NARRATE: A Multimodal Real-World Australian Driving Dataset for Human-Centred Explanations in Automated Driving

Ashkan Yousefi Zadeh[1,2] Zishuo Zhu[1,2] Xiaomeng Li[1,2] Andry Rakotonirainy[1,2] Sebastien Glaser[1,2] Ronald Schroeter[1,2] Patricia Delhomme[3] Zahra Mehraban[1,2]

[1] Queensland University of Technology (QUT), School of Psychology and Counselling, Australia · [2] ARC Training Centre for Automated Vehicles in Rural and Remote Regions (AVR3), Australia · [3] Université Gustave Eiffel, Laboratory of Applied Ergonomics and Psychology (LaPEA), France

Motivation

While existing driving-explanation datasets mostly describe the driver's actions from an external perspective, NARRATE allows drivers to articulate their own decision-making process.

Dataset	Real-World Data	Driver-Sourced	Situational Awareness
BDD-X, BDD-OIA – crowd, post-hoc	✓	✗	✗
LingoQA, DriveLM, OmniDrive – QA	✓	✗	✗
CoVLA, Alpamayo-R1 – generated	✓	✗	✗
LaMPilot – simulation	✗	✗	✗
NARRATE (ours)	✓	✓	✓

2,050 events · **2,383** explanations · **35** drivers · **8,200** clips

Qualitative

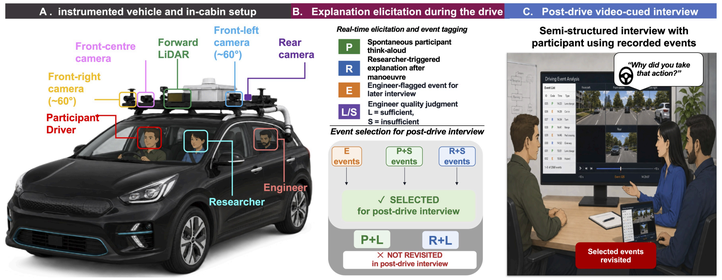
IN-VEHICLE "**I slowed down** because the car in front of me was slowing to turn right."
POST-DRIVE "**I slowed down** because it was turning across traffic.I had to give way, so I would not run into the back of it."

IN-VEHICLE "**I changed lane** because I thought the bus wanted to stop, so I wasn't happy to be there for a while."
POST-DRIVE "**I changed lane** to prepare for the turn ahead – I check the mirror, **decrease the speed, then change.**"

Highlight key – levels of situational awareness (SA)

L1 Perception · L2 Comprehension · L3 Projection
Driver action or reaction – no SA level

NARRATE — collection protocol



1 Tag the event, live
Driver, researcher or engineer flags a decision point as it happens.

804	385	861
think-aloud	researcher-prompted	engineer-flagged

2 Explain at the wheel
Spoken aloud in the moment, while still driving.

3 Explain again on replay
Same event, cued by a 15-second four-view video after the drive.

Both explanations are annotated span by span for L1 Perception, L2 Comprehension and L3 Projection.

Benchmarks

TASK 1 · SITUATIONAL AWARENESS

Can we tell L1/L2/L3 awareness from what the driver said?

0.914 macro-F1, RoBERTa-base

Baseline 0.881
Best model **0.914**

L1–L2 are nearly always present, so the real gain is on L3 Projection: **0.863 vs 0.773**.

TASK 2 · SCENARIO CONTEXT

Can we recover *why* the situation arose?

0.407 macro-F1, 6 groups

Baseline 0.000
Best model **0.407**

Six broad groups are partly recoverable; the 32-class version collapses to **0.035**.

TASK 3 · DRIVER ACTION

Can we predict the manoeuvre the driver took?

0.728 accuracy, RoBERTa-base

Baseline 0.476
Best model **0.728**

Text is the strongest single modality; vision + motion gives the best macro-F1 (**0.306**).

TASK 4 · EXPLANATION GENERATION

Can a model write the driver's explanation?

0.036 BLEU-4, T5-base

Best model **0.036**

Fine-tuned models beat every retrieval baseline, but scores sit near the floor – **still an open problem**.

Driver actions — 10 classes

Slow down	976 (47.6%)
Lane change	451 (22.0%)
Speed up	187 (9.1%)
Stop	150 (7.3%)
4 non-actions*	183 (8.9%)

Scenario contexts — top 5 of 32

Vehicle following	430 (21.0%)
Route-prep lane change	372 (18.1%)
Speed-limit adherence	301 (14.7%)
Signal compliance	232 (11.3%)
Right-of-way	135 (6.6%)

Top four contexts cover 65.1% of events. *Non-action = no stop, lane change, speed up or slow down.

Participant-disjoint splits

	Train	Val	Test	Total
Participants	24	5	6	35
Events	1,402	272	376	2,050
In-vehicle explanations	875	171	227	1,273
Post-drive explanations	750	155	205	1,110
Both timings	223	54	56	333

Words per explanation: 15.6 in-vehicle, 24.3 post-drive.

Route composition — 14.5 km, Brisbane

ROAD TYPE

Major arterial	7.2 km · 50%
Motorway / freeway	3.6 km · 25%
Collector road	1.9 km · 13%
Local street	1.7 km · 12%

Speed zones: 60 km/h 51% · 50 km/h 15% · 40 km/h 12% · under 10 km/h 19%.
~48 pedestrian crossings · ~85 junctions.

Route & language

One route driven by every participant, so events are comparable across drivers.

Think-aloud and prompted events were narrated at the wheel ~97% of the time; engineer-flagged deferrals only 13.2%, but 95.9% post-drive.

ECCV 2026 | MALMÖ | SEPT 8–12
DRIVEX

How do human drivers *explain* their driving actions?

SENSOR SUITE — RECORDED FOR EVERY EVENT

Vision 4 synchronised views · lossless · 10 fps
Ranging 128-channel LiDAR point clouds
Motion GNSS / INS pose, speed, acceleration
Language Driver speech, transcribed and span-annotated

Want to know more?

Scan for the dataset and the benchmark code.

DATASET · DOI
CODE · GITHUB