Transformer-Based Driver Behavior Recognition Using the UAH-DriveSet Dataset

Aristotelis Tsoutsanis[1,2] **Apostolos Ziakopoulos**[1] **George Yannis**[1]

[1]NTUA Dept. Transportation Planning & Engineering [2]OSeven Telematics

IRTAD Conference 2026
Better Road Safety Data for Better Safety Performance
www.nrso.ntua.gr/irtad2026
15-17 April 2026 Athens, Greece

1 INTRODUCTION

- Driver behavior analysis is critical for **road safety**, **insurance pricing**, and **fleet risk management**. Traditional approaches rely on dedicated hardware, but modern smartphones provide scalable telematics through built-in sensors.
- Smartphones offer scalable telematics via **accelerometers**, **gyroscopes**, **GPS**, and **cameras** — eliminating the need for dedicated on-board diagnostic devices.
- We propose a **multimodal transformer with cross-attention fusion** to classify driving behavior across motorway and secondary road contexts, combining inertial sensor data with visual information.
- Our key hypothesis: *visual context disambiguates identical sensor patterns* (e.g., braking on a motorway vs. a residential street).

2 DATASET — UAH-DriveSet

- **6 drivers** performing **3 behavior types** (normal, drowsy, aggressive) across **2 road types** (motorway, secondary). Total recording time exceeds **500 minutes** of naturalistic driving data.
- **Sensor pipeline:** All signals upsampled to a uniform 10 Hz rate, z-score normalized per channel, then segmented into 30-second sliding windows with 50% overlap.
- **Video pipeline:** 3 uniformly-sampled RGB frames per window, resized to 224×224 px, with standard ImageNet normalization applied.
- **Data split:** Leave-one-driver-out cross-validation to ensure generalization to unseen drivers and prevent data leakage across subjects.

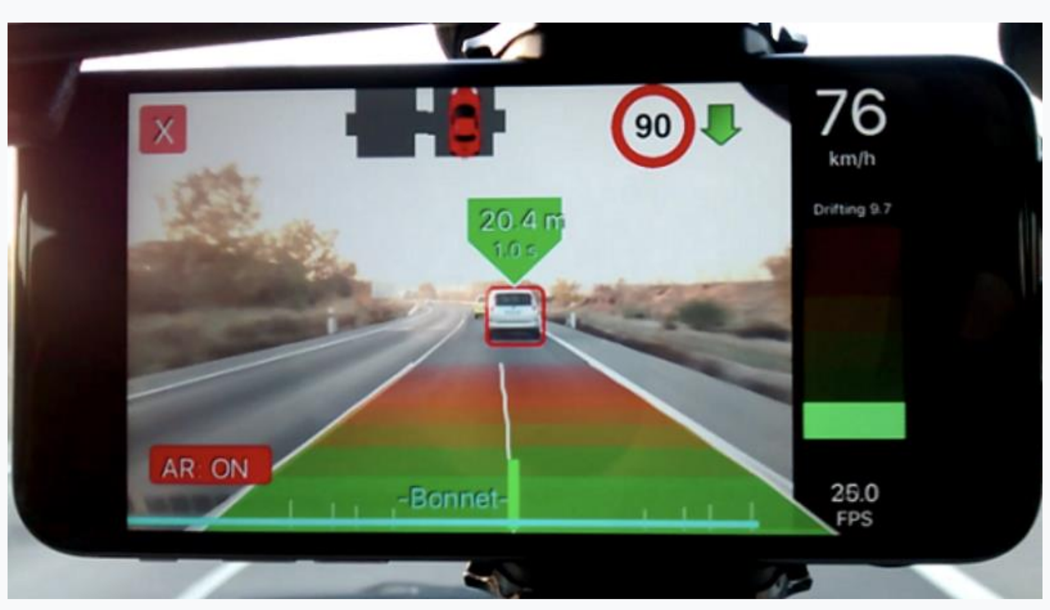

Fig. 1 — UAH DriveSet smartphone setup (Romera et al., 2016)

3 CONFIGURATIONS

Config 1: Sensor Only
Accelerometer + gyroscope signals processed through transformer encoder. Baseline for inertial-only classification.

Config 2: Sensor + GPS Speed
Adds GPS velocity dynamics to sensor stream. Tests whether contextual speed information improves behavior classification.

Config 3: Sensor + GPS + Vision
Full multimodal transformer with cross-attention fusion of inertial data and dashcam video. Also predicts road type.

4 TRAINING SETUP

- **Optimizer** — AdamW, lr = 1e−4, wd = 0.01
- **Schedule** — Cosine LR, 50 epochs, patience = 10
- **Loss** — Cross-entropy, batch 32/16
- **Grad Clip** — Max norm 1.0
- **Augmentation** — Random horizontal flip, color jitter on video frames; Gaussian noise on sensor channels
- **Validation** — Leave-one-driver-out (LODO) cross-validation, reporting macro-averaged F1 score

5 MODEL ARCHITECTURE

Sensor Data
Accel · Gyro · GPS

Video Frames
3 frames · 224×224 RGB

Preprocessing: Upsample → Normalize → 30s Rolling Window

Sensor Encoder
4× Transformer blocks
8-head attention
FFN-512 · 128-d output

Vision Encoder (ViT)
ViT-Small · ImageNet pretrained
384 → 128 dim projection

⊗ Cross-Attention Fusion
Motion + Visual context integration

MLP Classifier — 4 Classes

Normal Motorway	Normal Secondary	Aggressive Motorway	Aggressive Secondary

Fig. 2 — End-to-end multimodal fusion pipeline for driving behavior classification

The proposed architecture is an **end-to-end multimodal transformer** that jointly processes smartphone sensor streams and dashcam video to classify driving behavior. It comprises four stages:

- **Dual-stream encoding** — Inertial data (10 Hz) passes through a 4-layer transformer encoder with 8-head attention and FFN-512, yielding a 128-d motion embedding. Three dashcam frames are encoded by a **ViT-Small** (ImageNet-pretrained), projected 384→128d.

- **Cross-attention fusion** — The sensor embedding queries the visual embedding, enabling selective attention to road-scene features that disambiguate identical inertial patterns.

- **MLP classifier** — The fused representation predicts **four classes** (Normal/Aggressive × Motorway/Secondary) in a single forward pass.

This modular design allows sensor-only and sensor+GPS ablations to reuse the same backbone while bypassing the fusion layer.

KEY CONTRIBUTIONS

- **Cross-attention multimodal transformers** to naturalistic driving behavior classification
- Systematic ablation across **three modality configurations** quantifying each modality's contribution to classification accuracy
- Demonstrated that **visual context is the strongest disambiguator** when identical inertial patterns occur on different road types
- Joint prediction of **behavior class and road type** from a single forward pass, enabling efficient real-time deployment on smartphones
- Robust generalization validated through **leave-one-driver-out cross-validation**, ensuring the model generalizes to previously unseen drivers

6 RESULTS

64.4% — **Config 1: Sensor Only**
Accel + Gyro signals · Transformer encoder
F1 = 0.634

76.0% — **Config 2: Sensor + GPS Speed**
Adds velocity dynamics · +11.6pp gain over baseline
F1 = 0.747

★ BEST MODEL
92.3% — **Config 3: Sensor + GPS + Vision**
Cross-attention fuses motion + visual context · +27.9pp vs sensor-only
F1 = 0.980 (+34.6pp vs Config 1)
Also predicts **road type** (motorway vs. secondary)

KEY INSIGHT
The same sensor reading can be **normal** on a motorway or **aggressive** on a secondary road — visual context resolves this ambiguity.

7 CONCLUSIONS

- ✓ Multimodal fusion enables **context-aware safety decisions** and simultaneous road-type prediction, achieving 92.3% accuracy.
- ✓ Visual information provides the largest single improvement (+23.3pp when added to Config 2), confirming that **road scene understanding** is essential for accurate behavior classification.
- ✓ Directly applicable to real-time **ADAS** and **fleet monitoring** platforms using commodity smartphone hardware.

8 FUTURE WORK

→ Self-Supervised Pretraining: Leverage the vast amounts of available un-labeled driving data through self-supervised learning to further improve feature representation.

→ Larger Multi-Country Datasets: Move beyond the 6-driver population of the UAH-DriveSet to evaluate the model on more diverse datasets involving different driving cultures, vehicle types, and environmental conditions.

→ Temporal Tuning: Conduct experiments with varying window sizes and overlapping strategies to identify the optimal balance between classification latency and behavioral pattern recognition.

ACKNOLEDGEMENTS

This research has been conducted within the IVORY project. The project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101119590.

IVORY

Tsoutsanis · Ziakopoulos · Yannis

driver behavior · telematics · transformer · multimodal fusion · UAH-DriveSet

Contact information:
Aristotelis Tsoutsanis, MSCA Researcher
Department of Transportation Planning and Engineering, NTUA
OSeven Telematics
Email: atsoutsanis@oseven.io , a_tsoutsanis@mail.ntua.gr