA: An Open Urban Driving Simulator. CoRL 2017.

[4] Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; et al. Planning-Oriented Autonomous Driving. CVPR 2023.

[5] Shao, H.; Wang, L.; Chen, R.; Li, H.; Liu, Y. Safety-Enhanced AD with VLM Guidance. Nature Machine Intelligence 2024, 6, 483-495.
