Benchmarking Vision Foundation Models for Input Monitoring in Autonomous Driving

Mert Keser[1,2], Halil Ibrahim Orhan[2], Niki Amini-Naieni[3], Gesina Schwalbe[4], Alois Knoll[2], Matthias Rottmann[5]

[1]AUMOVIO SE, Germany*
[2] Technical University of Munich, Germany
[3]University of Oxford, United Kingdom
[4] University of Lübeck, Germany
[5] Osnabrück University, Germany

BMVC 2025

Motivation & Goal

Why Input Monitoring Matters?

- **Camera-based deep neural networks (DNNs)** drive critical autonomous driving (AD) functions: lane-keeping, sign recognition, and scene understanding.
- **DNNs in autonomous driving (AD)** lack robustness again **distribution shifts** and they cannot guarantee performance on **novel objects (semantic-shift)** or **conditions (covariate-shift).**
- **Safety-critical systems** require reliable safety monitoring.

Key Question: When can we trust perception, and when should we intervene?

Existing Solutions Fall Short:

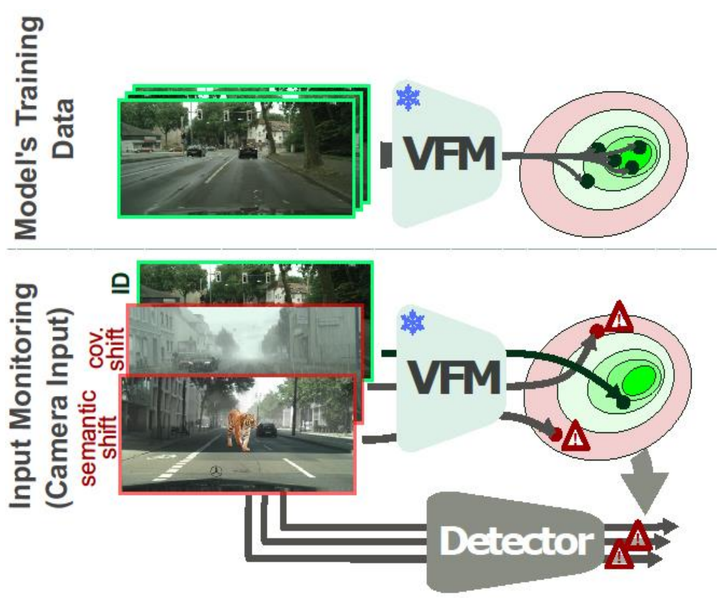
- Detect either semantic or covariate shifts
- Need model internals (impractical in AD)

Our Goal: **Unified**, **model-agnostic**, **scene-level OOD (out-of-training data) detection**.

Proposed Approach

Vision Foundation Models-Based Input Monitoring?

Step 1: Pre-trained VFMs encode training images into a fature space that captures the trainingdata characteristics.
Step 2: Density modeling techniques computes likelihood scores for new inputs, enabling automated detection of ID (in-distribution) or OOD inputs.



Experimental Setup

Vision Foundation Models (VFMs)

CLIP (8 Variants): ResNet + ViT backbones
DINO (5 Variants): Self-Supervised ViT + ResNet
DINOv2 (4 variants): ViT-S to Vit-G
Grounding DINO (2 variants): Swin Transformer

Datasets

Model's Train Data: Cityscapes (train), ACDC (train)
ID: Cityscapes (validation), ACDC (validation)
OOD Semantic Shift: Lost & Found, Bravo-Synobj
OOD Covariate Shift: Foggy Cityscapes, Bravo (Synrain/Synflare), ACDC (fog/night/rain/snow)
Semantic and Covariate Shift: Indian Driving Dataset (IDD), SegmentMeIfYouCan

Evaluation

Metrics: AUROC ↑, AUPR ↑, FPR95↓
Density Methods: APS, MFS, OC-SVM, GMM, Normalizing Flows (NF)
Baseline Methods: GradNorm, Energy, Entropy, MaxLogit, DICE, KNN, ODIN

Experiments

Key Results Summary: Top VFM configurations on semantic shift, covariate shift, and comparison with baseline OOD methods.

Semantic Shift

Detecting Novel Objects

Lost & Found (Real Hazards)

Model	Method	FPR95↓
G-DINO SwinB	**NF**	**0.16**
CLIP RN50x64	GMM	0.23
DINOv2 ViT-G	NF	0.46

Bravo-Synobj (Synthetic)

Model	Method	FPR95↓
G-DINO SwinB	**NF**	**0.02**
G-DINO SwinT	NF	0.20
CLIP RN50x64	GMM	0.53

Key Insights:
- NF most effective for semantic shifts
- Near-perfect on synthetic (0.02)
- Swin architecture excels

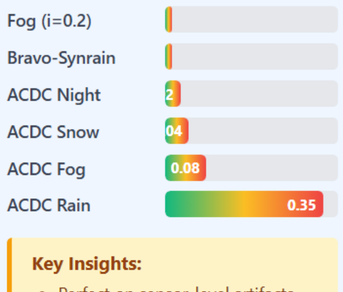
Covariate Shift

Detecting Environmental Changes

G-DINO SwinB + NF Performance

Condition	FPR95↓
Fog (i=0.2)	0.00
Fog (i=0.05)	0.36
Bravo-Synrain	0.00
Bravo-Synflare	0.05
ACDC Fog	0.08
ACDC Night	0.02
ACDC Snow	0.04
ACDC Rain	0.35

FPR95 by Condition



Key Insights:
- Perfect on sensor-level artifacts (fog, Bravo-Synrain)
- Excellent on ACDC night & snow (0.02-0.04)
- Severity impacts detection quality

vs. Baselines

VFMs vs. OOD Detection Methods

Lost & Found

Method	FPR95↓
VFM: SwinB-NF	**0.16**
VFM: RN50-GMM	0.36
GradNorm	0.43
KNN	0.78
Entropy	0.88
DICE	0.94
Energy	0.97

Foggy Cityscapes (i=0.2)

Method	FPR95↓
VFM: SwinB-NF	**0.00**
GradNorm	**0.06**
VFM: RN50-GMM	0.47
KNN	0.88
Energy	0.99

Key Insights:
- VFMs superior on semantic shift
- Perfect detection on covariate shift
- Consistent across both types

ℹ️ **Complete results in the paper:** Full evaluation across all VFMs (19 variants), density methods (5 types), baseline methods (7 approaches), both shift types, and ImageNet-trained counterparts of VFMs for fair comparison please refer to the paper.

Downstream Application

Improving Semantic Segmentation

Research Question: Can VFM-based OOD detection improve downstream task performance without accessing the task model?
Method: We use VFM likelihood scores to filter out high-risk OOD inputs before feeding them to pre-trained semantic segmentation models. We evaluate on IDD as it exhibits both semantic shift and covariate shift compared to Cityscapes. Baseline networks: DeepLabV3+ with ResNet50, ResNet101, and MobileNetV3 and trained on Cityscapes.

Filter Level	Keep	ResNet50	ResNet101	MobileNetV3
None (Baseline)	100%	43.58	45.85	34.22
Low	75%	44.73 +1.15	46.81 +0.96	35.40 +1.18
Medium	50%	46.86 +3.28	48.35 +2.50	36.59 +2.37
High	**25%**	**49.38** +5.80	**50.52** +4.67	**38.36** +4.14

Key Insights

- Up to +5.8 mIoU improvement without accessing task model
- Works across all architectures
- No labels from target domain (IDD) required
- Identifies high-risk inputs for fallback strategies

Conclusions & Future Work

Conclusions

We propose an **unsupervised, model-agnostic OOD detection framework** leveraging **VFM density modeling** to detect semantic and covariate shifts in autonomous driving. Our method **outperforms classical baselines** and enables **practical deployment** through model-agnostic input filtering that improves downstream tasks without accessing the task model.

Future Work

OOD mitigation strategies: Extend beyond detection to develop treatment methods (safe states, expert model queries, handover protocols)
Text-encoder integration: Leverage cross-modal VFMs (CLIP, Grounding-DINO) for human-interpretable descriptions of distribution shifts

Acknowledgments

M.K. and H.I.O. acknowledge that the research leading to these results is funded by the German Federal Ministry for Economic Affairs and Energy within the project "NXT GEN AI METHODS – Generative Methoden für Perzeption, Prädiktion und Planung". The authors would like to thank the consortium for the successful cooperation.
Furthermore, M.R. acknowledges funding from the "REFRAME" project (grant no. 16IS24073C), funded by the German Federal Ministry of Research, Technology and Space (BMFTR). G. S. acknowledges support through the project "chAI" funded by the BMFTR (grant no. 16IS24058). The authors are solely responsible for the content of this publication.

***This paper was submitted while AUMOVIO SE was still a part of Continental AG.**

Resources & Contact

Paper

Contact