The LAIA Dataset: Labelled Attention for Intelligent Automobiles

Ainoa Contreras, Diego Porres, Rubén Abad, Pau Cano, Gabriel Villalonga, Antonio M. López, Aura Hernández-Sabaté

***Abstract*—The development of autonomous vehicles (AVs) usually relies heavily on data-driven Artificial Intelligence (AI) models that require large volumes of sensor data with ground-truth annotations. While modular architectures are widely used, end-to-end driving paradigms offer a promising alternative by directly mapping sensor inputs to control actions. However, their adoption is limited due to challenges in interpretability and explainability. To address this, we present LAIA (Labelled Attention for Intelligent Automobiles), a novel synthetic dataset designed to enrich end-to-end driving research with human attention data. Collected using the CARLA simulator in closed-loop environments, LAIA comprises over 15 hours of driving from 44 participants across carefully crafted scenarios designed to evoke natural responses. Each sequence includes RGB images in six weather conditions, semantic and instance segmentation, depth, optical flow, CAN bus signals, and synchronized eye-tracking data. LAIA opens the doors to the use of human gaze driving for different applications such as training attention-aware end-to-end AI drivers, predict driver behavior, or develop methods to detect anomaly driver attention patterns, as well as to improve the explainability of the models.**

***Index Terms*—Autonomous vehicles, dataset, end-to-end driving, eye-tracking**

I. Introduction

DEPLOYING autonomous vehicles (AVs) as soon as they drive just better than humans (not when they are nearly perfect), could save hundreds of thousands of lives over 30 years [1] and reduce congestion, pollution, and time for delivering goods and moving persons. Artificial Intelligence (AI) drivers, which are central to autonomous vehicle development, usually rely on data-driven deep learning models [2] which require vast amounts of labeled sensor data to learn perception and planning tasks through modular pipelines.

As an alternative to the modular paradigm, end-to-end driving approaches [3] do not require manually annotated ground truth. In this framework, the AI driver is implemented as a holistic, deep learning-based sensorimotor model that directly processes raw inputs (typically images) and generates either control outputs such as steering angles, and acceleration/braking forces or a set of waypoints, which can be translated into control commands using conventional controllers. The training of such models relies on synchronously collected sensor data, and vehicle state variables (e.g., speed, acceleration, braking, steering angle), where the raw sensor data act as the input to the model, and the vehicle control signals serve as supervision. These datasets can be collected in a naturalistic and unobtrusive manner from fleets of human-driven vehicles. As a result, training end-to-end AI drivers under this paradigm becomes a process of behavior cloning, whereby the model learns to replicate human driving behavior from observed sensor-action pairs.

Despite the advantages of the end-to-end driving paradigm, its adoption is scarce. The main reason is the difficulty of always understanding the maneuver decisions taken by the DL sensorimotor model acting as an end-to-end AI driver [4]. Therefore, efforts towards having a better understanding of how to train and test end-to-end AI drivers contribute to democratizing the development of AVs.

So far, the development of end-to-end AI drivers has relied on raw sensor data paired with vehicle state variables, all collected during human-operated driving sessions. In this framework, in the training of the corresponding DL models by behavior cloning, human drivers act as latent agents, so that their maneuvers are indirectly captured in terms of vehicle state variables and their decisions are inferred solely through the resulting control signals, such as steering, acceleration, and braking.

However, the visual content that human drivers attend to while executing these maneuvers is entirely excluded from the training process. This omission raises concerns about whether end-to-end AI models genuinely learn the causal relationships that underlie human decision-making or merely replicate observable behavior patterns.

Incorporating human attention information into the training process might be essential for enhancing both the performance and interpretability of AI drivers. Specifically, aligning AI model attention with human gaze behavior could lead to more cognitively grounded and transparent driving models. Furthermore, leveraging human attention as a form of cognitive supervision may improve the model's ability to prioritize relevant scene elements, resulting in more robust and human-like driving behavior.

To explore the integration of human cognitive signals into end-to-end driving models, we present a novel dataset specifically designed to bridge the gap between human visual attention and autonomous driving behavior. We introduce **LAIA** (Labelled Attention for Intelligent Automobiles), the first public collection that couples eye gaze, control actions and full simulator annotations for large-scale autonomous-driving research, enabling reproducible large-scale studies of vision-guided driving. In particular, LAIA contains comprehensive driving scene data that reflect the inputs used by end-to-end AI driving models, including on-board sensor streams and vehicle control signals, all synchronized with human driver attention

All the authors are with the Computer Vision Center.

A.M. López and A. Hernández-Sabaté are also with the Computer Science Department of Universitat Autònoma de Barcelona.



Fig. 1. Platform Setup. *Left*: Dynamic driving platform. *Right*: Static driving platform. In both cases the simulation software is CARLA and the simulations used in human driving were based on the scenarios in Figure 2

data captured via eye-tracking technology. The data were collected in closed-loop driving simulation environments, where 44 participants drove through a series of carefully designed routes featuring sudden and unexpected driving events. The simulation has been conducted using the CARLA autonomous driving simulator [5], complemented by a monitoring system that tracks human behavior while navigating within CARLA's virtual towns. A total of 8 different CARLA routes with 18 different events carefully distributed to cause human reactions during their driving were implemented. RGB images in six meteorological variations, along with ground-truth annotations for semantic segmentation, scene depth, panoptic instance segmentation, and synchronized human eye-tracking data are publicly released.

Unlike existing driving datasets, LAIA uniquely integrates dynamic human attention signals into the driving loop, providing a foundation for more cognitively grounded training and evaluation of AI drivers. The dataset allows researchers to assess not only if an AI driver makes the correct decisions, but also if it pays attention to the most relevant parts of the scene—an important step toward building interpretable and reliable autonomous systems. Additionally, the attention maps extracted from human gaze data can serve as weak supervision or as an auxiliary task in multi-objective learning, supporting training strategies that improve both accuracy and robustness. Beyond learning, LAIA facilitates systematic benchmarking of attention-aware models and allows direct comparisons between human and model behavior under controlled, repeatable conditions.

The remaining of the paper is as follows. Section II describes all the elements necessary for creating the dataset, including the data recording protocol. Section III is devoted to explain the synchronization and mapping of the gaze data with the images generated by the simulator. Section IV describes the format and structure of the data while, Section V shows different examples of the use of the dataset. Finally, conclusions are summarized in Section V.

II. Data Acquisition Description

Data was collected in two different centres, the Institute of Biomechanics of Valencia[1] (IBV) and the Computer Vision Center[2] (CVC), but the only differences among the sub-datasets rely on the physical simulator environments.

During the human driving sessions, eye-tracker data, CARLA images and additional data were captured to replay the episodes offline with the aim of generating additional data without preventing real-time performance during human driving sessions. Adapted autopilots accessing CARLA insider information were also used to automatically generate data associated with driving episodes. In all cases, erratic data were removed.

In this Section, the main characteristics of the simulation environments, experimental scenarios (design, implementation, and experiment structure), and participants used to collect the database are described.

A. Physical Simulation Environments

The setup of the simulator environments consists of a system emulating a car, containing Hardware and Software Interfaces.

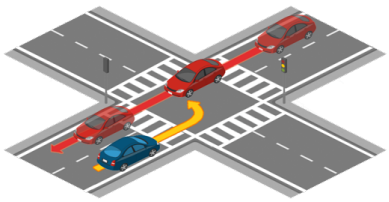
The simulator at IBV is mounted on a dynamic platform (Motion Systems PS-6TM-550) with a regulable seat that simulates the kinematic experience of a real vehicle (acceleration, braking, cornering, crashes and terrain irregularities) in response to manoeuvres executed by the human driver (Figure 1, left). The driving scene is displayed across three 55" 4K screens arranged at 46-degree angles between them.

The hardware simulated environment mounted at CVC is a static platform composed of a foldable aluminium frame, a Z-1 sports seat, the Thrustmaster T300 RS haptic simulation hardware including a GP-4 carbon offset steering wheel together with compatible turn signals and pedals (Figure 1, right). In this case, three articulated monitors of 28" and a resolution of 1920x1080, arranged at 46-degree angles between them, represent the front view of the car plus the two mirrors. In both simulators, a tablet displays the speed.

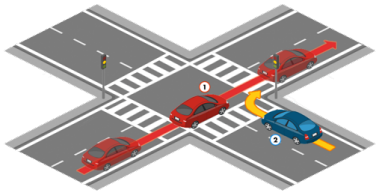
CARLA simulator [5], version 0.9.14 was used to design and run the scenarios containing sudden events. Experiments were run in a PC (ProArt X670E-creator wifi, Ryzen-9 7900X, 2 x 32GB PC5-5600U) with Windows 10 and NVIDIA GeForce 4090 Liquid plus NVIDIA GeForce 4090 INNO3D.

[1] www.ibv.org

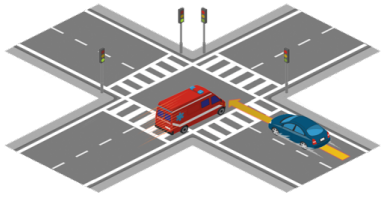
[2] www.cvc.uab.cat



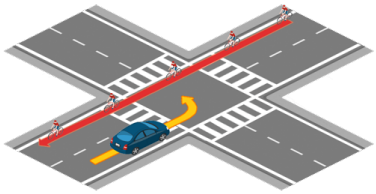
E2. Unprotected left turn at intersection



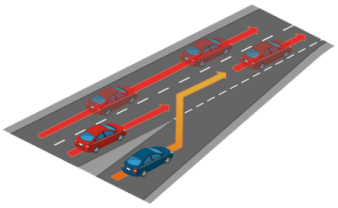
E3. Turn at intersection with crossing traffic



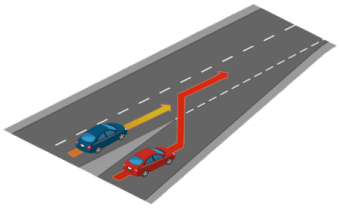
E5. Crossing traffic running a red light at an intersection



E6. Crossing with oncoming bicycles



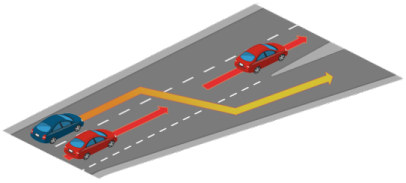
E7. Highway merge on ramp



E8. Highway cut-in from on-ramp



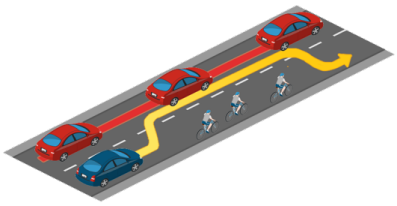
E9. Static cut-in



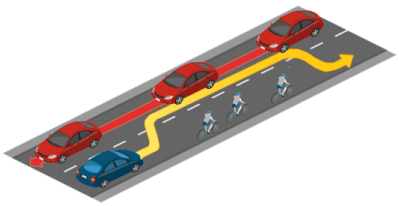
E10. Highway exit



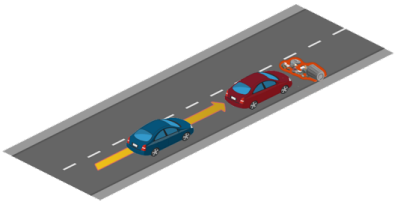
E12. Obstacle in lane



E14. Slow moving hazard at lane edge



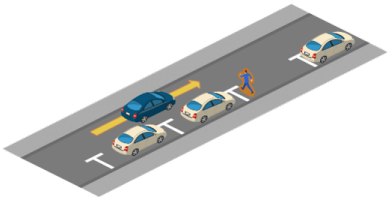
E14a. Slow moving hazard at lane edge



E16. Longitudinal control after leading vehicle's brake



E17. Obstacle avoidance without prior action



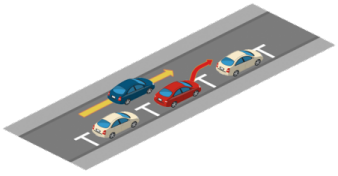
E18. Pedestrian emerging from behind parked vehicle



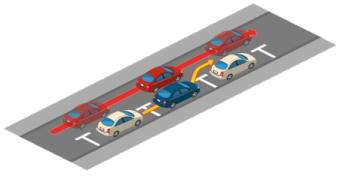
E19. Obstacle avoidance with prior action - pedestrian or bicycle



E19a. Obstacle avoidance with prior action - vehicle



E20. Parking cut-in



E21. Parking exit

Fig. 2. Implemented events

B. *Scenarios*

We define an **episode** as the human driving experience from the time the subject sits on the platform until the end of driving in a predefined scenario.

We define a **scenario**, named by TownXX_RYY, as the combination of a town and a route comprising several events properly triggered as each subject drives along the pre-defined route. The instructions to follow the route were given by an

automated voice (options: English/Spanish/Catalan).

To design the scenarios, four different CARLA maps, namely towns, were selected and, for each map, between one and three routes were designed. Different events that may occur both in a city and on a highway were taken into account in a total of four towns with Single Lane (SL), Multilane (ML), or Highway (HW) were taken into account. In particular, the events implemented are the following and can be grouped in:

- Traffic Negotiation: Unprotected left turn at intersection, Turn at intersection with crossing traffic, Crossing traffic running a red light at an intersection and Crossing with oncoming bicycles
- Highway: Highway merge on ramp, Highway cut-in from on-ramp, Static cut-in, Highway exit
- Obstacle Avoidance: Obstacle in lane, Slow moving hazard at lane edge, Slow moving hazard at lane edge
- Braking and Lane Changing: Longitudinal control after leading vehicle's brake, Obstacle avoidance without prior action, Pedestrian emerging from behind parked vehicle, Obstacle avoidance with prior action - pedestrian or bicycle, Obstacle avoidance with prior action - vehicle
- Parking: Parking cut-in, Parking exit

Figure 2 illustrates all the events implemented while Table I summarizes the distribution of the different events across the different scenarios, taking into account that SL is present in Towns 1 and 4, ML is present in towns 4 and 6 and HW can be found in towns 4, 6 and 10.

For a better illustration figures 3 and 4 show the different CARLA maps with the routes the human drivers have to follow, the simulated vehicles trajectories and the points where the events are implemented.

C. Eyetracker

To capture driver attention a wearable eye-tracker, as the one shown in figure 5, was used (Tobii glasses 3 [6]). These glasses are equipped with four eye cameras and a wide-angle scene camera, allowing for precise tracking of eye movements and visual attention. The glasses are lightweight and unobtrusive, ensuring natural behavior from participants during studies. Gaze data was recorded at 100Hz and the scene camera recorded at 25 fps with a resolution of 1920 x 1080.

D. Participants

A total of 44 participants (23 women and 21 men) were asked to drive during 20-30 minutes by 3-4 scenarios. Participants were healthy people without any condition that might have caused an imbalance in the data recorded:

- At least two years of driving license
- Normal or corrected to normal vision (without glasses)
- Very little tendency to suffer motion sickness
- Non professional drivers
- Driving at least once a week

The average age of participants was 36.59 ± 9.31 and the average of years of license was 16.23 ± 8.76. Four participants got sick during their driving, so, they had to stop the experiment. Even though, the recordings previous to getting sick were valid.

The characteristics of all the participants regarding age, years of license and number of scenarios run are detailed in table II-D:

E. Data Recording Protocol

The main steps of the recording protocol are the following:

1) The participant receives the information sheet about the project under which LAIA is developed and the informed consent to be signed.
2) As soon as (s)he agrees sits in the driving simulator and an eye-tracker glasses.
3) (S)he is asked to drive in the simulated environment relying on the driving simulator following acoustic instructions during about half an hour. These instructions are used to guide the driver along the specified route, in the same way as commercial navigation assistants do.
4) During human driving, the following data is collected:
    a) Driver's gaze captured by the eye-tracking glasses
    b) The images projected by CARLA to provide the driving experience
    c) Log of the driving simulation to allow offline playback of the driving episodes. In particular, CARLA data is captured using adapted autopilots that access all the CARLA privileged information needed for expert-level driving. That is, the autopilots know where the road is, lane lines, cars, pedestrians, traffic lights, etc. without the need to interpret images or any other sensory information.

The methodology of the experimentation procedure was approved by the "Comitè d'Ètica en la Recerca (CERec) de la Universitat Autònoma de Barcelona" (research ethics committee of the university, reference number: CEEAH 6936). For all the experiments, a written consent was obtained from each participant. The consent form explains the goal of the experiment and describes what kind of data are collected and the terms of privacy in the use of personal data. Additionally, it emphasizes that the data released to the general public does not contain information that can directly identify the subject and that any data and research results already shared with other investigators or the general public cannot be destroyed, withdrawn, or recalled. Each consent was hand-signed by each subject on the day of the first experiment.

The eye-tracking glasses were manually calibrated for each driver prior to the start of every scenario.Both human driving data and data captured by autopilots were reviewed to eliminate episodes of erratic driving, and in the case of erratic human driving, the experiment was repeated so as not to end up with fewer drivers than expected.

III. Data Preparation and Validation

To synchronize the driver attention data with the visual driving context, the eye-tracking videos captured during driving sessions are mapped to high-fidelity panoramic imagery generated using the CARLA simulator's playback functionality. The attention mapping pipeline proceeds through the following stages, as visualized in Figure 6, from bottom to top. The raw

TABLE I
EVENT DISTRIBUTION ACROSS SCENARIOS (SEE THE EVENT ID IN FIG. 2

	Traffic Negotiation				Highway				Obstacle Avoidance			Braking and Lane Changing					Parking	
	E2	E3	E5	E6	E7	E8	E9	E10	E12	E14	E14a	E16	E17	E18	E19	E19a	E20	E21
Town01_R01	x	x		x							x	x	x	x	x	x		
Town04_R01	x	x	x									x			x			
Town04_R02					x	x		x	x									
Town04_R03			x	x									x					
Town06_R01					x			x				x						
Town06_R02							x		x	x								
Town06_R03							x		x	x								
Town10_R01																	x	x



Fig. 3. Scenarios for CARLA's towns 01, 04 and 10. Blue lines show the route the human drivers have to follow, from the starting point in dark green to the ending red one. Red lines show the automatic vehicles trajectories, from the starting bright green points to the smallest red ones. Events are marked with a blue box.

		#Drivers	Age	License	Scenarios
IBV	Men	11	40 ± 5.99	20 ± 6.82	3
	Women	13	39 ± 9.02	19 ± 9.21	
CVC	Men	10	31 ± 12.2	12.4 ± 10.7	4
	Women	10	32.4 ± 9.5	11.9 ± 8.6	

visual data is recorded using eye-tracking glasses worn by the driver. This view includes all the displays visible to the participant, including the rearview mirrors, which are positioned to match their real-world locations inside the simulator (see the bottom layer). For consistency with the panoramic playback image generated later, the rearview mirrors are ultimately represented as the leftmost and rightmost sections of the resulting panoramic view.

Rendered square reference markers, with a unique color for each screen, are embedded into the borders of the simulator's physical displays. These markers serve as fixed reference points to geometrically align the eyetracker video frames with the composite CARLA-rendered panorama. Thin vertical strips where displays meet are not captured by any camera but are narrow enough to be considered negligible for practical purposes. Moreover, the gaps are filled in during the offline playback.

Using projective geometry and the aforementioned markers, an intermediate panoramic image is constructed offline. In particular, the eyetracker image is first rotated to correct for head tilt using data from the built-in Inertial Measurement Unit (IMU) of the glasses, ensuring alignment with the horizontal axis of the display setup. Color normalization is then applied to compensate for lighting variations between the eyetracker footage and the CARLA rendering. A color-based filtering approach, followed by a geometric analysis to identify the marker configuration corresponding to the simulator display enable the computation of a homographic transformation to map gaze coordinates from the eyetracker view to the CARLA playback image (see the intermediate layer). This image spatially aligns the eyetracker's visual field with the simulated environment but does not include regions outside the physical display bounds. It is used solely as a visual reference to assess where gaze



Fig. 4. Scenarios for town maps 06. Blue lines show the route the human drivers have to follow, from the starting point in dark green to the ending red one. Red lines show the automatic vehicles trajectories, from the starting bright green points to the smallest red ones. Events are marked with a blue box.

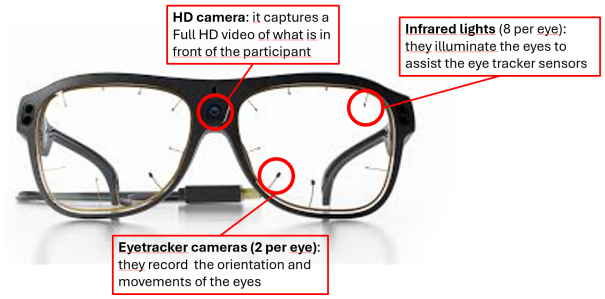


Fig. 5. Eyetracker device

data will be projected.

To project driver attention, gaze data are mapped from the eyetracker frame onto this intermediate panorama. A fixation is defined as a set of gaze points where the maximum angular velocity does not exceed 100 degrees per second and the duration ranges between 0.1 and 0.5 seconds. Fixations lasting longer than 0.5 seconds are divided into shorter fixations. Thus, attention is modeled as a two-dimensional Gaussian distribution centered at the gaze point, with the standard deviation σ accounting for multiple sources of uncertainty:

- The intrinsic fixation error of the eye-tracking device (typically < 5 pixels in the eyetracker image)
- Downsampling from the 100 Hz gaze stream to the 25 fps scene camera
- Temporal misalignment between eyetracker and simulator video playback
- Reduction of human attention to a single gaze coordinate per frame

Some regions remain without information in this intermediate image, as they fall outside the eyetracker's field of view (see the right side of the intermediate layer). These areas are recovered in the next stage.

The complete and accurate panoramic image is reconstructed offline using CARLA's playback system. This composite includes all six RGB views (frontal, lateral, rear, and mirrors), along with semantic overlays and depth-based intensity projections. The CARLA simulator, in playback mode, renders these perspectives using ground-truth data from the original driving session, independent of the driver's head pose or visible field.

Gaze data previously mapped to the intermediate image is now accurately projected onto the full panoramic scene. Attention projections retain their Gaussian spread to encapsulate uncertainty across all synchronization and projection stages.

This rigorous multi-step process ensures that human attention is faithfully mapped onto the original simulator-grounded visual scenes. To qualitative support the proper generation of Gaussian gaze attention, Figure 7 shows the comparison between two Gaussian-based representations of driver gaze, the one provided directly by the eye-tracker and the Gaussian distribution computed in our pipeline. The inner crosses indicate the instantaneous gaze point recorded by the eye-tracking device and the center of our Gaussian showing an almost perfect concentricity.

Figure 8 shows different examples of qualitative validation of the whole mapping.

IV. Data Format and Structure

The dataset described in this paper has been made publicly available on https://cloningdcb.org/. No registration is required

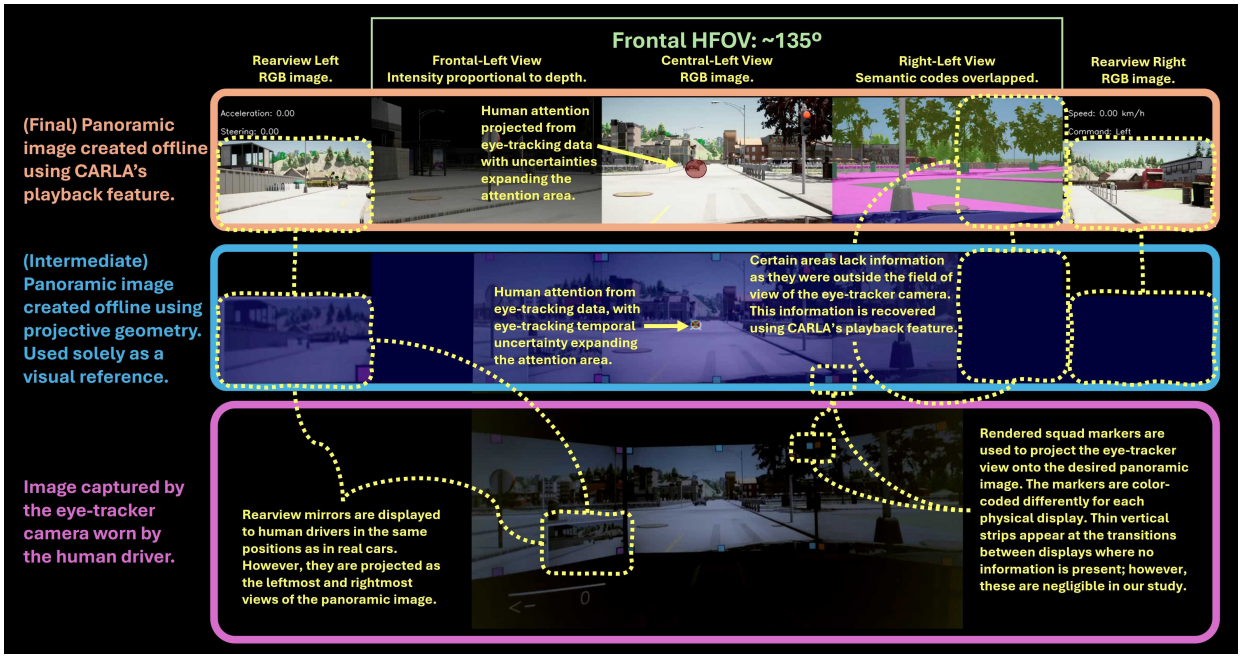


Fig. 6. Attention mapping pipeline. From bottom to top: image captured by the eyetracker camera, intermediate panoramic image using projective geometry, and final panoramic image generated using CARLA's playback feature. Human attention is projected and expanded with uncertainty modeled by a Gaussian distribution.



Fig. 7. Visualization of gaze uncertainty modeling and comparison to the original Gaussian provided by the eye-tracking device.

and users can select and download specific data through a customizable download configurator. This interface will allow users to tailor their dataset selection based on their needs, ensuring flexible and efficient access to the data.

For each driver and each scenario the data available is the following:

1) RGB images obtained from central, right and left cameras, as well as left and right mirrors. These images are generated in 6 different weathers: clear noon, clear sunset, Hard rain noon, wet noon, cloudy noon, mid rain sunset.
2) Pixel-level Semantic Segmentation labels in color, in PNG format.
3) Depth: Pixel distance obtained per camera using a CARLA Depth camera at the same position and resolution as the associated RGB camera.
4) Optical flow data.
5) CANBUS information at every timestamp, at a frequency of 25 Hz. It contains a .json file with the information of speed, acceleration, brake, throttle, steer, ego position, gyroscope, accelerometer, hand_brake, rear_gear, blinker, and direction
6) Attention map generated from eye-tracker mapped on the aligned RGB images.

Internally, data is structured as follows:

```
driver
└─ map
   └─ route
      └─ data_modality
         └─ camera
            └─ weather
               └─ frame
```

Figure 9 shows some examples of images of the data available. On the top an original image with the Gaussian from human attention. On the bottom,

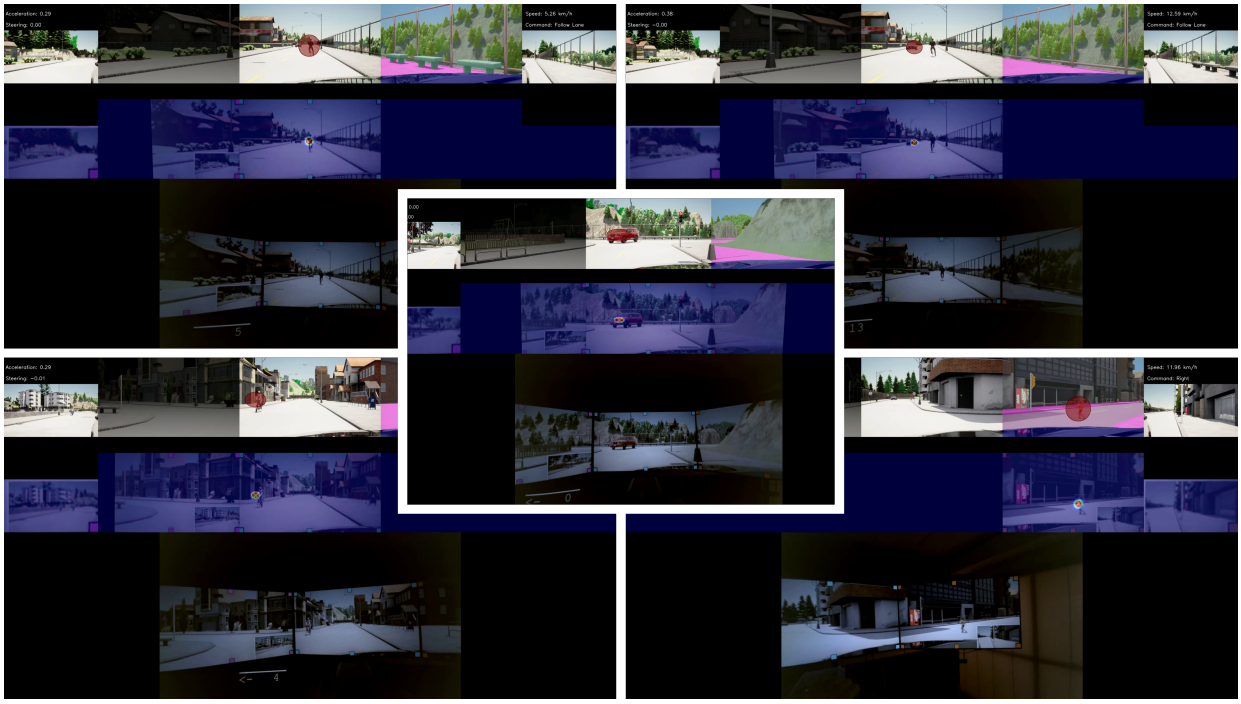


Fig. 8. Qualitative validation of the eye-tracking mapping onto the original simulator-grounded visual scenes. Each grouped image contains the original image captured by the eye-tracking on the bottom, the intermediate panoramic image with a heat map visualizing the projected human attention, together with the instant gaze produced by the eye-tracker device, and the final panoramic image with the gaze data accurately projected.

V. Examples of Application

The LAIA dataset enables a wide range of applications aimed at advancing both the performance and interpretability of AI-based driving systems. In particular, the availability of synchronized gaze data, control signals, and rich scene annotations opens new directions for cognitive modeling and human-centered AI development. Below, we outline several representative use cases. For example, the dataset can be used to train end-to-end driving models guided not only by control signals but also by human attention maps, serving as cognitive supervision, improving model interpretability and robustness. As well, LAIA enables systematic comparison between the attention of AI drivers and that of human drivers, which might support the evaluation of whether models focus on semantically meaningful regions (e.g., pedestrians, vehicles, traffic signs) and foster the development of more explainable AI systems. The design of specific events in the routes enables the study of how gaze patterns correlate with driving maneuvers in the presence of specific events (e.g., a pedestrian emerging, a vehicle cutting in). Thus, driving behavior modelling and prediction could be approached. Finally, by combining human attention with ground-truth annotations, the dataset enables robust benchmarking of perception models across diverse weather conditions, viewpoints, and occlusion levels.

These applications demonstrate the versatility of LAIA as a resource for improving autonomous driving systems.

VI. Conclusion

In this paper, we presented LAIA, a novel dataset designed to support the development of attention-aware end-to-end autonomous driving systems. By combining raw sensory input, control actions, and high-resolution human gaze data in a wide variety of simulated scenarios, LAIA offers a unique resource for studying the interplay between perception, decision-making, and visual attention.

The dataset enables new research directions focused on improving the transparency, interpretability, and cognitive alignment of AI drivers.

References

[1] M. Bauman, "Why waiting for perfect autonomous vehicles may cost lives," *RAND Corporation*, 2017.

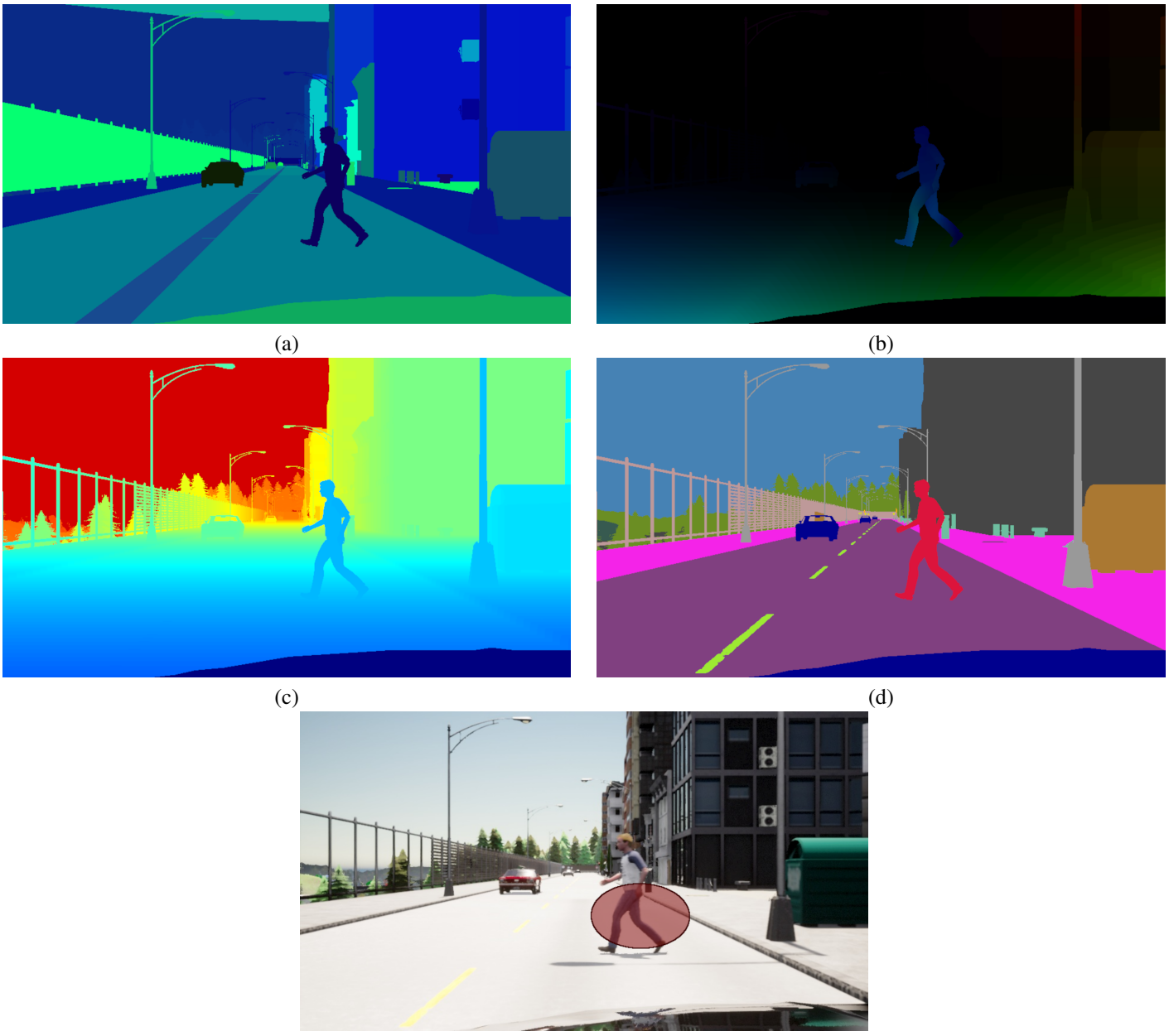
[2] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *IEEE Trans. on ITS*, 2020.

[3] F. Codevilla, M. Müller, A. López, V. Koltun, and A. Dosovitskiy, "End-to-end driving via conditional imitation learning," in *ICRA*, 2018.

[4] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, "End-to-end autonomous driving: Challenges and frontiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[5] A. Dosovitskiy, G. Ros, F. Codevilla, A. López, and V. Koltun, "Carla: An open urban driving simulator," in *CoRL*, 2017.

[6] "Tobii pro glasses 3 — latest in wearable eye tracking," accessed: 2024-11-04. [Online]. Available: https://www.tobii.com/products/eye-trackers/wearables/tobii-pro-glasses-3



(a) (b) (c) (d) (e)

Fig. 9. Illustration of available data. a) Panoptic Instant Segmentation b) Optical flow, c) Depth, d) Semantic Segmentation e) Gaze overlapped on RGB image