TimelyTale: A Multimodal Dataset Approach to Assessing Passengers' Explanation Demands in Highly Automated Vehicles

GWANGBIN KIM, Gwangju Institute of Science and Technology, Republic of Korea
SEOKHYUN HWANG, Gwangju Institute of Science and Technology, Republic of Korea
MINWOO SEONG, Gwangju Institute of Science and Technology, Republic of Korea
DOHYEON YEO, Gwangju Institute of Science and Technology, Republic of Korea
DANIELA RUS, Massachusetts Institute of Technology, United States
SEUNGJUN KIM*, Gwangju Institute of Science and Technology, Republic of Korea

Explanations in automated vehicles enhance passengers' understanding of vehicle decision-making, mitigating negative experiences by increasing their sense of control. These explanations help maintain situation awareness, even when passengers are not actively driving, and calibrate trust to match vehicle capabilities, enabling safe engagement in non-driving related tasks. While design studies emphasize timing as a crucial factor affecting trust, machine learning practices for explanation generation primarily focus on content rather than delivery timing. This discrepancy could lead to mistimed explanations, causing misunderstandings or unnecessary interruptions. This gap is partly due to a lack of datasets capturing passengers' real-world demands and experiences with in-vehicle explanations. We introduce TimelyTale, an approach that records passengers' demands for explanations in automated vehicles. The dataset includes environmental, driving-related, and passenger-specific sensor data for context-aware explanations. Our machine learning analysis identifies proprioceptive and physiological data as key features for predicting passengers' explanation demands, suggesting their potential for generating timely, context-aware explanations. The TimelyTale dataset is available at https://doi.org/10.7910/DVN/CQ8UB0.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; **Ubiquitous computing**; **User studies**.

Additional Key Words and Phrases: automated vehicles, explanation, explainability, intelligibility

ACM Reference Format:
Gwangbin Kim, Seokhyun Hwang, Minwoo Seong, Dohyeon Yeo, Daniela Rus, and SeungJun Kim. 2024. TimelyTale: A Multimodal Dataset Approach to Assessing Passengers' Explanation Demands in Highly Automated Vehicles. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 3, Article 109 (September 2024), 60 pages. https://doi.org/10.1145/3678544

1 Introduction

The integration of automated vehicles into our daily lives is expected to bring benefits for urban mobility, such as increased safety, reduced traffic congestion [31, 84], and enhanced accessibility for various socioeconomic

*Corresponding author.

Authors' Contact Information: Gwangbin Kim, gwangbin@gm.gist.ac.kr, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea; Seokhyun Hwang, anoldhsh@gm.gist.ac.kr, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea; Minwoo Seong, seongminwoo@gm.gist.ac.kr, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea; Dohyeon Yeo, ing.dohyeonyeo@gm.gist.ac.kr, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea; Daniela Rus, rus@csail.mit.edu, Massachusetts Institute of Technology, Cambridge, United States; SeungJun Kim, seungjun@gist.ac.kr, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea.

ACM 2474-9567/2024/9-ART109
https://doi.org/10.1145/3678544

groups, particularly those who are currently disadvantaged or unable to drive [165]. Although the timeline for widespread adoption of fully automated vehicles (SAE Level 5) remains uncertain, with some studies predicting market readiness by 2030 [140], building user trust is a critical factor in the successful deployment and acceptance of automated vehicles, as it affects users' attitudes towards the technology [18]. However, public anxiety and hesitation towards automated vehicles [79], combined with the potential for low trust, which causes worry and decreased acceptance [80], highlight the challenges in fostering user trust. Explanations in automated vehicles can help address these challenges by alleviating negative experiences and providing an increased sense of control [129]. By improving understanding of the vehicle's capabilities [74, 128] and enhancing situation awareness [64], explanations can improve passenger trust [116] and acceptance [64].

To support trust and situation awareness as intended, providing explanations in automated vehicles requires the consideration of nuanced contexts. Since poorly designed explanations can adversely affect the passenger experience [46], explanations should convey information with sufficient intelligibility under rapidly changing road environments [48]. Thus, previous studies have explored various explanation presentation methods, focusing on enhancing the passenger experience by building trust, comfort, and acceptance while reducing fear, anxiety, and cognitive load [19, 20, 22, 74, 92, 127, 147]. These studies have considered design factors, such as explanation types and timing, as well as modality, visualization methods, and information quantity, across different driving scenarios and non-driving related tasks (NDRTs).

Alongside other design factors, the timing of explanations should be considered for effective information delivery and for enhancing the passenger experience in automated vehicles. For instance, Haspiel et al. [49] found that providing explanations before an automated vehicle takes action enhances user trust. The experiment by Du et al. [32] has underscored that explanations given before an automated vehicle's actions are more effective in building trust than those delivered afterward. Despite their importance, the optimal timing for explanations and the actual passenger demand in real road environments have yet to be widely explored, particularly outside of simulated environments [92]. This work aims to contribute to this area by creating a multimodal dataset that captures real-time demands for in-vehicle explanations during automated driving, aiming to align explanation timing with passenger needs better.

Current machine learning approaches to explanation generation, which are often designed for developers and regulators, may not fully align with passenger requirements. They tend to focus on high-risk scenarios, such as accidents [104] or offer explanations for every scenario [67], which may not always be necessary and could be disruptive. Few studies [77, 113, 114] have considered human perspectives. A primary obstacle to developing passenger-centric explainable artificial intelligence (XAI) black models is the lack of datasets that account for passenger contexts [33]. Due to the limited norm of data sharing in automotive studies, most naturalistic driving studies do not share their data [33]. Dataset for passengers' actions, states, and demands for explanations are keys to aligning XAI models with end-user services that provide both proactive and adaptive explanations. Our research aims to address situations in which explanations can compete for passengers' attention during NDRTs under naturalistic driving scenarios.

In identifying the right moments for explanations, we explored the concept of in-vehicle interruptibility. In manually driven vehicles, AI-initiated interactions aim to support safe driving [56], and such interventions must not distract drivers, necessitating a focus on the driver's availability—a concept referred to as interruptibility [5, 72]. However, in automated vehicles, where passengers are generally not engaged in driving tasks, the emphasis shifts from identifying moments for secondary tasks during manual driving to pinpointing moments for driving-related information during NDRTs [73]. Our study addresses this shift in focus, exploring the concept of passengers' demand for explanations, a transition from the traditional concept of in-vehicle interruptibility, to support passengers' situation awareness during automated driving situations.

Our study seeks to bridge the gaps in datasets for in-vehicle explanations, with a focus on capturing in-situ passenger experiences and collecting real-time data across environmental, driving, and passenger contexts. This

dataset includes real-time demands for explanations from passengers in automated vehicles under actual road conditions. By incorporating the concept of interruptibility, our dataset and analysis is designed to support the development of proactive XAI services that can adapt to passenger needs and provide timely interactions.

1.1 Present Work

In this study, we present a multimodal dataset for explanation demands in automated vehicles, focusing on highly automated vehicles with SAE levels 4-5 and passengers who are not already familiar with these vehicles. Our dataset provides passenger explanation demand timing and content, contributing to the fields of interruptibility and XAI in the following ways:

- We offer explanation demand timing and content in the form of textual driving explanations required by in-situ passengers, synchronized with timestamped exteroceptive, proprioceptive, and interoceptive data for environmental, driving, and passenger contexts. For a detailed description of the composition and formatting of each data channel, refer to Appendix B.
- We analyze explanation demands in relation to geographical data using co-occurrence analysis, GPS-tagged clustering, and textual analysis to understand the most common demand scenarios.
- We develop a preliminary machine learning model that predicts opportune moments for providing explanations, extending the concept of interruptibility to the domain of XAI in automated vehicles.

2 Related Works

2.1 Explanations and Trust Calibration

As Norman discussed, inattentiveness to automation or inappropriate understanding of the automation's status and actions may lead to unawareness and inability to intervene in the automation, even in critical situations [111]. To prevent this, automation systems, particularly automated vehicles, should communicate their state to users in a comprehensible manner, making their processes and status transparent [16]. Explanations in automated vehicles enhance transparency and help passengers understand the vehicle's decision-making process. This increased transparency contributes to passenger trust and improves the overall experience in automated vehicles [64]. When passengers have a clear understanding of the vehicle's capabilities and limitations, they can trust the system to handle the driving tasks without constantly monitoring its functions [59]. Providing information about the system's potential limitations helps prevent temporary declines in trust [76].

However, the primary goal of explanations in automated vehicles should not be merely to increase passenger trust regardless of the vehicle's actual capabilities. Instead, the true purpose of explanations is to calibrate user trust. Explanations help passengers understand the vehicle's genuine capabilities, enabling them to trust the vehicle at an appropriate level, preventing both undertrust and overtrust [142]. Explanations can help users calibrate their trust in the system, leading to a more appropriate level of trust that matches the system's actual capabilities [52].

Overtrust in automated vehicles can lead to complacency, where passengers may fail to remain sufficiently aware of the road environment and react appropriately in cases of system malfunction. Although highly automated vehicles with SAE levels 4 are designed to operate without human intervention, they function within predefined operational design domains (ODDs) [75]. While passengers are not expected to take control of the vehicle and may not need to be constantly aware of the road environment, they should understand the vehicle's operation, the situations it can handle, and its boundaries. Overtrust could cause users to believe that the vehicle can operate under any condition without properly assessing whether the current situation falls within the ODD [142]. In rare cases, Level 4 AVs might encounter situations beyond their capabilities, such as extreme weather events or unexpected road hazards. Overtrust could lead to a false sense of security and lack of preparedness for emergencies and lead to the misuse of the system, where users rely on the automated vehicle in situations

beyond its capabilities [142]. Explanations help passengers improve their situation awareness, understand the vehicle's state and capabilities, and trust the vehicle to the appropriate level [74].

To calibrate user trust without degrading passenger experience, the interaction with these explanations should consider the content and timing. Trust in AI systems is influenced by their ability to provide justifiable explanations [120]. Poorly designed explanations in AI can lead to trust calibration errors, such as irrational agreement or disagreement with the AI's decisions [105]. Therefore, the content of explanations should be tailored to the specific context and user needs. The timing of explanations also influences trust levels. Explanations provided before an action tend to foster more trust than those given afterward [49]. Risk-adaptive explanations have been found to be effective in promoting passenger experience and acceptance, especially when the information provided is overwhelming [64]. Explainability contributes to trust in AI by fostering justified and warranted trust, particularly when it enables users to rely on the system without constant monitoring [36].

In summary, well-designed explanations delivered at opportune moments can help passengers calibrate their trust in automated vehicles by providing transparency about the vehicle's current state and actions. Effective explanations should be tailored to the user's needs, provide information about the system's decision-making process, and be delivered at appropriate times to foster trust and acceptance. This study aims to contribute to the development of intelligible explanations that deliver relevant information at appropriate times, enabling passengers to maintain a well-calibrated level of trust in automated vehicles.

2.2 Non-driving Related Tasks and In-vehicle Explanations

In highly automated vehicles with SAE levels 4 or higher, passengers are free to engage in NDRTs such as reading, watching videos, or working on their laptops, as the vehicle handles all driving tasks without human intervention [8, 106]. While passengers are not required to actively monitor the driving environment or take over control, maintaining a certain level of situation awareness is still important for a safe and comfortable experience [19, 21]. This shift in the primary task from driving to NDRTs creates new challenges in understanding how the passenger interacts with the vehicle and the driving environment, as well as how to support their situation awareness during NDRTs in automated vehicles. Engaging in NDRTs can divert passenger attention from the driving situation [95]. For example, a passenger immersed in watching a movie or working on a complex task may not notice changes in the vehicle's speed, direction, or surrounding traffic conditions. Thus, engaging in NDRTs can result in reduced situation awareness [95], potentially leading to discomfort or anxiety when the vehicle encounters unexpected situations or takes actions that the passenger does not understand [145]. To attain situation awareness, people often self-interrupt their NDRT and watch the immediate surrounding environment [43]. As self-interruption halts the ongoing NDRT, passenger self-interruption poses trade-offs between situation awareness and NDRT engagement. Additionally, in some studies, intermittent views of the road did not suffice to maintain situation awareness [124]. In this regard, explanations from the vehicle about its actions and the driving environment can be helpful in promoting situation awareness [74].

While explanations can provide situation awareness in vehicles, they can also interrupt passengers engaged with NDRTs. Thus, automated vehicles should support passengers in deciding whether to interrupt an NDRT and facilitate the re-uptake of the NDRT [107]. In highly or fully automated vehicles, although passengers don't have to participate in the driving decisions, Yusof et al. [160] express concern that interruptions to ensure passenger comfort may shift attention from the NDRT and affect the passenger experience and performance in tasks such as reading. Indeed, an experimental study assessed the average demand that different NDRTs pose and classified tasks into those that require visual and cognitive demand (smartphone, computer, read, internet, texting), visual and psychomotor demand (eat, drink, makeup, dress up, cleaning), and almost no demand (sleep, nothing) [66]. Similarly, the effectiveness of explanations is affected by the NDRT engagement [162] and the modality channels they occupy [15]. For example, visual explanations may not be sufficient to provide information while passengers

are engaged with an NDRT [15, 157]. In contrast, auditory explanations may be more effective when the passenger is visually engaged with an NDRT, as they do not compete for the same attentional channel [108, 162].

In summary, while NDRTs are a key benefit of highly automated vehicles [106], they can compete for the passenger's attention and working memory, potentially reducing situation awareness and affecting the passenger experience. Explanations from the vehicle can support situation awareness, but the timing and modality of these explanations should be carefully chosen to minimize interference with the passenger's ongoing NDRTs. As our demand for explanations and the effect of explanations may change depending on the engagement and the type of NDRTs, this study aims to investigate this relationship. Specifically, this study seeks to collect passenger demand datasets for explanations amid their naturalistic NDRTs, understand the distribution of NDRTs in a naturalistic driving environment, and examine the extent to which passengers demand explanations during such tasks.

2.3 Wizard-of-Oz (WoZ) Method for Automated Vehicles Research

Despite the need for studies assessing passenger behavior in automated vehicles, given their anticipated prevalence in the near future, current regulatory and legislative issues have not been fully addressed, making it challenging to study passenger behavior in vehicles with higher SAE levels. Existing automated systems rely on drivers for monitoring and supervision, which hinders the ability to study driver behavior during their disengagement from driving tasks. To overcome these regulatory and ethical issues associated with testing actual vehicles with SAE levels 4 or higher, researchers often employ the WoZ methodology [24], a technique that creates the illusion of automated driving for participants by concealing an experimenter acting as a hidden driver [93].

Researchers have utilized the WoZ method to study human behavior and interaction in automated vehicles by concealing the wizard' driver behind a partition, creating the illusion of an automated vehicle [9, 144]. This approach has been used to investigate preferred driving styles [68], passenger experiences in automated vehicles [34], robo-taxis' [96], AV-pedestrian interaction [123], AR interfaces [38], and automated driving training [135]. In some cases, partitions have been replaced with video-see-through screens that hide the front driver side [29]. When the study requires the augmentation of automotive interfaces rather than the naturalness of the environment itself, extended reality environments have been incorporated to conceal the wizard driver [135, 155, 156]. These approaches allow researchers to investigate user behavior and interaction in a more realistic setting without compromising safety or violating current regulations.

When investigating user aspects of future automated vehicles, researchers face challenges regarding the naturalness or technical fidelity of their studies on either the car side or the environment side. Pai et al. [117] compared passenger perceptions of automation in two conditions: one with actual algorithmic automation and another employing a WoZ approach. The study found no statistically significant differences in participant perception and behavioral performance between the two conditions, demonstrating the validity of the WoZ method in studying human behavior in automated vehicles [117]. Although driving simulators offer a safer way to simulate driving situations with proven behavioral validity [44], they have limitations owing to their artificial lab environment, which is inherently safe and provides different kinesthetic experiences from actual driving, potentially leading to a different perception of driving [29]. The WoZ method can be considered when the feeling of being on actual roads should be provided [29].

However, the WoZ method can present several methodological challenges. Firstly, the obtained results should be independent of the driving wizard to ensure objectivity. Thus, different driving wizards should be able to reproduce the same driving style consistently [103]. This can be achieved through training and the use of predefined driving scripts or scenarios. Secondly, to ensure reliability, each individual driving wizard must be able to reproduce the same driving behavior across different sessions consistently [9, 103]. Lastly, to maintain validity, the simulation should appear and behave like real automation scenarios and create the illusion of riding

an automated vehicle [103]. In other words, participants must be under the impression that they are interacting with a genuine automated vehicle for the study to yield valid results [13].

In our study, we chose an on-road experimental environment to elicit passenger demands for explanations while interacting visually and kinesthetically with simulated automated vehicle scenarios. We used the WoZ method to maintain safety and adhere to ethical and legislative regulations while providing the illusion of riding an automated vehicle (specific implementation measures in response to the methodological challenges are described in section 3). However, the limitation of the WoZ method that user behavior in actual automated vehicles might still differ from actual automated vehicles in the absence of the experimenter should be acknowledged (e.g., due to the presence of an experimenter).

2.4 Predicting In-vehicle Interruptibility

The growing use of mobile computers for in-vehicle information and entertainment has heightened interest in identifying opportune moments to 'interrupt' drivers with interactions as these systems can distract drivers and challenge their cognitive resources [143]. Indeed, the effectiveness and acceptance of proactive interactions in vehicles are sensitive to environmental and user contexts [161]. In manually driven vehicles, interaction design focuses on identifying moments suitable for AI-initiated interruptions that do not distract drivers from safe driving [5]. However, in automated vehicles, the context changes as passengers are no longer responsible for driving and are more likely to engage in NDRTs. This shift requires a new focus, moving from managing NDRTs during manual driving to providing relevant driving-related information during NDRTs [73].

Passenger interruptibility could differ depending on the task the passenger is performing [54]. Moreover, explanations that adapt to risk levels are preferred to constant explanations when the information is excessive [64]. In this context, AI in vehicles could identify the most appropriate moments to offer explanations about the road environment, enhancing passenger awareness while allowing them to enjoy NDRTs without excessive preoccupation with the vehicle's driving status. Our study draws from in-vehicle interruptibility research and aims to shift the focus from modeling moments of interruptibility during NDRTs to moments of explainability for passengers engaged with NDRTs.

2.4.1 Sensor-based approaches to model in-vehicle interruptibility. As interruptibility aims to model opportune moments to interrupt the driver, research in this field focuses on modeling both the timing and content of proactive vehicle interactions, considering driving situations and driver characteristics [5]. Since driving sensors or passenger sensors can capture events or states in driving and physiology, various sensor-based approaches have been employed to gather data on driving and passenger contexts for modeling these situational and user-related features. For instance, Kim et al. [61] utilized On-Board Diagnostics (OBD), Controller Area Network (CAN), and dashcam data to determine the appropriate timing for proactive audio-verbal tasks in cars, aiding in the creation of a model that predicts the need for such tasks across various driving conditions [62]. Similarly, Semmens et al. [130] developed an on-road driving dataset, incorporating GPS, CAN, and video data, to assess a driver's readiness for proactive interactions with in-vehicle voice assistants. Wu et al. [148] employed Inertial Measurement Unit (IMU), CAN, and interoception sensors modeling passenger behavior and states to determine moments for proactive interactions. These studies demonstrate the efficacy of sensor-based approaches to capture relevant contextual information for modeling in-vehicle interruptibility. Building on this approach, our study utilized cameras, automotive sensors, and physiological sensors to gather data on vehicles' operation and passenger behavior and state to model in-vehicle demand for explanations.

2.4.2 Related Datasets: Dataset for In-vehicle interruptibility. The development of datasets could advance in-vehicle interruptibility research, as the prediction of opportune moments predominantly relies on machine learning algorithms to interpret driving and passenger-related sensor data. These datasets enable a wider exploration of

situational and user-related features, informing the design of proactive vehicle interactions. For example, the public on-road driving dataset for driver interruptibility created by Semmens et al. [130], which incorporates GPS, CAN-bus, and video data, has enabled further research, such as Wu et al. [148] utilizing the dataset to assess a driver's readiness for proactive interactions with in-vehicle voice agents. Despite the potential benefits datasets could bring, only a few datasets for modeling in-vehicle interruptibility are publicly available (e.g., INAGT dataset [130]), which deals with manually driven vehicles. In our study, we built upon these approaches and utilized cameras and physiological and driving-related sensors to gather comprehensive data on passenger behavior and state, with the aim of creating a dataset that can be used to model the demand timing for explanations in automated vehicles.

2.5 Generating Explanations for Automated Vehicles

As automated vehicles are safety-critical applications of machine learning, interpretable models to make driving decisions transparent have been developed. They were designed to support various stakeholders, such as developers, regulators, and accident investigators, in developing, inspecting, and establishing guidelines by providing accountability in decision-making [116]. In this study, we focus on end-user passengers as the main stakeholders and aim to model the actual demand for explanations from the passenger perspective, measuring exteroception, proprioception, and interoception data (see Table 1 for comparison with other datasets for in-vehicle interruptibility and explainability).

2.5.1 Machine Learning Approaches for Explanation Generation. To explain the decisions of machine learning algorithms for automated driving, explainable models have been proposed. They generate visual or class-wise reasoning behind driving decisions [65, 149]. Some studies introduced textual explanations for more detailed reasoning [67]. Textual reasoning capabilities of transformer models have led to driving scene-based captions that are aware of the driving actions [57]. Recent advances in large language models (LLMs) are even incorporating these models in the reasoning and generation of explanations [150, 152].

Despite the direct impact passengers have with these explanations, most studies focus on generating accurate explanations causal to driving decisions rather than the intelligibility and effective delivery of these machine learning-generated explanations. Indeed, passengers' needs for explanations can vary depending on their behavior and their cognitive and emotional states within the vehicle. This is mainly attributable to the fact that the currently available datasets for explanations are designed for general purposes covering most of the stakeholders, rather than focusing on the end-user passengers. Only a few studies [77, 113, 114] have incorporated passenger aspects in their experiments. In response, despite the known effect of explanations on passenger experiences, the impact of these algorithms on actual passengers under in-situ driving environments has been underestimated.

2.5.2 Related Datasets: Dataset for Explaining Automated Vehicles. Most datasets for generating explanations in automated driving focus on data and annotations for detecting failures or interpreting driving decisions. Typically, these datasets provide explanations in the form of cause-and-effect for accidents [153, 154, 158], centered on driving videos that account for the outcomes of various driving scenarios. Some datasets have been developed to generate textual explanations for automated vehicle passengers, detailing the actions taken by the vehicle and the reasons behind them [65, 67, 149], developed from selected segments of the Berkeley Deep Drive dataset [159]. The introduction of datasets specifically designed for driving instructions [28] and traffic scene explanations [125] has accelerated the development and application of natural language processing techniques in these domains. These data offer detailed annotations for each moment of the included drives, which makes them particularly well-suited for generating continuous or on-demand explanations. However, these datasets may be less effective in generating selective explanations in critical situations where trust and anxiety can be challenged.

Although driving videos are an efficient and reliable data collection method that enables the integration of multiple annotator inputs, they may not fully capture the nuanced demands of actual passengers within automated vehicles or mirror the complexities of real-world driving experiences. Shen et al. [132] introduced the concept of necessity scores for explanations, annotated by individuals watching driving videos. While these studies effectively assessed the need for explanations, watching videos may not fully capture the risk that passengers of automated vehicles might experience [137]. Given the impact of risk levels on passenger experiences and attitudes towards explanations [48, 81], especially among first-time users of automated vehicles [138], there is a need for further investigation into passengers' real-time demands and experiences on the road to tailor explanations and strategically align them with passengers' needs and concerns in automated vehicles. To address this, we extended the data collection environment to actual roads where passengers can experience anxiety, seek situation awareness, manage working memory, experience cognitive load to maintain it, and perform non-driving related tasks that compete with it. In this work, we suggest a dataset in a more ecologically valid setting by incorporating on-road experimental settings and involving passengers in the automated driving scenarios to model their in-situ demand for explanations.

3 Method

In this study, we aimed to gather contextual information that included details about the surrounding environment, the vehicle's driving status, and various aspects of the passengers' states, such as physiological responses or body posture, in relation to their need for explanations of situation awareness. To achieve this, we outfitted a vehicle with a set of sensors and conducted experiments with human subjects to determine their need for explanations. The selection of sensors for our study was guided by a literature review [136] and a design space analysis for in-vehicle interaction [53], which highlighted the importance of kinesthetic, electrodermal, tactile, thermal, and cardiac modalities, alongside visual and auditory input. This data collection took place while the vehicle was believed to be driven via automation but was actually driven by a wizard driver. In the following subsections, we describe the sensor sets that were used and the format in which they were recorded, as well as the detailed data collection procedure, including the experimental protocols and driving scenarios. Implementations used in the paper is available at https://github.com/GWANGBIN/timelytale.

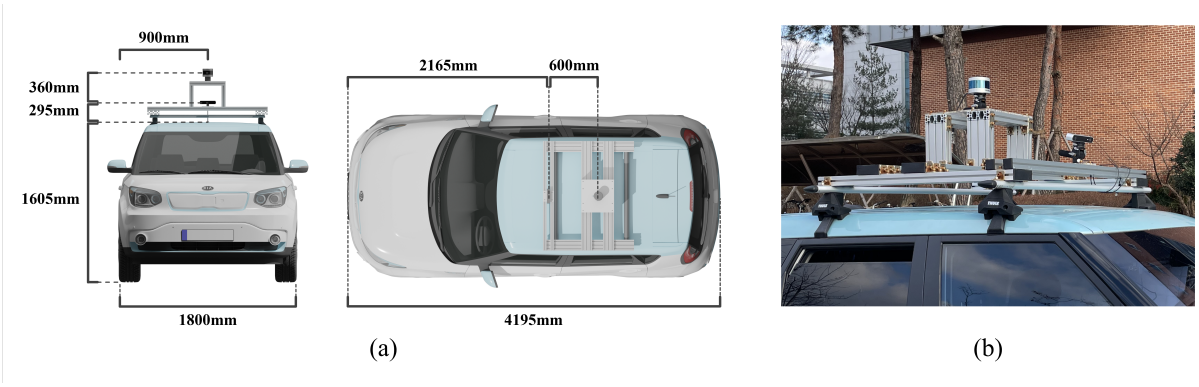


Fig. 1. (a) Model illustrating the dimensions of the exteroception sensors installed in the vehicle. (b) Actual vehicle with the exteroception sensors installed.

Dataset	Purpose	Exteroception	Proprioception	Interoception	Annotation	Size
INAGT [130, 148]	Predicting the optimal moments for proactive driver interruptions	Camera	GPS, CAN, IMU	Camera, Physiological Responses	Yes/No for driver interruptibility	50 hours
BDD-X [67]	Text explanations for AV actions	Camera	GPS	No	Text explanations for video segments	77 hours
BDD-OIA [149]	Associating objects with actions and reasons in driving	Camera	GPS	No	Action categories and text explanations	31.8 hours
Sampled BDD-A [132]	Modeling explanation necessity for AV videos	Camera	GPS	No	Gaze data, necessity scores for explanations	3.1 hours
DoTA [153, 154]	Predicting traffic anomalies from driving videos	Camera	No	No	Annotations for traffic anomalies	20.3 hours
CTA [158]	Identifying unusual traffic events	Camera	No	No	Cause and effect in scenarios	9.5 hours
HDD [122]	Detecting driver behavior and interactions in manual driving scenarios	Camera, 3D LiDAR	GPS, CAN, IMU	No	Driver behavior and causal relationships	104 hours
DRAMA [89]	Visual reasoning of driving risks associated with important objects	Camera	CAN, IMU	No	Important object bounding boxes, free-form caption for risks and interactions	9.8 hours
Rank2Tell [125]	Predicting reasoning and importance in traffic scenes	Camera, 3D LiDAR	GPS, CAN	No	Questions and interpretations of scenes	0.64 hours
Talk2Car [28]	Developing AVs that understand and execute natural language commands	Camera, 3D LiDAR, RADAR	GPS, IMU	No	Bounding box annotations, written commands	4.72 hours
Ours	Modeling the explanation type and timing passengers in automated vehicles demand	Stereo Camera, 3D LiDAR	GPS, OBD-II, IMU	Depth Camera, LiDAR Camera, Physiological Response, Thermal Image, Seat Pressure	Timing and contents of passenger's demand for explanations	15.2 hours

Table 1. Summary of the currently available datasets relevant to in-vehicle explanations and related services, such as interruptibility (INAGT) and language commands (Talk2Car). Following the classification by Omeiza et al. [116], we categorize sensor types as exteroception, proprioception, and interoception. Additionally, we introduce the concept of interoception to account for attempts to model passengers' actual behaviors or states.

3.1 Apparatus and Sensor Settings for Data Acquisition

3.1.1 Exteroception Sensors for Surrounding Environments and Objects. In explainable driving, machine-generated explanations primarily utilize camera data. This type of data has become increasingly vital for producing vision-language explanations, particularly with recent advancements in large-scale vision-language models tailored for driving contexts. The generation of driving explanations is a typical downstream application of these models. To comprehensively capture the environmental contexts in our dataset, we employed a stereo camera and a LiDAR sensor to capture object recognition data, encompassing data for traditional computer vision, stereo vision, and 3D vision with a point cloud. The LiDAR sensor was mounted on top of the vehicle's roof so that other parts of the vehicle did not occlude it. The stereo camera was mounted on top of the roof of the vehicle, positioned directly in front of and below the LiDAR sensor to ensure that neither device obstructed the other (refer to Figure 1 for the detailed dimensions of the installed sensors).

Specifically, a *ZED* stereo camera from *Stereo Labs* was installed to capture the driving contexts. The camera provides a 90 ° horizontal and 60 ° vertical field of view. We recorded the stereo image of the driving context at a resolution of 1344×376 and a frame rate of 15fps. ZED camera images were stored as separate PNG files, each named according to the UNIX time at which they were captured.

We included a 3D LiDAR sensor in our dataset for richer 3D environmental information, particularly in challenging scenarios, such as night conditions or sudden lighting changes in tunnels. This addition provides direct depth information. Specifically, we utilized the *VLP-16* from Velodyne, which offers a 360 ° degree horizontal and 30 ° vertical field of view with a 0.2 ° horizontal and 2 ° vertical angular resolutions. Consequently, our dataset comprised a point cloud with 16 vertical layer channels and 1808 horizontal scanning channels. The point cloud data were collected at a refresh rate of 10 Hz, with each point in the cloud defined by the x, y, and z coordinates, including light reflectance intensity data for each point.

3.1.2 Proprioception Sensors for Vehicled's Movements. We measured GPS data during driving to track changes in the vehicle's latitude and longitude, aiding in the modeling of the demand for explanations with respect to the geolocations. We used a *ZED-F9R dual-band GNSS* module from *u-blox*, which was integrated with a PC via a microcontroller unit (MCU). With the antenna positioned in the middle left of the car's roof, we collected data on latitude, longitude, height above mean sea level, velocity north (toward the north), velocity east (toward the east), velocity down (toward the Earth's center), and heading (direction of movement in degrees from true north) at a 5 Hz refresh rate, each with a UNIX timestamp.

Linear accelerations and angular velocity are used to monitor the vehicle's driving behavior, driving-related events, and road conditions [99]. Also, abrupt changes in motion can indicate sudden vehicle movements and their impact on in-vehicle interruptibility [148]. To capture these data, we installed a 9-axis *HWT905* IMU sensor from *Wit Motion* in the middle of the vehicle. We collected 3-axis acceleration, 3-axis angular velocity, and 3-axis angle data at 50Hz via UART/USB to the PC, with each data point marked with a UNIX timestamp.

To collect the vehicle's operational data, such as speed, throttle, brake, and steering angles, we used the *ELM327* OBD-II module. These features, while traditionally associated with driver interruptibility in manual driving [61, 69], still represent the vehicle's operational behavior and dynamics in the context of automated driving. The OBD-II data were read via Bluetooth to the MCU and then to the PC via a serial port and recorded on the PC at a frequency of 5 Hz, along with the corresponding UNIX timestamp.

3.1.3 Interoception Sensors for Passenger's Pose and States. ***Action Recognition Sensors.*** To provide information about the passenger's head and body posture as well as the NDRTs the passenger was performing, action recognition sensors were installed in front of the passenger. We used two cameras with RGB and depth information: an *D435* depth camera from *Intel RealSense* and an *L515* LiDAR camera from *Intel RealSense*. The RGB and depth images from the *D435* camera were collected at 20 fps in 640x480 resolution, while the depth image stream from

the *L515* camera was gathered at 20 fps in 768x1024 pixels, all in PNG format. Considering the wider field of view *L515* camera has, we installed the camera on the dashboard to capture relatively whole body parts. Meanwhile, *D435* camera was placed higher to focus on the upper body parts, such as the torso, head, and arms, which generally exhibit greater movement than lower body parts in cars.

As this setting involves continuously recording participants whose IDs are mapped with all other sensors and behaviors, we aimed to respect potential privacy issues and conform to the ACM code of ethics [45]. We provide a grayscale depth image to measure or recognize driver postures, actions, and behaviors. Regarding the RGB image, we offer different levels of data depending on participants' consent levels. We provide the original data for those who agreed to the release of raw RGB images. For other participants who refused the release of raw RGB images but agreed to the use of processed data, we provide skeletal tracking and facial mesh data extracted from the RGB data using the Mediapipe framework [87]. This approach minimized the amount of data retained and eliminated potentially identifiable images to ensure participant privacy.

Thermal Imaging In our study, we recorded thermal images of passengers' faces, considering the relationship between nasal and forehead temperature differences and their cognitive load [2, 86], arousal [30], and attention [1]. For this purpose, we used the *Lepton 3.5 module* from *FLIR*, capable of measuring temperatures between -10 to +450 $^{\circ}C$ with a 5% accuracy, suitable for facial imaging [50]. Data from this module were collected at a rate of 8.7 Hz with a resolution of 160×120, covering a 57 ° horizontal and 71 ° vertical field of view. The thermal camera images were stored as separate PNG files, each named according to the UNIX time they were taken.

Tactile Pressure Seat.

The touch, weight, and centroid placement on the car seat vary with the torso, head, and lower body positions. Consequently, pressure sensing on the driver's seat can serve as an unobtrusive indicator of driver posture [164]. Indeed, since the pressure also varies depending on the position of the user's hands, it can be used to identify the passengers' NDRTs such as entering, leaving, drinking, and accessing the glovebox [60] or forward gaze, cell phone use, and sleeping [112]. Pressure sensing can also assess a driver's availability to take over control during automated driving [121]. As shared mobility is predicted to be one of the most imminent forms of automated vehicles [91], privacy-preserving sensing for services in these shared environments is becoming increasingly important, and tactile seats can also contribute to enhancing privacy while providing relevant pose information.

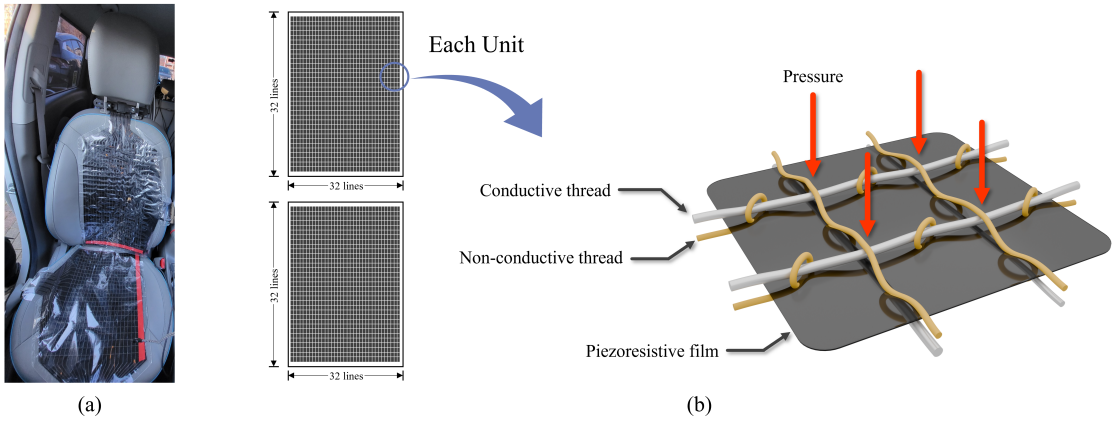


Fig. 2. (a) Resistive tactile sensors installed on the seat and backrest of the car seat. (b) Principle underlying the pressure measurement of these sensors through resistive changes across the matrix.

Given these relationships, we collected pressure sensing data in our study to offer a resource for potential analyses that model seat pressure patterns and passenger behavior in automated vehicles.

Following Zhao et al. [164] and Khazar et al. [60], we installed a resistive-sensing-based tactile seat in our vehicle, covering both the seat and backrest areas (see Figure 2 (a) for reference). However, it should be noted that while previous studies have demonstrated the feasibility of using tactile seats as sensors to measure driver posture and activity, the pressure data could also reflect road bumps and vibrations in on-road driving environments. This could introduce noise when classifying NDRTs, although it may also contribute to passengers' demand for explanations.

This tactile seat consists of a 32×32 electrode array on each side (1024 sensors), designed to sense pressure from electrical signals. We assembled the tactile seat using piezoresistive carbon polyolefin film (Velostat, Desco), silver-plated conductive thread (HC40, Madeira) for detecting pressure, and non-conductive thread for secure attachment. We used the Loen LE-6040 embroidery machine in the fabrication process to ensure the robustness and reliability of the tactile sensors, as compared to hand-crafting.

As depicted in Figure 2, each sensing array includes orthogonally aligned electrodes on both sides of the piezoresistive films, securely stitched together with non-conductive thread. At the intersection where the two conductive threads cross, the piezoresistive film between them changes its resistance under pressure, allowing us to measure the pressure at each contact point. We employed a non-inverting operational amplifier (op-amp) with a reference voltage, coupled with an analog-to-digital converter, for measurement of slight resistance alterations. When pressure is exerted on a tactile seat, it modifies the resistance between two layers of conductive wire arrays. This change in resistance varies the voltage output from the op-amp, enabling the identification of horizontal and vertical conductive threads at the pressure point. The identified coordinates were then sent to the computer via a microcontroller unit. Each sensor can measure pressures up to 14kPa, with the highest sensitivity of 0.3kPa achieved through resistance changes [88].

To capture and record the pose and physiological data, we adopted the modified version [131] of the ActionSense dataset framework [27]. This framework is particularly adept at reliably recording heterogeneous data, including array-based tactile resistive sensing signals, allowing continuous monitoring of the passenger's physical state and interactions with the seat. Each data frame was timestamped using the UNIX timestamp format at the time of acquisition and saved in HDF5 file format with two separate CSV files for the backrest and seat.

Physiological Responses Our dataset was designed to provide real-time data on passengers' actions and states, including their physiological responses that are indicative of driver's cognitive [71] and emotional arousal [70], driver situation awareness [63], and event-related responses (e.g., event-related electrodermal activity (EDA) [25, 82]). For this purpose, we utilized the E4 wristband, which features reliable data acquisition capabilities in environments where motion artifacts are common (e.g., in-vehicle environments) owing to its reduction of and compensation for motion artifacts [42]. This system captured various physiological metrics, including passenger galvanic skin response (GSR) at 4 Hz, interbeat interval (IBI) at 1 Hz, blood volume pulse (BVP) at 64 Hz, temperature at 4 Hz, and XYZ raw acceleration at 32 Hz. These measures are indicators of user arousal in cars [86]. Each data point was timestamped using the UNIX timestamp format at the time of acquisition and saved as CSV files.

The collected data was stored in two formats: a comprehensive HDF5 file for integrated analysis and separate CSV files for each data channel. In addition, the streaming of physiological data was recorded in a video file, alongside the seat pressure data (refer to Figure 3). This video recording was intended to serve as a visual reference for the experimenter to identify any sudden changes in the passenger's physiological response recording because the E4 wristband requires its own server to stream data (by uploading and fetching), while all other sensors are streamed natively. Therefore, we used videos to monitor any disconnect from the network, considering that the experiment was conducted in a vehicle driving on an actual road environment. While designed for data

collection purposes, the video recording can also serve as a visual reference to help identify any abrupt changes in passengers' physiological states during the use of the data.

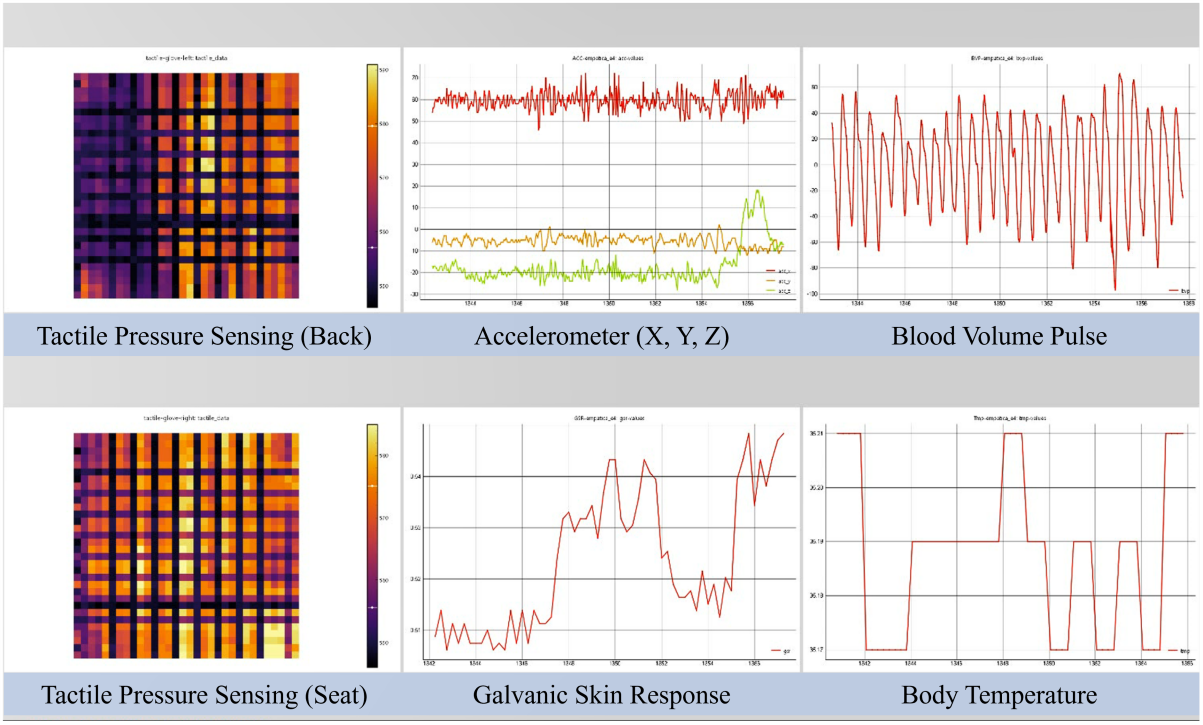


Fig. 3. Video recorded of the streaming of seat pressure and physiological sensors. Tactile pressure sensing data and physiological responses captured from the E4 wristband were recorded as a video plot during the experiments. Note that subtitles in blue boxes were added to the figure for improved legibility, although they were not present in the video itself.

3.2 Automation Wizard

To navigate the ethical and regulatory concerns associated with automated vehicles, we implemented the WoZ methodology [24], which creates the illusion of automated driving for participants with an experimenter acting as a hidden driver [93]. We adopted the RRADs method by Baltodano et al. [10], which conceals the 'wizard' driver behind a partition. To ensure the validity of the WoZ method, we took measures to make the presence of the wizard driver unnoticeable and maintain the illusion of automated driving for our participants.

We installed sensors both outside and inside the vehicle to contribute to the illusion of an actual self-driving vehicle. To hide the driver, we installed a half-mirror glass partition between the driver and passenger compartments, a technique used in deceiving users considering the automation of AI speakers [4]. The glass partition featured a half mirror in the upper part and a full mirror in the lower part. To protect the identity of the 'wizard' driver, we illuminated only the passenger side, ensuring that the half-mirror was half-transparent only when viewed from the driver's side. The seating arrangement of the driver and passenger was planned to complement this setup, ensuring that the passenger could not see the driver, even when the upper half of the glass was transparent. Conversely, the driver could see the right side mirror through the transparent upper half for safe driving (see Figure 4 for the viewpoints of the passengers).

(a) Experimental Settings from the Passenger's Perspective

(b) Experimental Settings from the Wizard Driver's Viewpoint

Fig. 4. (a) Experimental settings from the passenger's viewpoint: the wizard driver is concealed by a partition. (b) Experimental settings from the wizard driver's perspective: the wizard driver has an unobstructed view of the side mirror for safe driving.

We instructed participants that the wizard driver was present only due to regulatory requirements, and they were led to believe that the driver would only take control before entering and after exiting the experimental route. After the experiment, no participant reported noticing the wizard driver driving as it was unseen from their side, and the presence itself was explained. For the experiment's objectivity, the role of automation wizard was done by two instructed drivers: a 26-year-old male with 6 years of driving experience and a 30-year-old male with 10 years of driving experience. Both drivers were informed about the study's objectives and were tasked with the role of wizard driver, with their primary responsibility being to follow the designated route. For a reliable and reproducible experiment for different participants, these drivers were trained on the driving route.

3.3 Procedure

The driving scenarios and experimental protocol outlined below were used to conduct the user study.

3.3.1 Protocol. Before commencing the experiment, the participants were equipped with an E4 wristband and asked to complete questionnaires regarding their age and driving experience. The experiment involved approximately 30 minutes of naturalistic driving.

To encourage naturalistic passenger behavior, we did not impose any specific NDRTs during driving. Instead, we provided snacks, books, and magazines for potential interactions, and the participants were allowed to use their own cell phones. While engaging in these NDRTs, the participants were instructed to immediately inform the experimenter whenever they required additional situation awareness of the car's decisions. They were asked to specify the exact explanation needed, which could include the vehicle's driving actions and/or the underlying reasons. To accurately record the timing of these requests, participants were asked to say 'now' when they needed more explanation, followed by a marker from the experimenter, before detailing the specific explanation required. The explanations required by the participants were coded by consensus between two experimenters, who were

fully knowledgeable about the study and seated at the back. While we acknowledge that it could interfere with a more natural experience in automated vehicles to understand the passenger's in-situ demand, follow-up questions were sometimes necessary to ascertain the specifics of the 'what' and 'why' behind each request when initial details were insufficient. An experimenter reviewed these interactions and annotations post-experiment against the recordings for confirmation. Although the participants were advised that they could stop the experiment at any time if they experienced discomfort, such as motion sickness, no such requests were made.

The coding of explanation data within the context of thematic analysis followed a deductive approach. While we did not restrict the type or content of explanations provided by participants, the experimenters prepared an initial set of codes based on the explanation annotations from Kim et al. [67]. These explanations were refined and unified when they referred to the same concept, despite variations in participant expression. However, when the semantic content or the type of explanation differed—whether addressing the 'what' aspect or the 'why' aspect—they were maintained as distinct codes. A detailed description of the explanation coding is available in the Appendix A.

3.3.2 Driving Scenarios. To facilitate naturalistic experiments in a controlled environment, we set a 14.8km experimental driving route that traverses urban roads, arterial roads, and highways (refer to Figure 5). The route features diverse driving events, potentially leading to various situations influenced by traffic lights and traffic conditions. However, subtle lane changes are not predefined in the driving events of the route, because they are heavily reliant on in-situ environments and interactions with other road users.

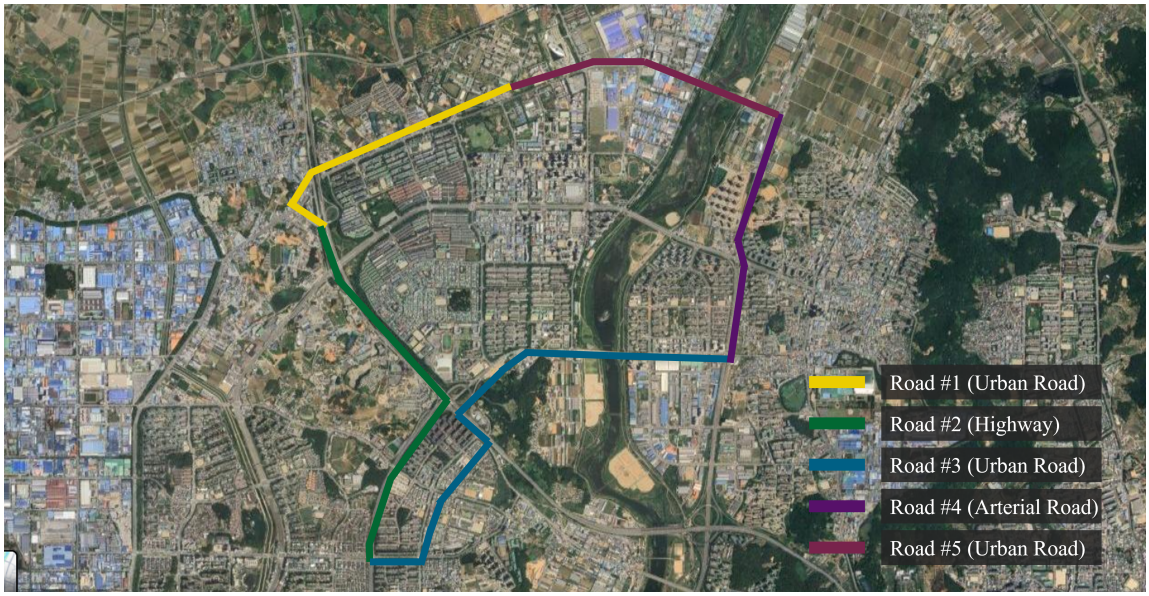


Fig. 5. Experimental driving route encompassing urban road, highway, and arterial.

- Road #1: Urban Road (speed limit: 60 km/h) - Right turn at an intersection with two crosswalks and traffic light, school zone, straight travel at three intersections with traffic lights and two crosswalks each, one underground tunnel, left turn at an intersection with two crosswalks and a traffic light.
- Road #2: Highway (speed limit: 100 km/h) - Uphill drive and right turn for entry, sharp right turn for exit.

- Road #3: Urban Road (speed limit: 60 km/h) - Two left turns at intersections with traffic lights and two crosswalks, three instances of going straight at intersections with traffic lights and two crosswalks each, two left turns at intersections with two crosswalks and a traffic light, school zone, right turn at an intersection with a crosswalk and traffic light, two instances of going straight at intersections with a single crosswalk and traffic light, going straight at an intersection with two crosswalks and traffic light, crossing a river via bridge, two instances of going straight at intersections with two crosswalks and traffic light, left turn at an intersection with two crosswalks and a traffic light, straight travel at an intersection with two crosswalks and a traffic light.
- Road #4: Arterial (speed limit: 80 km/h) - Uphill drive and right turn for entry, sharp decline for exit.
- Road #5: Urban Road (speed limit: 60 km/h) - Two left turns at intersections with two crosswalks and a traffic light, crossing a river via bridge, two instances of going straight at intersections with two crosswalks and a traffic light, three instances of going straight at intersections with a single crosswalk and traffic light, right turn to enter the destination.

3.4 Participants

We recruited 29 participants (10 females and 19 males) with an average age of 30.1 (SD = 10.2, min = 20, and max = 67). Since we assumed highly automated vehicles with SAE levels 4 and 5, we did not restrict our participants to driver's license holders. Of the participants, 20 had driver's licenses, with an average of 5.9 years of driving experience (SD = 5.4, min = 1, max = 20).

3.5 Ethics

All procedures conformed to the principles of the Declaration of Helsinki and were approved by the Institutional Review Board. Prior to the experiment, participants were briefed about the study. They agreed to participate and provided consent for the use and distribution of collected data for scientific research purposes. Given the importance of privacy in data collection, especially with video data where personal identity can be specified, we prioritized the anonymity of our participants. To achieve this, the raw RGB video data is provided only for the participants who provided consent; otherwise, we provide extracted skeleton and facial landmark positions and thermal and depth imaging only where individual identity is not discernible.

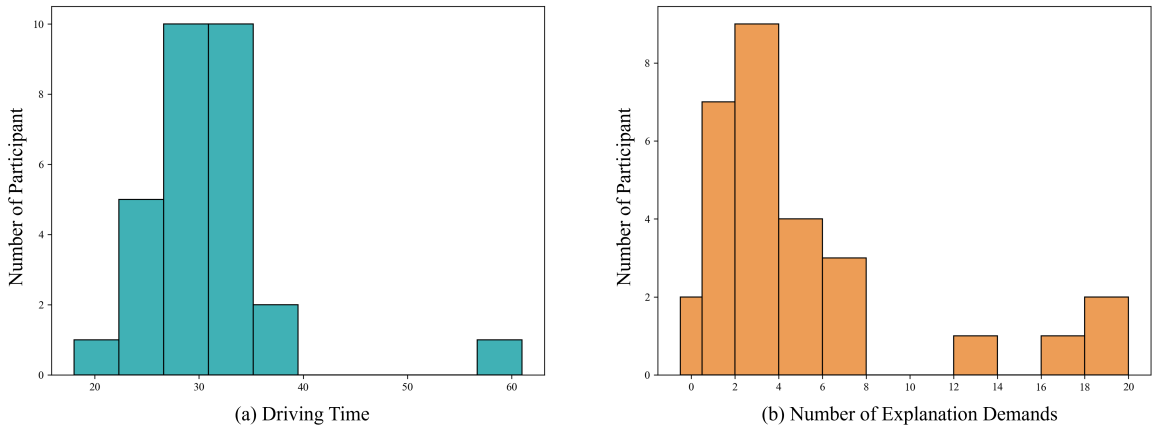


Fig. 6. (a) The driving time each participant experienced and (b) the number of explanation demands during the drive.

4 Result

4.1 Descriptive Statistics for the Data

4.1.1 Demand for Explanations. On average, the driving experiment involved the collection of multimodal exteroception, proprioception, and interoception data, lasting approximately 30.5 minutes for each of the 29 participants, with a standard deviation of 7.1 minutes. The shortest duration recorded was 18 minutes for one participant to complete the experimental route, whereas the longest was 61 minutes, attributed to varying traffic conditions from daytime to nighttime, particularly during evening rush hours. Figure 6 (a) illustrates the dataset's composition in terms of the duration.

Participants requested explanations for an average of 4.5 times during the ride, with a standard deviation of 5.3 (as shown in Figure 6 (b)). Two participants did not seek any explanations throughout their ride, whereas four participants requested more than 10 explanations, reaching up to 19 times during their drive. However, there was no significant correlation between the frequency of explanation demands and driving duration, as indicated by a Pearson correlation coefficient of $r(27) = -0.058$, $p = .766$.

4.1.2 Passenger NDRTs during the Drive. In this section, we report on the non-driving related tasks (NDRTs) that participants performed during the experiment. Figure 7 illustrates the NDRTs each participant engaged in during the drive, with the duration normalized over the total driving time. Although participants were instructed to behave naturally and engage in NDRTs at their own discretion, they performed 11 types of NDRTs (accessing and managing the items in and on the dashboard compartment, drinking water, eating a snack, having a phone call, interacting with a mobile phone, reading a book, reading a magazine, reading a paper, relaxing in the passenger's seat, watching inside the car, and watching outside the window, detailed format is illustrated in subsection B.5).

Figure 8 presents the time allocation for various NDRTs in our study. We compared these findings with those of Detjen et al. [29], who conducted a similar on-road study observing NDRT of passengers in automated vehicles. Our results show that participants spent the majority of their time interacting with mobile phones (42.38%), followed by watching outside the window (28.72%), reading books (10.70%), and relaxing (3.65%). While specific distributions varied to some extent, both our study and that of Detjen et al. [29] found that participants primarily allocated their time to watching out of the window and using smartphones, followed by reading and listening to books.

We compared the frequency of NDRTs in terms of occurrence during the experiment with Pfleging et al. [119], who used an in-situ survey to elicit anticipated preferred NDRTs, and Detjen et al. [29], who employed an on-road experiment to observe NDRTs. To align with the labels used in previous studies, we merged some of our NDRT labels (e.g., 'eating a snack' and 'drinking water' were combined into 'Eating/Drinking').

Figure 9 shows the occurrence of NDRTs in terms of participant frequency. Despite differences in setup, instruction, and the demographic background of the participants among the experiments, 7 out of the 10 NDRT labels mutually reported in both Pfleging et al. [119] and Detjen et al. [29] were also observed in our study. These activities exhibited similar frequencies to those in Detjen et al. [29], which also used an on-road experiment (e.g., watching out of the window, calling, reading).

However, some tasks differed depending on the setup and labeling conventions. For instance, our study involved only one participant per experimental session and did not include interactions with multiple passengers, whereas such interactions were reported in surveys [119] and other studies [29]. Additionally, our experiment specifically labeled activities such as watching inside the car and interacting with the dashboard compartment. While both Pfleging et al. [119] and Detjen et al. [29] differentiated between general smartphone use and smartphone typing, our study did not make this distinction in categorizing mobile phone interactions.

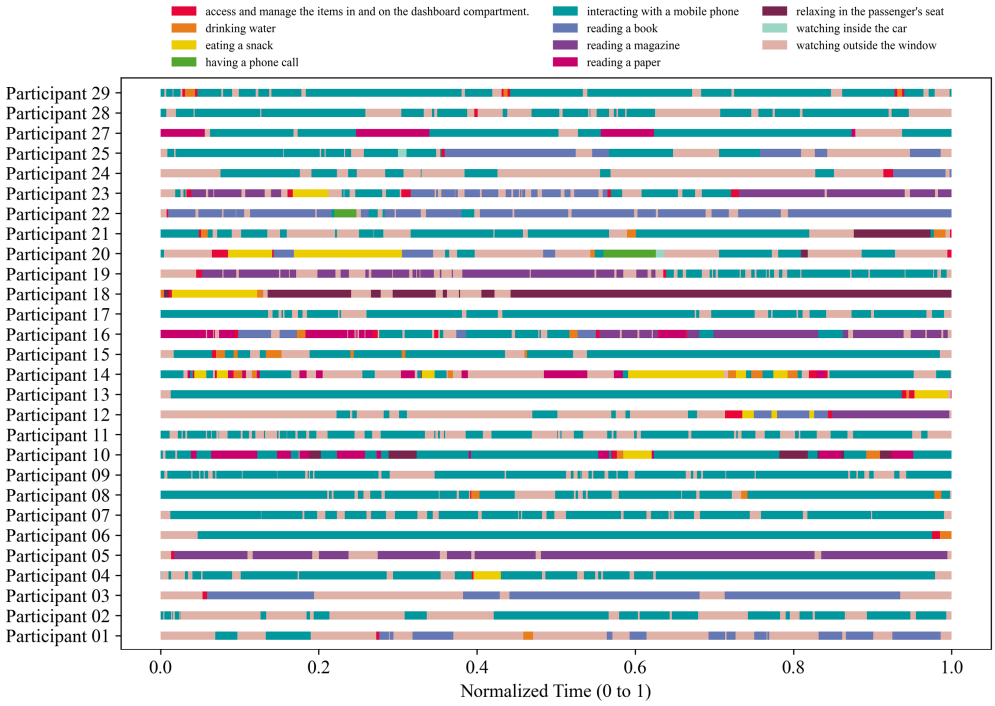


Fig. 7. Timeline of NDRTs performed by each participant during the automated driving experiment.

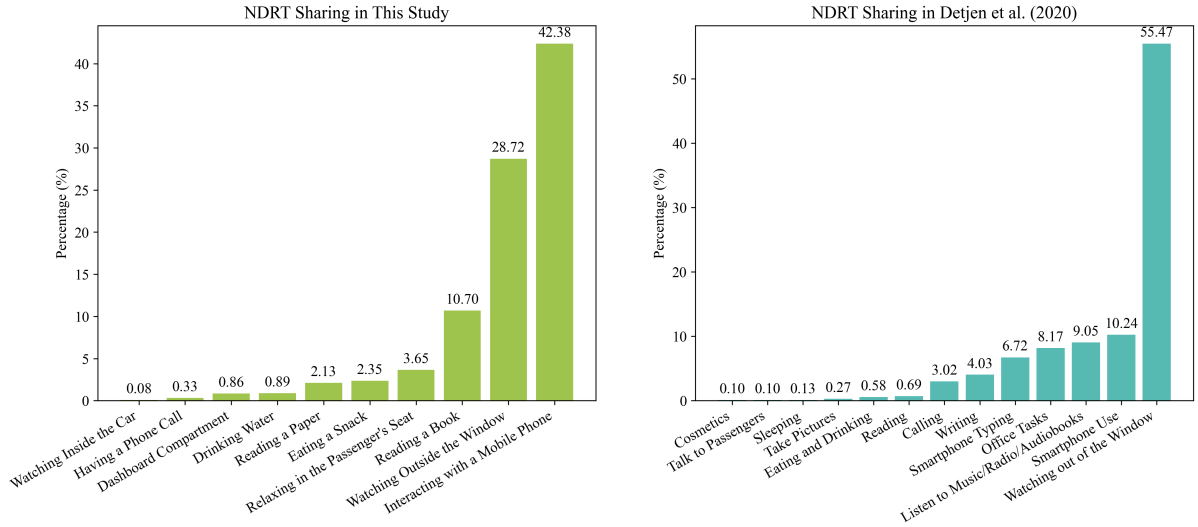


Fig. 8. Time shares of NDRTs performed by participants during the automated driving experiment.

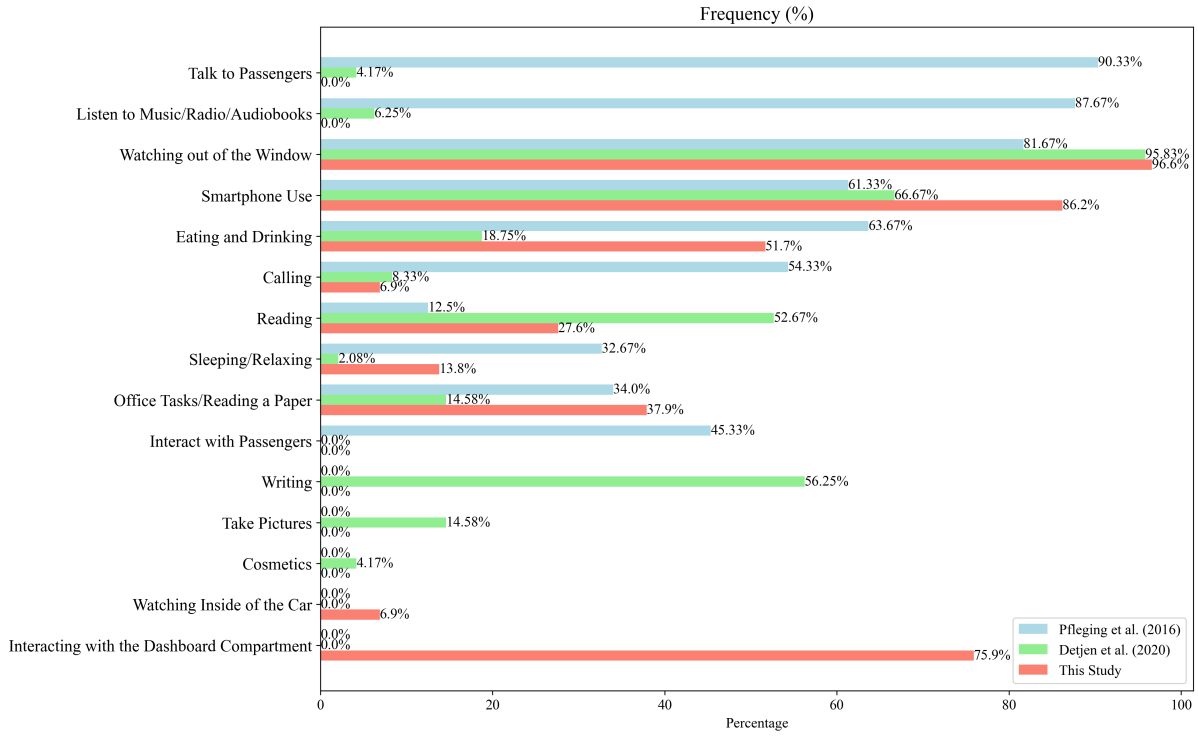
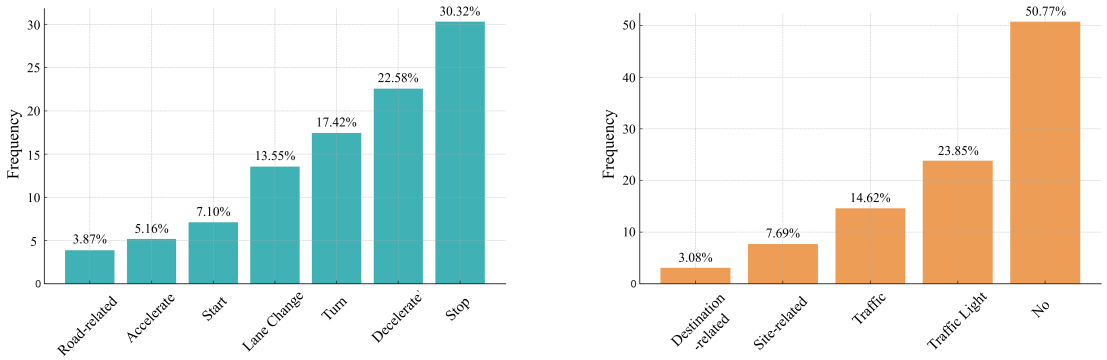


Fig. 9. Frequency of NDRTs in terms of participant occurrence during the automated driving experiment, compared to the findings of Pfleging et al. [119] and Detjen et al. [29].



(a) Distribution of 'What' Explanations

(b) Distribution of 'Why' Explanations

Fig. 10. Types of explanation demands required by participants, classified as 'what' and 'why' explanations.

4.2 Type and Content of Explanation Demands

To understand the type of explanation demands, we classified the collected requests into 'What' and 'What+Why' categories, following the 'action' + 'justification' framework used by Kim et al. [67], which generated explanations for driving behavior from videos. It is also consistent with the 'simple' (content) and 'attributional' (content and reasoning) explanations described by Ha et al. [48]. Despite differences in the framing wording, our classification aligns with the frequently used 'How' (how the vehicle would behave or respond; 'what' in our case) and 'Why' (the reason for it) framework [74, 90]. Of the collected explanation demands, 50.7% were of the 'What' type, focusing on actions, while 49.3% were of the 'What+Why' type, requiring reasoning behind actions. Although we did not apply the framing during the data collection process and were open to 'Why' only explanations, after the classification, we found no demands solely for 'Why' justification without an accompanying action explanation.

We also classified the type of explanation needed depending on the actions to be described within 'what' and 'which' types of explanations (Figure 10). In the specific requirements for 'what' explanations, participants most commonly asked about stopping behaviors (30.32%). This was followed by inquiries about deceleration (22.58%), turning (17.42%), and lane changes (13.55%). There were also questions about starting (7.10%) and accelerating (5.16%), as well as requests for information on specific road features such as highways or merging areas. Regarding 'why' explanations, 50.77% of the time, participants did not seek additional reasons after the action was explained. When they did, the most common queries were about traffic lights (23.85%), traffic conditions (14.62%), and site- or destination-specific details (7.69% and 3.08%, respectively). These findings are consistent with Wiegand et al. [146], where participants often required explanations forecasting the vehicle's movement. Similarly, another study [145] found that participants in simulated driving environments needed explanations in unexpected driving situations, such as abrupt inertial movements due to turns, stops, lane changes, accelerations, etc.

To further investigate the interrelation between the contents of 'what' and 'why' explanations, we utilized a Sankey diagram to connect 'what' and 'why' explanations (see Figure 11). Explanations regarding decelerating, accelerating, stopping, and starting behaviors often necessitate additional reasons, such as traffic status or lights (E.g., 'The car slows down and stops due to traffic', 'The car stops because the traffic light is red'). Conversely, explanations of actions such as lane changes or turns typically do not require accompanying reasons. Moreover, the demand for road-related explanations was not frequently associated with the reasons for these actions.

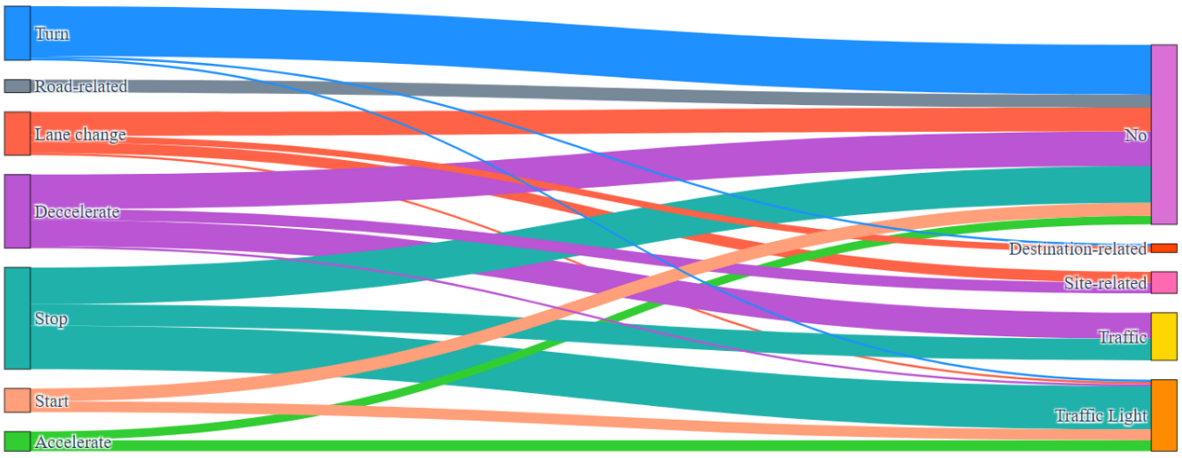


Fig. 11. Sankey Diagram relating 'what' (action) and 'why' (justification) aspects of in-vehicle explanation demands. About half (50.77%) of the explanation demands are for 'what' explanations only, without requiring 'why' justifications.

4.3 Taxonomy of Required Explanations

4.3.1 Data Processing and Clustering Approach. To provide a comprehensive understanding of the types of content in the required explanations within our collected data, we illustrate the composition of these explanations according to the themes they address. To analyze the taxonomy of the explanations, we translated the original deductive coding (within the context of thematic analysis) for passenger demand into binary coding with presence-absence within the context of content analysis. This presence-absence coding was designed to quantitatively illustrate the frequency of specific word labels from the original deductive-coded explanation demand, allowing multiple labels to be present for a single explanation (multi-label encoding). We expanded the number of label codes for each explanation, coding the explanations with 21 distinct labels: turn, right, left, speed, accelerate, decelerate, lane change, start, stop, exit, enter, road, highway, school zone, merging, traffic, traffic light, red, yellow, green, and destination.

For quantitative analysis of label frequency within explanation demands, we then clustered the required explanations that were coded into presence-absence labels using the K-means clustering algorithm. For visualization and analysis, the dimensionality of the data was reduced using uniform manifold approximation and projection (UMAP), with the target dimensionality set to 2 for visualization in a two-dimensional space. The optimal number of K-means clusters was determined using silhouette analysis, which measured the compactness and separation of clusters, with a higher silhouette score (close to 1) indicating better clustering. Based on the silhouette analysis, the optimal number of clusters for the given dataset was found to be 6, with a silhouette score of 0.958. We describe the six distinct types of frequently required explanations as visualized in (Figure 12).

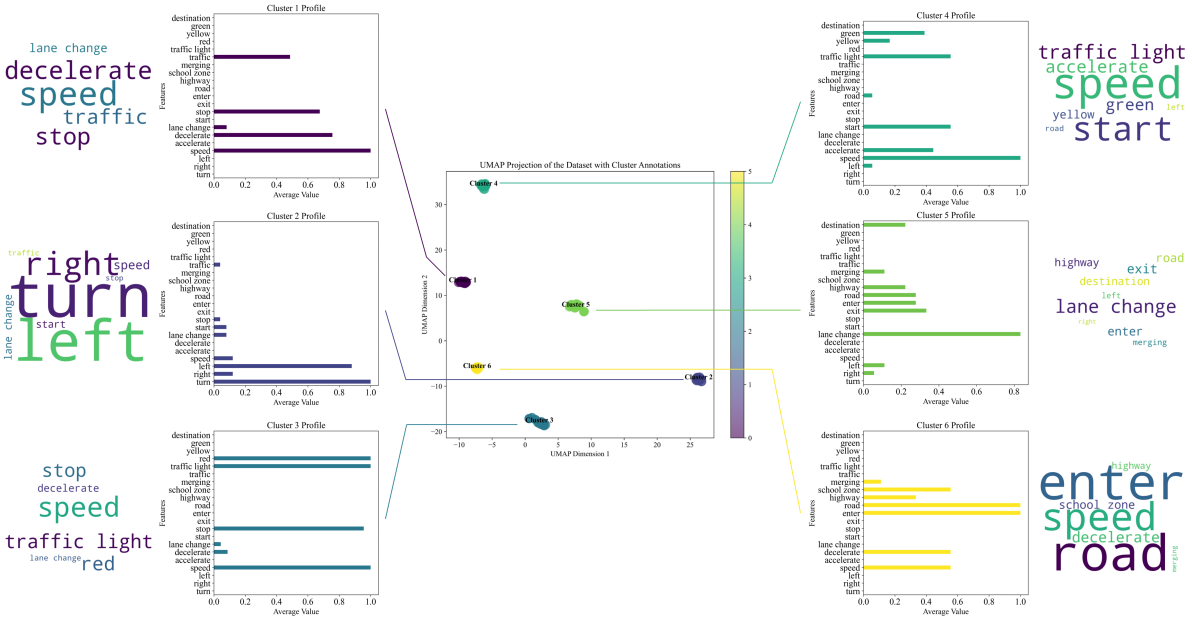


Fig. 12. Taxonomy of the required explanations clusters among the collected data, mapped on the UMAP with frequency count and word clouds for each cluster presented.

4.3.2 Clusters of Required Explanations. ***Cluster 1****: Decelerate and stopping behavior due to traffic.* This type shows the need for explanations in scenarios where the car decelerates or stops in response to traffic conditions. Passengers asked for explanations for the slowing down or stopping behavior. In some situations where the vehicle slows down and then stops, some people demanded explanations about the deceleration, some people demanded explanations for the stopping only, while others asked for both.

Cluster 2*: Turning right and left.* This type shows the demand for explanations for abrupt changes in rotational inertia, which is the turn the car makes on roads. Some passengers required these turns to be preceded by explanations in front of traffic lights, which included simultaneous explanations of the vehicle's start, stop, and lane change during its wait at the intersection, along with the future turn explanation. Referring to the Sankey diagram (Figure 11), we observed that most turn-related explanations did not require accompanying 'why' type justifications. Only a few cases included explanations for the justification, such as the traffic status.

Cluster 3*: Stopping behavior due to a red light.* This type explains the need for explanations for the vehicle's stopping behavior when it stops in front of a red light to wait for traffic. While passengers mostly wanted an explanation about the stop or the stop and waiting behavior, some of them asked about the deceleration and the stop.

Cluster 4*: Starting/accelerating behavior due to a green/yellow light.* This type illustrates the need for explanations for abrupt starts and acceleration. Passengers asked to be informed about the change in acceleration when the traffic light turns green and the vehicle starts, as well as when the vehicle accelerates a bit when the traffic light turns yellow during its crossing. Some demanded explanations were solely for the speed, such as precaution, without the need for reasons for acceleration. This cluster captures urban driving experiences characterized by frequent stop-and-go movements controlled by traffic lights. It focuses on a car's response to traffic signals and specifically explains its transition from a complete stop to a starting movement.

Cluster 5*: Lane change behavior due to entering/exiting a highway/merging/destination.* This type addresses the need for explanations for lane changes. The need for explanations for lane changes included various road type situations such as entering or exiting a highway, merging points, or reaching a destination.

Cluster 6*: Decelerating in a school zone.* This type of explanation shows the need for an explanation of the vehicle's slowing down behavior when it enters a school zone. This cluster also addresses the car's adaptive maneuvers, such as slowing down in response to entering or merging points. These maneuvers are often in response to varying traffic road situations or involve actions to enter, exit, or merge onto different types of roads, such as highways or merging points.

4.4 Explanation Demands According to NDRTs

To understand how the NDRT a passenger was performing affected their demand for explanations, we analyzed the NDRT at which passengers demanded explanations. In doing so, we used the NDRT data and applied a 10-second window around the moment the explanation demand was captured. We only designated the NDRT as "watching outside the window" if the participant was labeled as watching outside the window before, at the moment of, and after the explanation demand. This was done to confirm the true NDRT that the passenger was engaged in when they felt the need for explanations, excluding the moment-wise heading up for quick glance or reporting of explanation demand.

Figure 13 illustrates the explanation demand ratios per hour. The demand ratio was highest when passengers were relaxing in the passenger's seat and watching outside the window. These activities are categorized as requiring low visual, auditory, cognitive, and psychomotor demand [66]. Accessing and managing items on and in the dashboard compartment and eating a snack can be classified as NDRTs that require medium visual and psychomotor and low auditory and cognitive demand [66] while reading and interacting with a mobile phone belonging to NDRTs with high visual and cognitive and medium psychomotor demand [66]. The need for

explanations was lower in these two scenarios than in scenarios where passengers were engaged with relatively low-demanding NDRTs. While designing intelligible explanations in these scenarios for successful delivery is still important [108, 162], the demand for explanations itself was higher when passengers' attention resources were less occupied with NDRTs.

The result is in line with the concept of self-interruption of drivers in conditionally automated vehicles, where passengers' self-interruption is meant to enhance their situation awareness [43]. This happens more often when they have some level of situation awareness to develop further, as perception is the first stage followed by comprehension and projection in Endsley's situation awareness model [35]. Similarly, in our results, participants were assumed to demand more explanations when they had some level of situation awareness regarding driving compared to being disconnected from it owing to NDRTs. However, it should be acknowledged that our analysis is preliminary, with a limited number of participants and explanation demands, and that the results should be confirmed and could be strengthened with longitudinal studies.

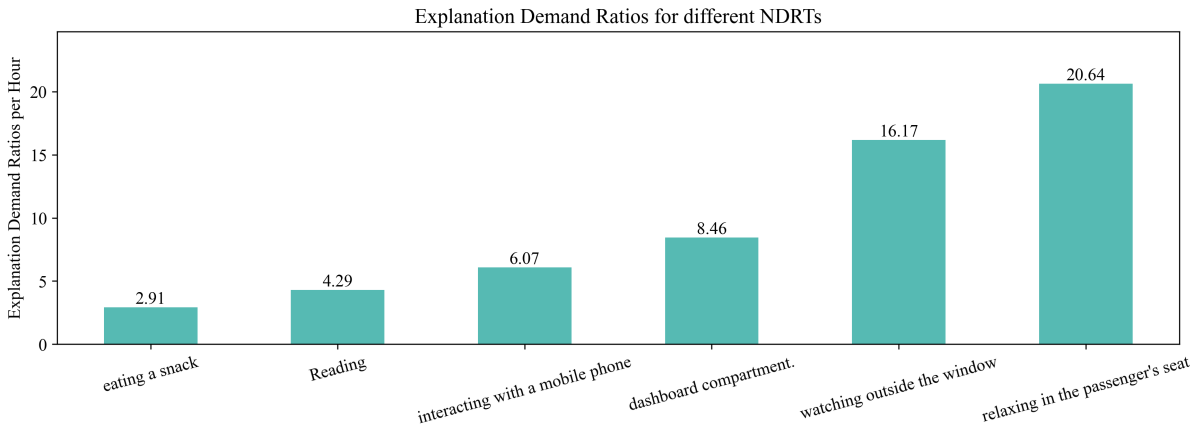
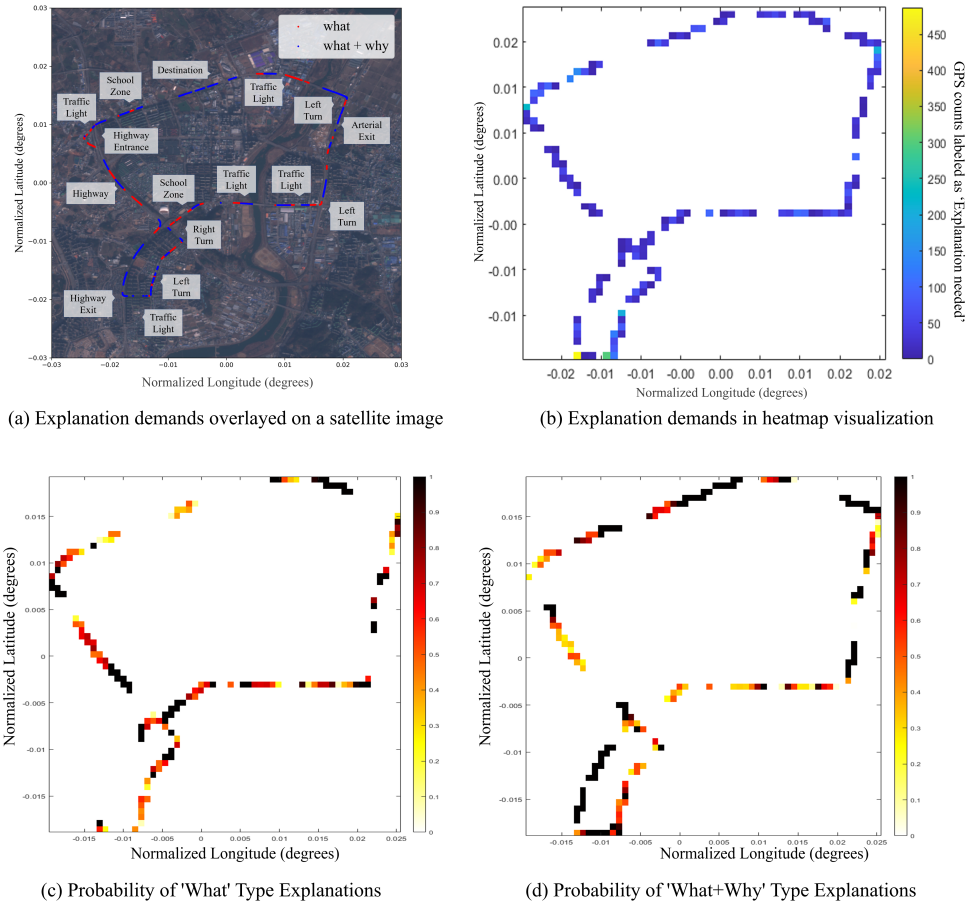


Fig. 13. Explanation demand ratios per hour for different NDRTs.

4.5 Geo-Positioned Explanation Demands

As explanation demands were gathered with proprioception sensors, the demand can be visualized over an actual map using GPS coordinates Figure 14 (a). We plotted the route characteristics to help understand the sources of explanation demand. The analysis identified specific explanation-sensitive areas, particularly at major corners or locations with significant changes in driving routes, path types, or areas with important traffic lights or intersections. Comparing with Figure 5, key areas for explanation demand included transitions from urban roads to highways, highways to urban roads, arterial roads to urban roads, and other junctions involving major turns or changes in the route.

To visualize the area-specific sensitivity with intensity information, we created a grid-based heatmap (see Figure 14 (b)). This heatmap illustrates the distribution of expressed explanation demands across various geolocations, along the designated route. The numbers on the heatmap represent the count of GPS signals (10 Hz) annotated as 'explanation needed', plotted over a geographical grid. We also observed varying demands for explanation types across different zones. As shown in Figure 14, certain areas, highlighted in black, demonstrated a clear preference for either 'what' type (Figure 14 (c)) or 'what+why' type explanations (Figure 14 (d)).



(a) Explanation demands overlayed on a satellite image

(b) Explanation demands in heatmap visualization

(c) Probability of 'What' Type Explanations

(d) Probability of 'What+Why' Type Explanations

Fig. 14. (a) GPS data plotted on a satellite image of the experimental site. (b) Explanation demands using heatmap visualization. (c) heatmap visualization for the 'what' type explanations. (d) heatmap visualization for the 'what+Why' type explanations. The total grid comprises 60 tiles in each direction, covering 0.0381° of latitude (2' 17") and 0.0451° of longitude (2' 42"). The plotted GPS value has been normalized by subtracting 35 degrees from latitude and 127 degrees from longitude.

Vision-language models employed in explanation generation often utilize different algorithms to cater to diverse explanation needs. For instance, a significant number of 'what' explanations can be efficiently addressed through visual question-answering tasks [6], whereas some require direct reasoning using language models that guide these models in interpretable end-to-end driving [150].

Action justifications, on the other hand, are often conveyed through interpretable textual explanations [67, 149], but they can also be effectively generated by vision-language models [57]. The choice of the explanation generation method, and consequently the data type, should be informed by the specific demand for explanation types. Our data can be used to align the algorithm's intended scope of explanation with the specific explanation needs identified.

5 Baseline Machine-learning Analysis: Proprioceptive and Interoceptive Modeling of Explanation Timing

We conducted baseline machine learning analyses to validate our dataset's utility in predicting the timing of passengers' explanation demands. These models were designed to classify passengers' demand for explanations within automated vehicles using data related to vehicle dynamics and passengers' physiological responses. Our ablation study was structured into three key areas:

- **Abation 1: Model Performance and Feature Contribution.** We performed time-series classification to determine the timing for explanations to provide the baseline accuracy and demonstrate the potential of the dataset. We investigated the contribution and importance of different features in determining demand timing for explanations.
- **Ablation 2: Multimodal Fusion Methods.** We assessed the impact of the method for combining multimodal data on the accuracy of our models to guide the preferred input structure.
- **Ablation 3: Temporal Window Sizes.** We investigated how the input and annotation windows for explanation demand timing affect the model's accuracy.

5.1 Ablation 1: Model Performance and Feature Contribution.

5.1.1 Dataset Preparation. Given the role of abrupt motion in interruptibility [148] in cars and attitude towards explanations in cars [145], we used the IMU sensor data as a basis input feature channel and conducted ablation study by adding multimodal contexts, such as OBD-II and physiological responses to see how the added contexts may or may not contribute to the enhanced prediction of passenger's demand for explanations.

Specifically, we used 9-axis acceleration data from IMU, speed, throttle, brake usage, and steering angles from OBD-II. Additionally, inter-beat interval and galvanic skin response data, including phasic and tonic features (skin conductance level and response), were used. The data were preprocessed to clean and normalize by applying filtering conditions to ensure the data falls within ranges for each. Data sequences were labeled to reflect participants' explanation demands within a -5s to +5s window, following the methodology of Semmens et al. [130]. This labeling approach accommodated the reaction times of the participants and experimenters in our study.

5.1.2 Training and Evaluation Method. For our analysis, we employed K-fold cross-validation with $K = 5$ to guarantee training variance and mitigate neighborhood bias. In doing so, data from a particular instance of explanation demand were exclusively assigned to either the training or validation stage to keep the model's generalizability across various explanation instances. Given the predominance of non-explanation scenarios in the dataset, class weights were adjusted during training to rectify the imbalance in the 'explanation needed' class.

The models were created and compiled with the Adam optimizer, focal loss function, and accuracy as the evaluation metric. We trained the models using a mini-batch approach with model checkpointing and implemented learning rate decay to aid convergence. Early stopping and regularization were also applied to prevent overfitting. However, the frequent activation of early stopping in earlier epochs indicated regularization challenges in our dataset, suggesting some degree of overfitting to the training data despite these preventive measures.

Acknowledging the presence of 'normal' situations without explanation demands, our focus remains on the precise prediction of 'anormal' situations with explanation demands. Consequently, we report balanced accuracy metrics, calibrated to maintain a balance between 'explanation needed' and 'not needed' labels, with a baseline accuracy established at 50%.

5.1.3 Model Selection. We compared six models that have shown performance in sequential data prediction tasks, each of which uses Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), and these combined

with Attention or Convolutional Neural Network (CNN) layers, all combined with a dense layer: LSTM-Dense, Bi-LSTM-Dense, LSTM-Attention-Dense, Bi-LSTM-Attention-Dense, LSTM-CNN-Dense, and Bi-LSTM-CNN-Dense.

The precise number of layers, as well as the number of nodes in each layer, were tuned based on the balanced accuracy achieved when utilizing all available features. Detailed information about the architectures of the employed models can be found in Appendix C. The selection of layers in each model architecture was based on the following considerations.

- **LSTM:** LSTM is a specific form of recurrent neural network (RNN) for sequential data that decreases the vanishing gradient problem of RNN by employing gates that regulate the flow of information [47]. We considered LSTM as base model given its performance in time-series classification [133]. We used the configurations that Chen et al. [17] employed in classifying opportune moments to interrupt in VR using gaze and interaction data as a baseline and modified them to fit our specific problem.
- **Bi-LSTM:** Bi-LSTM can process data in both forward and backward directions, which allows capturing context from both sides of a sequence point where the meaning of a sensor data can depend on the preceding data and the following response. We considered the use of Bi-LSTM due to its potential in decreasing error in time-series classification [133]. As Bi-LSTM has been shown to diminish overfitting compared to simple LSTM in tasks estimating driver behavior [100], we used Bi-LSTM model to tackle the regularization challenge of the task and dataset this paper tackles.
- **LSTM-Attention:** The attention mechanism, introduced by Vaswani et al. [141], allows the model to consider the weighted impact of each step of the input sequence on the output when combined with LSTM [101]. We considered the use of the Attention layer with LSTM due to the reported increase in performance in driver behavior estimation from CAN-bus [100, 101] and driver emotion recognition [102]. We positioned the attention layer between LSTM and Dense layers to refine the features processed by the LSTM by selectively focusing on salient elements of the sequential input.
- **LSTM-CNN:** The combined use of LSTM and CNN model helps capture both temporal dependencies and local spatial features within the time-series data [58]. This characteristic of temporal and spatial feature extraction has led to its performance in recognizing driver behavior from a vehicle's proprioception measures [55] and VR interruptibility [17]. We positioned the CNN layer between LSTM and Dense layers to extract features across different segments of the input data from different channels such as IMU, OBD-II, and physiological responses.

5.1.4 Loss Function. We incorporated the focal loss to ensure that training wasn't dominated by the most frequent labels. The focal loss adjusts the contribution of each example to the overall loss during training (Equation 1), where:

- α_t is a weighting factor that balances the importance of positive and negative examples.
- p_t is the estimated probability for the ground truth class.
- γ is the focusing parameter that controls the degree of down-weighting for easy examples.
- $\log(p_t)$ is the standard cross-entropy loss.

$$\text{Focal Loss} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{1}$$

We tuned the hyperparameters α_t from 0 to 1 and γ from 2 to 5 and set them as $\alpha_t = 0.95$ and $\gamma = 2$.

5.1.5 Baseline Machine Learning Result. As reported in Table 2, the Bi-LSTM-CNN network demonstrated the highest balanced accuracy at 86.41%, closely followed by the LSTM-CNN model with an accuracy of 85.94% when

Table 2. Performance metrics of models on predicting moments for explanation demands

IMU Only					
Model	Accuracy	Bal. Acc.	Precision	Recall	F1-score
Majority Baseline	95.24	5.00	-	-	-
Random Guess	90.90	49.60	3.95	3.99	3.97
LSTM	96.50	67.22	79.72	34.91	47.89
Bi-LSTM	96.36	65.82	79.23	32.06	45.46
LSTM-Attention	96.68	66.95	89.45	34.12	49.28
Bi-LSTM-Attention	96.85	68.43	92.43	37.02	52.82
LSTM-CNN	96.74	77.10	72.98	55.43	62.07
Bi-LSTM-CNN	**97.32**	**75.63**	**85.18**	**51.70**	**64.24**
IMU + OBD					
Model	Accuracy	Bal. Acc.	Precision	Recall	F1-score
Majority Baseline	95.33	50.00	-	-	-
Random Guess	90.91	49.77	4.33	4.28	4.31
LSTM	97.43	80.91	78.66	62.66	69.70
Bi-LSTM	97.33	79.11	80.12	58.97	67.76
LSTM-Attention	97.25	76.43	82.22	53.44	64.52
Bi-LSTM-Attention	97.49	77.65	85.98	55.75	67.49
LSTM-CNN	97.49	85.38	73.89	72.01	72.91
Bi-LSTM-CNN	**97.63**	**84.33**	**77.85**	**69.64**	**73.45**
IMU + OBD + Physiological Response					
Model	Accuracy	Bal. Acc.	Precision	Recall	F1-score
Majority Baseline	95.36	50.00	-	-	-
Random Guess	91.06	49.95	4.60	4.56	4.58
LSTM	97.52	85.43	74.46	72.08	73.23
Bi-LSTM	97.63	84.24	78.17	69.46	73.35
LSTM-Attention	97.59	82.26	80.21	65.34	71.95
Bi-LSTM-Attention	97.32	83.69	73.37	68.63	70.76
LSTM-CNN	97.72	85.94	77.58	72.92	75.14
Bi-LSTM-CNN	**97.69**	**86.41**	**76.46**	**73.95**	**75.16**

all features were used. The result implies that the combined use of LSTM and CNN allowed for the extraction of temporal and spatial features that relate one sensor or data channel to another.

The Bi-LSTM-CNN network achieved its highest accuracy when used with IMU, OBD-II, and physiological responses, allowing for the extraction of inter-relational features among them. Similarly, in our ablation study comparing this and other models, the addition of driving- and physiological response-related features revealed an increasing trend in accuracy. These results suggest that our multimodal approach can capture nuanced contexts of explanation demands than using only proprioceptive sensing measures.

Although using all features resulted in the highest balanced accuracy, the combined use of IMU and OBD also provided a reasonable level of accuracy with our baseline model, achieving balanced accuracies of 84.33%, respectively (baseline: 50%). These findings imply that, while the incorporation of E4 wristband data could lead

to increased accuracy, our method can be integrated with existing vehicle settings without requiring passengers to wear obtrusive sensors to be subject to timing-opportunistic explanations.

The validation results demonstrate the reliability of driving and physiological features as indicators of explanation demands, as substantiated by the balanced accuracy (compared to the baseline balanced accuracy of 50%) and F1-score. However, the F1-score is lower than the accuracy, indicating that training on the imbalanced dataset was not entirely successful. While precision and recall are similar in all feature cases, recall is smaller than precision when only IMU or IMU and OBD-II are used. A lower recall than precision suggests that the model is more conservative in predicting the 'explanation needed' class, resulting in a higher number of false negatives compared to false positives.

This disparity suggests less balanced training results, a consequence of the task's nature and its reflection on the dataset composition, where most driving situations do not require explanations and are not UX-vulnerable scenarios. Models specifically designed for imbalanced scenarios, such as anomaly detection models, could address the data imbalance issue. Alternatively, employing a more regularized protocol for annotating explanation demands could be beneficial, although this might compromise ecological validity.

5.1.6 Feature Importance in Explanation Demand Prediction. To verify the importance of each feature from multimodal sensor channels in determining the moment for explanation demands, we conducted a permutation test using the Bi-LSTM-CNN model that incorporated all feature channels (IMU+OBD+Physiological responses). The permutation importance was calculated by iterating over each feature, randomly permuting its values, and measuring the decrease in the model's performance (F1-score) on the permuted data compared to the original, unperturbed data. This difference represents the importance of the permuted feature, as features that cause a larger performance drop when permuted are considered more important for the model's predictions. The permutation importance for all features was obtained by repeating this process for each feature. The results are illustrated in Figure 15.

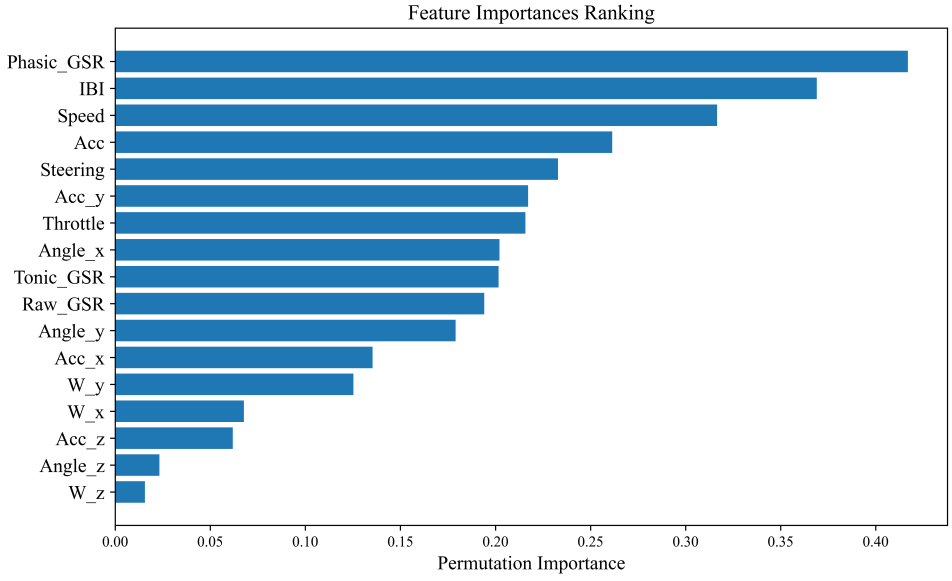


Fig. 15. Feature importance measured from permutation study using Bi-LSTM-CNN model with all features.

The result from feature importance analysis indicated that physiological responses from the E4 wristband, specifically phasic GSR and inter-beat interval (ibi), are the most indicative features in determining the demanding timing of in-vehicle explanations. This finding partly echoes the results from a machine-learning ablation result (Table 2), which showed an increase in demand prediction accuracy with the adoption of physiological data. Following these, speed, acceleration, steering, and throttle from OBD-II were identified to have a high influence on the model predictions, aligning with the ablation test's findings that incorporating these data enhanced the accuracy beyond an IMU-only model.

Despite the known effect of abrupt motion on in-vehicle interruptibility [148], the IMU, a direct motion sensor, was less indicative of the passenger's demands for explanations regarding the driving actions. Conversely, OBD-II features were generally more indicative. This is likely because passenger demands for explanations often related to the vehicle's operational maneuvers. OBD-II data, providing direct measurement into specific actions that vehicles take, such as deliberate acceleration and deceleration at crucial moments (e.g., stopping at a traffic light or speeding up to pass before it turns red), had a more importance in determining the timing for explanations.

5.2 Ablation 2: Multimodal Fusion Strategies

Our feature ablation study highlighted the importance of a multimodal approach in improving the accuracy of predicting the moment for a passenger's explanation demand. This is because multimodal fusion can contribute to a fused representation and increased certainty in data, leading to improved accuracy and reduced error [41]. To understand how different fusion strategies could affect the accuracy of multimodal prediction, we compared early and late fusion approaches.

- **Early Fusion**: In the early fusion approach, we used the all-feature models described in subsection 5.1. The model takes the IMU, OBD-II, and physiological response data as input to a single LSTM layer, which undergoes further processing by the subsequent layers to produce the final decision. In this method, the LSTM layer is expected to learn the temporal patterns and interactions across all sensor modalities simultaneously [41].
- **Late Fusion**: For the late fusion approach, we used models identical to the all-feature models in subsection 5.1, except for the input and input LSTM structure. By having separate Bidirectional LSTM layers for each sensor channel, the model is designed to capture the unique characteristics and dynamics of each modality independently. This method treats each sensor modality as a distinct input stream, allowing the model to learn modality-specific temporal patterns and features [41].

We compared the accuracy of different models with respect to the fusion strategy employed (Table 3). The balanced accuracy and F1-scores were higher for the early fusion approaches, while the model architecture other than the input structure remained the same. Additionally, while the Bi-LSTM-CNN model showed the highest accuracy with the early fusion method, the accuracy of the Bi-LSTM model was higher than that of the Bi-LSTM-CNN model with late fusion. This result is attributable to the fact that the convolutional layers, which capture statistical correlations between different modalities [40], could extract less inter-feature information when these inputs were fed with separate LSTM streams.

Our ablation study on multimodal fusion showed that the early fusion strategy performed better than the late fusion approach for this problem. This indicates that the correlation between different modalities, such as the vehicle's actions and the passenger's physiological response, could be more important in analyzing the passenger's demand for explanations in automated vehicles.

5.3 Ablation 3: Temporal Window Sizes

We conducted an ablation study to investigate the effects of the annotation window size and input window size on the model's performance. The input window size was explored since our models require a certain number of

Table 3. Performance Comparison of Early- and Late-Fusion Strategies

	Early Fusion					Late Fusion				
Model	Accuracy	Bal. Acc.	Precision	Recall	F1-score	Accuracy	Bal. Acc.	Precision	Recall	F1-score
Majority Baseline	95.36	50.00	-	-	-	95.26	50.00	-	-	-
Random Guess	91.06	49.95	4.60	4.56	4.58	90.94	50.06	4.86	4.90	4.88
LSTM	97.52	85.43	74.46	72.08	73.23	96.52	75.13	67.56	51.51	58.16
Bi-LSTM	97.63	84.24	78.17	69.46	73.35	**97.45**	**83.74**	**75.36**	**68.61**	**71.70**
LSTM-Attention	97.59	82.26	80.21	65.34	71.95	96.27	73.42	63.65	48.20	54.83
Bi-LSTM-Attention	97.32	83.69	73.37	68.63	70.76	96.93	78.82	71.28	58.83	64.36
LSTM-CNN	97.72	85.94	77.58	72.92	75.14	96.66	82.62	64.17	67.12	65.50
Bi-LSTM-CNN	**97.69**	**86.41**	**76.46**	**73.95**	**75.16**	97.43	83.18	75.45	67.45	71.09

time steps in the input sequence for making predictions. Specifically, we compared input window sizes of 3, 5, 7, and 10 seconds. The annotation window size was also considered because the annotation process naturally involves some level of delay. The default annotation window was set to -5 to +5 seconds around the recorded UNIX timestamp. However, we varied the annotation window size to 3, 5, 7, and 10 seconds in both directions to account for potential variations in the delay. The rationale behind exploring different annotation window sizes was to capture the temporal range within which a passenger might perceive the need for an explanation, as this delay can depend on individual participants and specific situations.

We compared the performance of the Bi-LSTM-CNN model with varying levels of annotation and input windows. While all other data preparation conformed to the previous ablation studies, the explanation data was annotated according to the annotation window size, which resulted in differences in data distribution. Although the majority baseline for an annotation window of 3 seconds was 97.98%, the majority baseline for an annotation window of 10 seconds was 94.68%. However, we compare balanced accuracy (baseline: 50%) and F1-score, which are not affected by the change in distribution. The results indicate that the balanced accuracy and F1-score generally increase within the range we searched as we accommodate larger input and annotation windows (Figure 16).

5.3.1 Larger Annotation Window Yields Higher Accuracy. Within the investigated range of annotation windows, a larger window yielded better prediction results, with the 10-s annotation window showing the highest accuracy of 86.75% when used with a 10-second input window. This result is attributable to that a larger annotation window increases the likelihood of capturing delayed reactions. Conversely, a smaller annotation window may exclude the delayed response due to individual differences, participants' potential hesitancy that induced delay, and complex cases of each reporting. While the benefits of a larger annotation window used in our experiment seemed to have worked more dominantly, a smaller annotation window can tightly focus on the most relevant temporal proximity to the event, whereas a larger annotation window may introduce noise by including irrelevant data points that dilute the model's ability. As our ablation study indicated the impact of annotation window size on model performance, the trade-off between the size of the annotation window should be considered when designing models to predict opportune moments to provide explanations, though the explored range preferred a larger annotation window.

5.3.2 Larger Input Window Yields Higher Accuracy. Within the investigated range of input windows, a larger window yielded better prediction results, with the 10-second input window showing the highest accuracy of 86.75% in the 10-second annotation window case. This is mainly because a larger input window provides a broader context and sequences, which could involve some level of demand being developed in the long term. By considering a wider time frame, the model can learn from a richer set of data points and identify patterns that

may not be apparent within a smaller window. This increased contextual awareness can lead to more accurate predictions of explanation demand timing. While the benefits of a larger input window used in our experiment worked more dominantly, a smaller input window can provide quicker predictions suitable for real-time prediction, whereas a larger input window could lead to slower response times due to increased model complexity. As our ablation study indicated the impact of input window size on model performance, the trade-off between the size of the input window should be considered when designing models to predict opportune moments to provide explanations, although the explored range preferred a larger input window.

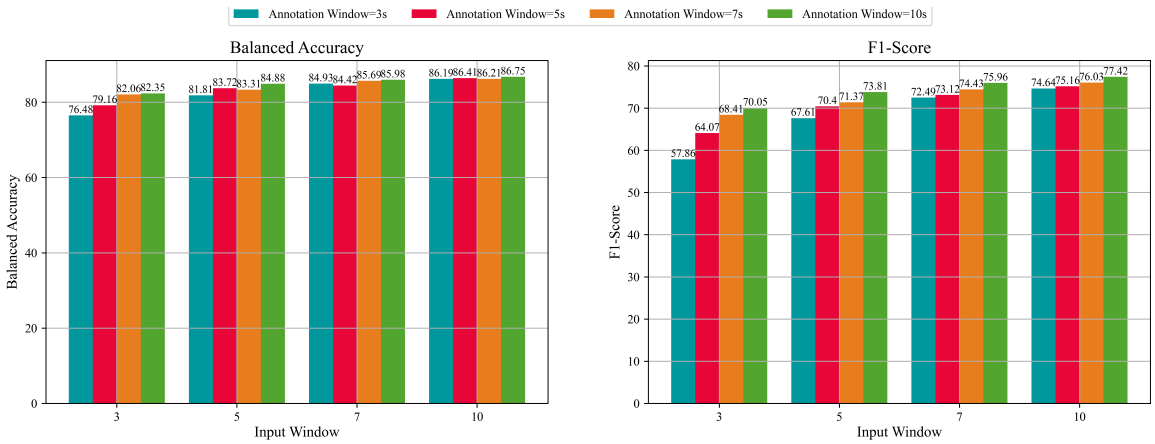


Fig. 16. Balanced accuracies and F1-scores for Bi-LSTM-CNN models with different annotations and input windows.

6 Discussion

6.1 Potential Applications of the TimelyTale Dataset

While our ablation study focused on predicting the demand timing for explanations, the multimodal composition of our dataset opens possibilities for extended analysis. The open nature of our dataset could enable further applications at both the passenger application level and city level when facilitated for wider data collection.

6.1.1 City-wide Modeling of Explanation Demands. Our research primarily focused on constructing multimodal datasets that include environmental, driving, and passenger-related contextual data. We conducted experiments under controlled route settings to gather this data. Given the widespread adoption of vehicle proprioception [61, 85, 148], smartphones [56, 61], and wrist-worn devices [56, 148], our approach could benefit from drawing on the methodologies and insights from previous research in the field of interruptibility. For example, considering the capability of OBD-II to transfer data via Bluetooth, our machine learning approach can be applied to use only OBD-II data in conjunction with a smartphone application for GPS and IMU data collection and explanation demand annotation for data collection at scale. This approach will enable large-scale data collection, such as crowdsourcing, to identify city-wide explanation needs or potentially other types of demands related to urban planning, including traffic, road type, and regulation design. It would aid in pinpointing critical areas for explanations and support the development of services to meet passengers' informational needs when designing services and facilities for which drivers and passengers may have specific types of demands.

6.1.2 In-vehicle Explanation Application: Combining Demand Timing Prediction with Explanation Generation. While our study primarily focused on identifying moments for delivering in-vehicle explanations related to a

vehicle's decision-making or driving behavior, its inclusion of exteroception data can be used to extract object- and environment-specific features surrounding the vehicle and relate them to the explanation demands in automated vehicles. Given that a significant portion of the explanation requests in our dataset relates to the vehicle's decision-making process and its environmental context, these demands are likely to be addressed by the downstream tasks of vision-language models, such as those involving GPT-4 for decision-making and reasoning in automated driving [23, 118, 150, 152]. The prediction of explanation demand timing can be combined with these explanation generation models to create a system for timely and contextually relevant explanations. This integration can be achieved through a cascading approach, where the demand timing prediction model determines the provision of explanations generated by visual-language models at the appropriate moments, or through an end-to-end model that jointly learns to predict the timing and generate the explanations.

Large language models for generating vehicle and traffic scenario explanations have emerged with the dual aim of describing traffic situations, navigation, and driving operations [166]. These models, especially those employing visual question-answering frameworks, are increasingly oriented towards explaining traffic scenarios to share an understanding of current traffic conditions between vehicles and passengers, often without accompanying causal justifications for specific actions. For example, Atakishiyev et al. [6] demonstrated the effectiveness of visual question-answering models in generating detailed explanations of traffic scenarios. Sima et al. [134] used these models for language-driven perception and end-to-end automated driving tasks. The effectiveness of these approaches in generating accurate explanations could be enhanced by integrating contextual passenger contexts regarding their needs and states, to tailor in-vehicle explanation services that meet actual user demands.

The multimodal composition of our dataset, which includes LiDAR, GPS, OBD-II, front-view stereo images, and real-time passenger explanation requests, offers exteroceptive and proprioceptive contextual information in which explanations are sought. It thus holds the potential for creating driving-adaptive explanations. For example, it can support GPS-specific contextual explanation demands, utilizing methodologies similar to those used in graph-based models for driving-purpose prediction studies [83]. Similarly, the exteroception, proprioception, and interoception data can be used in multimodal transformer models to generate explanations that are contextually pertinent and attuned to the vehicle's ego-motion behavior and proprioception-influenced decisions [23, 57], beyond traffic situation explanations.

6.1.3 Proof-of-Concept: City-Wide Geolocational Modeling of Textual Explanation Demands. We analyzed the textual explanations with regards to the positions where they were required to provide a proof-of-concept for two proposed applications: city-wide modeling and textual explanation. First, we clustered the GPS positions into ten groups using k-means clustering (Silhouette score = 0.61). For each geographical cluster (GeoCluster), we applied the TextRank algorithm [97] to identify the most important words and understand the explanation demands modeled from the raw text of required explanations. This approach balances the previously introduced coded labels with an objective method based on the original text label. The TextRank algorithm constructs a graph representation of the text, where nodes represent words or phrases, and edges represent their co-occurrence or semantic similarity. We plot TextRank importance scores for each word or phrase based on their connections to other words to identify the most salient or representative information in the GeoClusters.

In Figure 17, most GeoClusters shared common explanation demands related to the vehicle's starting, stopping, and slowing down behaviors. However, unique demands emerged in certain areas. For instance, participants frequently sought explanations for merging and entering highways. The transition between different road types, such as urban roads to highways or vice versa, at GeoClusters 4, 5, and 10, was also a trigger for explanation demands. Likewise, in GeoCluster 1 (Figure 17), corresponding to the departure and arrival points, there were specific inquiries about lane-changing behaviors undertaken by the vehicle to reach its destination.

The application of the TextRank algorithm to our dataset serves as a proof-of-concept, demonstrating the potential for scalability to larger datasets and more advanced natural language processing methods. TextRank's

graph-based approach captures the inherent structure and relationships within the text, making it adaptable to various text lengths and complexities. As the dataset grows in size, the method can be extended to complex language models. Furthermore, while each GeoCluster in our proof-of-concept covers large areas, which may not be uniform in terms of explanation demands, larger datasets will allow for the grouping of more condensed GeoClusters. This will enable an understanding of city-wide demand at specific locations, aiding traffic and urban planning, policymaking, and the development of in-vehicle services for geography-aware explanations.

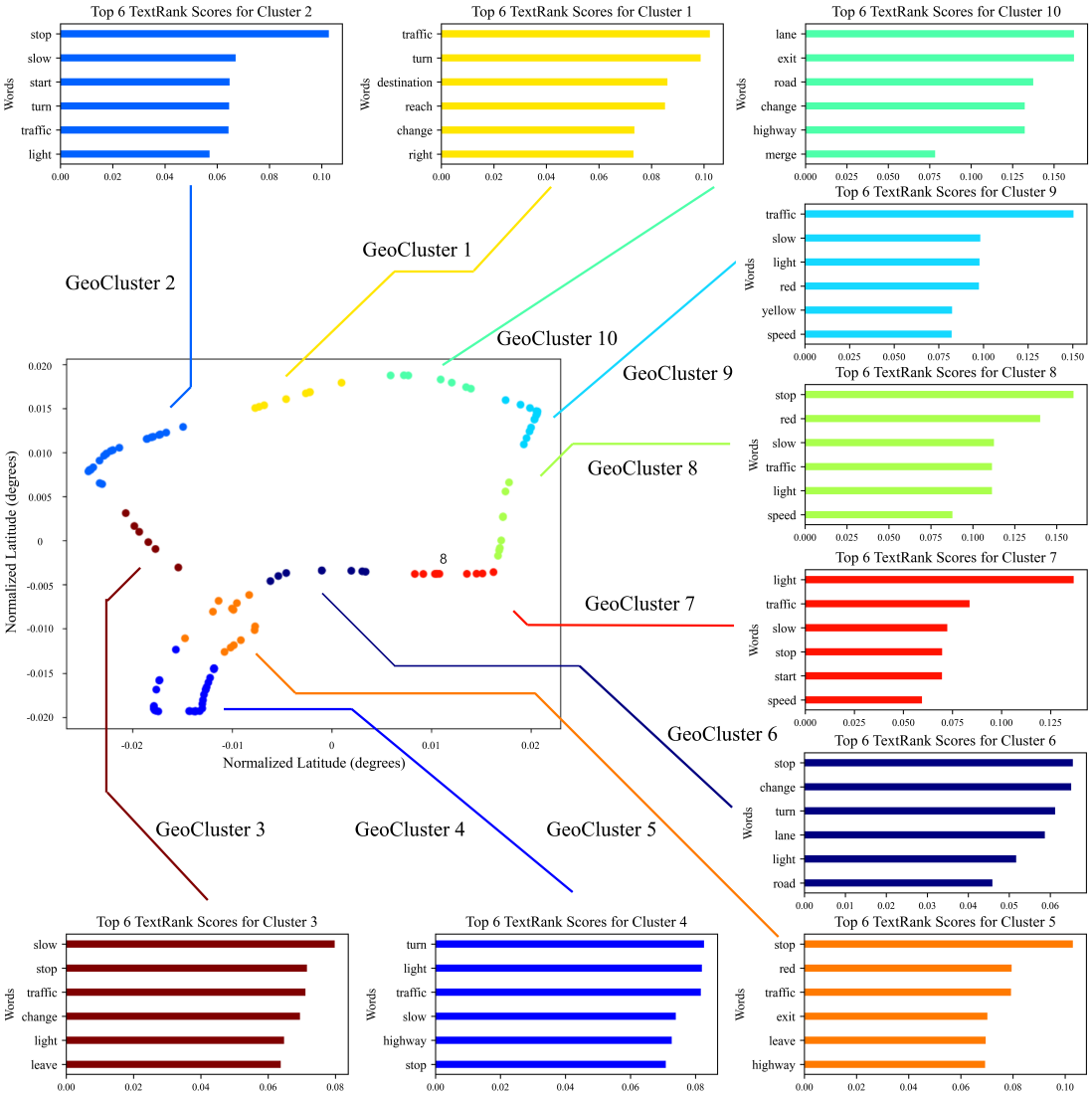


Fig. 17. TextRank analysis on the k-means clustering of each GPS-tagged explanation demand.

6.2 Explanation Requires Explaining Situations and Decisions Rather Than Focusing Solely on Reasons

Passenger reports in simulated driving environments emphasize the importance of receiving explanations for unexpected driving behaviors [145]. This need is also reflected in research on visual explanations, which demonstrates that passengers' trust increases when an automated vehicle visualizes its perception of the surrounding objects and environments [19, 21, 64]. Similarly, our results show that explanation demands in automated vehicles primarily focus on operational driving decisions, such as turns, lane changes, accelerations, starts, and stops, which are influenced by the surrounding objects and environments, rather than complex tactical driving strategies or long-term navigational goals.

Interestingly, about half of the requests (50.77%) for explanations did not seek the underlying reasons for driving behaviors; instead, passengers were interested in understanding operational decisions. The other half of the explanation demands asked for reasons, but no demand asked for the reason, only without explaining the vehicle's actions. The demanded reasoning-related explanations were mostly related to traffic lights and traffic conditions, explaining the vehicle's reaction to the changing road environment. This result partially explains the mixed findings from previous studies assessing the impact of providing action explanations and their reasoning on passenger experience [163]. However, it should be noted that our result was conducted under normal driving conditions without encountering safety-critical situations or engaging in prolonged driving that might require insights into tactical driving decisions, destination planning, or route awareness.

Most XAI models focus on explaining the rationale behind a vehicle's decision-making process [67]. While these provide relevant information to the stakeholders who seek insights into a vehicle's reasoning processes, such as developers, AI experts, insurers, and policymakers, the impact of these models on in-vehicle passengers is not as well-defined [115]. Building on the groundwork laid by in-vehicle explanation studies focused on describing the vehicle's actions and justifications that focused on safety-critical environments, our study contributes to the in-vehicle interruptibility and explainability by exploring passengers' informational needs and sources of explanation demands within routine automated driving scenarios.

6.3 Proprioceptive and Physiological Responses as Reliable Indicators for Passenger Explanation Demand Timing

Our machine learning analysis included IMU, OBD-II, and physiological responses as features capturing passengers' demand for explanations. The results also showed that a vehicle's proprioceptive sensors alone can serve as indicators of moments requiring explanation, though it may be improved with passenger sensing data. This suggests the possibility of our method being integrated with vehicles without additional hardware use or obtrusive sensing. The ablation study demonstrated that adding relevant features can enhance the accuracy of timing classification. For example, introducing additional sensors to monitor the passenger's physiological state increased the accuracy. Similarly, while not included in the current model, incorporating additional features such as exteroceptive sensing data, as used in Liu et al. [85] and Wu et al. [148], could further improve accuracy.

However, it is important to consider the trade-offs between the validity of the settings in terms of the likelihood of use and integration in actual scenarios, the obtrusiveness of the setup, and the richness of the data. As part of our dataset initiative, we collected extensive data while maintaining minimal obtrusiveness to ensure a naturalistic experience for study participants. Nevertheless, some sensors pose additional installation, and others require the passenger to wear them, which could limit their usage. When designing models for real-world applications based on our dataset, careful consideration should be given to their integration with the daily environment.

6.4 Validity of the WoZ Driving Experience

The WoZ methodology is used in automated vehicle research for its safety and regulatory advantages. However, there are concerns about how authentic the passenger experience it provides is compared with that of real automated vehicles. Schneider et al. [129], in their WoZ study conducted on actual roads, found that knowing about the wizard driver did not significantly impact the perceived authenticity of the automated driving simulation. Despite this finding, many researchers, such as Detjen et al. [29] and Andrea et al. [103], have emphasized the importance of maintaining the illusion of an automated vehicle and have raised concerns about the potential effect of discovering a wizard driver on passengers' perceptions and behavior, suggesting that participants who do not believe the simulation should be excluded to ensure the validity of the results.

In our study, the vehicle was fitted with sensors similar to those in automated vehicles, which enhanced the impression of automated operation. No participant reported noticing the driver at the end of the experiment, and thus none were excluded from the experiment report. This suggests that the WoZ implementation was effective in maintaining the illusion of an automated vehicle. However, it is important to acknowledge that passengers' awareness of the experimenter's presence might have given them a sense of safety, knowing that the wizard might be ready to take over, though they were informed and believed the experimenter did not drive. This awareness could influence their behavior and reactions compared to what they would experience in actual automated vehicles.

Additionally, the partition used for the experiments may have limited passengers' peripheral vision and situation awareness. As we aimed to elicit passenger demand for explanations under automated driving scenarios, we prioritized a realistic visual and kinesthetic driving environment. When testing specific services based on our study, driving simulators can also be used to provide more natural interface experiences without the need for partitions to hide the driver.

6.5 Data Limitations

6.5.1 Focused User Composition and Driving Scenarios. Our dataset could be further enriched by including a more diverse range of user familiarity, driving scenarios, and a larger number of participants. Our study specifically focused on users unfamiliar with automated vehicles, acknowledging the public anxiety and reluctance to accept automated vehicles [51], which could be alleviated by transparency and explainability [64]. Although some users had experience with SAE level 2 vehicles with advanced driver assistance systems (e.g., Tesla's Autopilot [151]), the subjects of our study had no prior experience with highly automated vehicles.

With the upcoming prevalence of automated vehicles, our approach could be expanded to include scenarios involving experienced passengers. The demand for explanations may be lower when the driving automation is perceived as flawless, as we typically don't expect extensive explanations when riding in a vehicle driven by a human driver. Nevertheless, given that explanations can alleviate passengers' anxiety when vehicles behave in unexpected ways [145], the value of explanations can also lie in the alignment and sharability of in-situ decisions rather than building and calibrating trust in the automation as a whole. Further research is needed to determine the persistence of these demands with experienced passengers. Additionally, while the dataset includes comparable hours of driving data, it involved a relatively limited number of participants ($N = 29$), suggesting the potential to better represent individual differences by including a larger number of participants with diverse age, gender, and cultural backgrounds.

When automated vehicles become more pervasive, new interaction scenarios may arise, such as multiple user situations or shared mobility [91]. For example, the impact of social dynamics on demand in public shared mobility [14] and the specific information needs in shared vehicle scenarios [37] could be further explored. Also, given that road types can impact passengers' perceptions of automated vehicles [39], expanding the study to include rural road environments would enhance the dataset's representativeness and usability.

6.5.2 Potential Overestimation of Explanation Demands. We acknowledge that the number of explanation demands reported in this study could be overestimated owing to the research methods employed. Although we did not impose a specific number of explanations and explicitly informed participants that they could report no explanations if they needed none during the drive, our experimental settings for assessing explanation demands might have introduced certain biases. The Hawthorne effect [94], where people become more aware of and attentive to the vehicle's actions when they know they are being studied, could have been intensified by the priming effect of describing explanations. This might have led participants to think more about the concept and its importance [98], potentially increasing their perceived need for explanations. This effect could have been further amplified by the examples of in-vehicle explanations for automated vehicles provided to aid participants' understanding of the study. These examples may have led participants to believe they needed more explanations than they otherwise would have, potentially limiting or biasing their perceptions of the types of explanations available and resulting in oversampling of certain explanation types.

Moreover, demand characteristics [110] or social desirability [109] bias could have influenced participants to align their behavior with the researcher's expectations [26] or general beliefs. We attempted to minimize this effect by emphasizing that expressing the absence of demand (by not reporting any demand) is just as important as expressing demand. However, the study's experimental nature might have led to an overestimation of explanation demands. Furthermore, the framing effect [139] of presenting our study as an investigation of in-vehicle explanation demands might have emphasized the importance of explanations and encouraged participants to express a greater need for them. This framing could have inadvertently influenced participants' responses and contributed to an overestimation of explanation demands.

While we took steps to mitigate these biases, such as providing clear instructions and avoiding leading questions, it is important to consider these limitations when interpreting the results of this study. Future research could employ alternative methods, such as long-term studies or crowdsourcing, to further investigate the demand for explanations in real-world settings and validate the findings of this study.

6.5.3 Variability in Explanation Demand Report and Regularization Challenges. While our dataset aimed to standardize the collection of explanation demands through a fixed protocol and detailed instructions, the annotation of individuals' inner intentions inherently relied on self-reporting. This approach introduces variability in two aspects: first, the demand for explanations varies among individuals, and second, the threshold at which one reports a demand for an explanation differs from person to person. While this approach could serve as formative research, from a machine learning perspective, it could complicate the solution by inducing generalization challenges that result from differences in experience and reporting among individuals, and even within the same participant across different explanation cases.

This challenge produces data that is difficult to regularize and prone to overfitting due to its characteristics. In such cases, the individual and session-specific differences among explanation cases make it hard to generalize, causing trained models to work less effectively on datasets not used for training. Therefore, our dataset approach should be complemented with more standardizable methods for scalable data collection at large.

The data collection method could be adapted to include digital experience sampling methods [7], or a crowdsourcing approach with driving videos, simplifying the process for participants to express their explanation demands. However, these methods may induce actions not typically present in actual automated vehicles, reducing ecological validity. Alternatively, monitoring self-interruptions [43] may be another viable approach, depending on the objectives of the explanation system (e.g., minimizing the disruption of NDRTs).

Implementing such modifications would standardize the annotation of explanation demands, yielding data more suitable for machine learning predictions while better reflecting the varied needs and behaviors of passengers in automated vehicles. These approaches will provide a good dataset for baseline models, while experiments involving passengers could serve to fine-tune individual models.

7 Conclusion

In this study, we collected data on passengers' demand for explanations in automated driving environments while gathering environmental driving-related and passenger response (interoceptive) measurements, which can be used for contextual understanding. We used 3D LiDAR, stereo cameras, GPS, OBD-II, and IMUs for exteroceptive and proprioceptive data. Interoceptive data pertaining to the passenger's state were captured using a depth camera, LiDAR camera, E4 wristband, thermal imaging, and seat pressure sensors. The results of our data identified both the timing and frequency of passengers' demands for explanations, as well as the specific in-situ explanations that passengers demand in driving situations. Our methodology, equipped with GPS, can be applied to encompass city-wide analysis of the need for explanations. Despite some dataset imbalance and regularization issues, our preliminary analysis indicated its potential utility in determining the passenger's demand timing for in-vehicle explanations. By integrating exteroceptive, proprioceptive, and interoceptive sensor data and multimodal vision-language models, our dataset could be used for the end-to-end generation of textual explanation content that is pertinent to environmental, driving-related, and passenger-specific contexts.

Acknowledgments

This work was in part supported by the GIST-MIT Research Collaboration grant funded by the GIST in 2024. This work was in part supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (RS-2024-00343397). This work was in part supported by the Artificial Intelligence Graduate School Program (GIST) of the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2019-0-01842).

References

[1] Yomna Abdelrahman, Anam Ahmad Khan, Joshua Newn, Eduardo Velloso, Sherine Ashraf Safwat, James Bailey, Andreas Bulling, Frank Vetere, and Albrecht Schmidt. 2019. Classifying Attention Types with Thermal Imaging and Eye Tracking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 69 (sep 2019), 27 pages. https://doi.org/10.1145/3351227

[2] Yomna Abdelrahman, Eduardo Velloso, Tilman Dingler, Albrecht Schmidt, and Frank Vetere. 2017. Cognitive Heat: Exploring the Usage of Thermal Imaging to Unobtrusively Estimate Cognitive Load. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 33 (sep 2017), 20 pages. https://doi.org/10.1145/3130898

[3] Madina Abdrakhmanova, Askat Kuzdeuov, Sheikh Jarju, Yerbolat Khassanov, Michael Lewis, and Huseyin Atakan Varol. 2021. SpeakingFaces: A Large-Scale Multimodal Dataset of Voice Commands with Visual and Thermal Video Streams. *Sensors* 21, 10 (2021). https://doi.org/10.3390/s21103465

[4] Wataru Akahori, Asuka Miyake, Hiroaki Sugiyama, Masahiro Watanabe, and Hiroya Minami. 2019. Paired Conversational Agents for Easy-to-Understand Instruction. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312794

[5] Christoph Draxler Anna-Maria Meck and Thurid Vogt. 2023. How May I Interrupt? Linguistic-Driven Design Guidelines for Proactive in-Car Voice Assistants. *International Journal of Human–Computer Interaction* 0, 0 (2023), 1–15. https://doi.org/10.1080/10447318.2023.2266251 arXiv:https://doi.org/10.1080/10447318.2023.2266251

[6] Shahin Atakishiyev, Mohammad Salameh, Housam Babiker, and Randy Goebel. 2023. Explaining Autonomous Driving Actions with Visual Question Answering. arXiv:2307.10408 [cs.CV]

[7] Aya Ataya, Won Kim, Ahmed Elsharkawy, and SeungJun Kim. 2020. Gaze-Head Input: Examining Potential Interaction with Immediate Experience Sampling in an Autonomous Vehicle. *Applied Sciences* 10, 24 (2020). https://doi.org/10.3390/app10249011

[8] Aya Ataya, Won Kim, Ahmed Elsharkawy, and SeungJun Kim. 2021. How to interact with a fully autonomous vehicle: Naturalistic ways for drivers to intervene in the vehicle system while performing non-driving related tasks. *Sensors* 21, 6 (March 2021), 2206. https://doi.org/10.3390/s21062206

[9] Sonia Baltodano, Srinath Sibi, Nikolas Martelaro, Nikhil Gowda, and Wendy Ju. 2015. The RRADS Platform: A Real Road Autonomous Driving Simulator. In *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Nottingham, United Kingdom) *(AutomotiveUI '15)*. Association for Computing Machinery, New York, NY, USA, 281–288. https://doi.org/10.1145/2799250.2799288

[10] Sonia Baltodano, Srinath Sibi, Nikolas Martelaro, Nikhil Gowda, and Wendy Ju. 2015. The RRADS Platform: A Real Road Autonomous Driving Simulator. In *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*

(Nottingham, United Kingdom) *(AutomotiveUI '15)*. Association for Computing Machinery, New York, NY, USA, 281–288. https://doi.org/10.1145/2799250.2799288

[11] Mathias Benedek and Christian Kaernbach. 2010. A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods* 190, 1 (2010), 80–91. https://doi.org/10.1016/j.jneumeth.2010.04.028

[12] Mathias Benedek and Christian Kaernbach. 2010. Decomposition of skin conductance data by means of nonnegative deconvolution. *Psychophysiology* (2010). https://doi.org/10.1111/j.1469-8986.2009.00972.x

[13] Klaus Bengler, Kamil Omozik, and Andrea Isabell Müller. 2019. The Renaissance of Wizard of Oz (WoOz)–Using the WoOz methodology to prototype automated vehicles. *Proceedings of the Human Factors and Ergonomics Society Europe* (2019), 63–72.

[14] Alexandra Bremers, Natalie Friedman, Sam Lee, Tong Wu, Eric Laurier, Malte Jung, Jorge Ortiz, and Wendy Ju. 2023. (Social) Trouble on the Road: Understanding and Addressing Social Discomfort in Shared Car Trips. arXiv:2311.04456 [cs.HC]

[15] Marine Capallera, Leonardo Angelini, Quentin Meteier, Omar Abou Khaled, and Elena Mugellini. 2023. Human-Vehicle Interaction to Support Driver's Situation Awareness in Automated Vehicles: A Systematic Review. *IEEE Transactions on Intelligent Vehicles* 8, 3 (2023), 2551–2567. https://doi.org/10.1109/TIV.2022.3200826

[16] Stephen M. Casner, Edwin L. Hutchins, and Don Norman. 2016. The challenges of partially automated driving. *Commun. ACM* 59, 5 (apr 2016), 70–77. https://doi.org/10.1145/2830565

[17] Kuan-Wen Chen, Yung-Ju Chang, and Liwei Chan. 2022. Predicting Opportune Moments to Deliver Notifications in Virtual Reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (, New Orleans, LA, USA,) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 186, 18 pages. https://doi.org/10.1145/3491102.3517529

[18] Jong Kyu Choi and Yong Gu Ji. 2015. Investigating the importance of trust on adopting an autonomous vehicle. . *International Journal of Human-Computer Interaction* 31, 10 (2015), 692–702. https://doi.org/10.1080/10447318.2015.1070549

[19] Mark Colley, Benjamin Eder, Jan Ole Rixen, and Enrico Rukzio. 2021. Effects of Semantic Segmentation Visualization on Trust, Situation Awareness, and Cognitive Load in Highly Automated Vehicles. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 155, 11 pages. https://doi.org/10.1145/3411764.3445351

[20] Mark Colley, Svenja Krauss, Mirjam Lanzer, and Enrico Rukzio. 2021. How Should Automated Vehicles Communicate Critical Situations? A Comparative Analysis of Visualization Concepts. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3, Article 94 (sep 2021), 23 pages. https://doi.org/10.1145/3478111

[21] Mark Colley, Max Rädler, Jonas Glimmann, and Enrico Rukzio. 2022. Effects of Scene Detection, Scene Prediction, and Maneuver Planning Visualizations on Trust, Situation Awareness, and Cognitive Load in Highly Automated Vehicles. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 49 (jul 2022), 21 pages. https://doi.org/10.1145/3534609

[22] Mark Colley, Oliver Speidel, Jan Strohbeck, Jan Ole Rixen, Jan Henry Belz, and Enrico Rukzio. 2024. Effects of Uncertain Trajectory Prediction Visualization in Highly Automated Vehicles on Trust, Situation Awareness, and Cognitive Load. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 4, Article 153 (jan 2024), 23 pages. https://doi.org/10.1145/3631408

[23] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. 2024. Drive As You Speak: Enabling Human-Like Interaction With Large Language Models in Autonomous Vehicles. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. 902–909.

[24] N. Dahlbäck, A. Jönsson, and L. Ahrenberg. 1993. Wizard of Oz Studies — Why and How. *Know.-Based Syst.* 6, 4 (dec 1993), 258–266. https://doi.org/10.1016/0950-7051(93)90017-N

[25] Yannick Daviaux, Emilien Bonhomme, Hans Ivers, Étienne de Sevin, Jean-Arthur Micoulaud-Franchi, Stéphanie Bioulac, Charles M. Morin, Pierre Philip, and Ellemarije Altena. 2020. Event-Related Electrodermal Response to Stress: Results From a Realistic Driving Simulator Scenario. *Human Factors* 62, 1 (2020), 138–151. https://doi.org/10.1177/0018720819842779 arXiv:https://doi.org/10.1177/0018720819842779 PMID: 31050918.

[26] Nicola Dell, Vidya Vaidyanathan, Indrani Medhi, Edward Cutrell, and William Thies. 2012. "Yours is better!": participant response bias in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1321–1330. https://doi.org/10.1145/2207676.2208589

[27] Joseph DelPreto, Chao Liu, Yiyue Luo, Michael Foshey, Yunzhu Li, Antonio Torralba, Wojciech Matusik, and Daniela Rus. 2022. ActionSense: A Multimodal Dataset and Recording Framework for Human Activities Using Wearable Sensors in a Kitchen Environment. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 13800–13813. https://proceedings.neurips.cc/paper_files/paper/2022/file/5985e81d65605827ac35401999aea22a-Paper-Datasets_and_Benchmarks.pdf

[28] Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie-Francine Moens. 2019. Talk2Car: Taking Control of Your Self-Driving Car. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/d19-1215

[29] Henrik Detjen, Bastian Pfleging, and Stefan Schneegass. 2020. A Wizard of Oz Field Study to Understand Non-Driving-Related Activities, Trust, and Acceptance of Automated Vehicles. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Virtual Event, DC, USA) *(AutomotiveUI '20)*. Association for Computing Machinery, New York, NY, USA, 19–29. https://doi.org/10.1145/3409120.3410662

[30] Carolina Diaz-Piedra, Emilo Gomez-Milan, and Leandro L. Di Stasi. 2019. Nasal skin temperature reveals changes in arousal levels due to time on task: An experimental thermal infrared imaging study. *Applied Ergonomics* 81 (2019), 102870. https://doi.org/10.1016/j.apergo.2019.06.001

[31] Bart van Arem Dimitris Milakis and Bert van Wee. 2017. Policy and society related implications of automated driving: A review of literature and directions for future research. *Journal of Intelligent Transportation Systems* 21, 4 (2017), 324–348. https://doi.org/10.1080/15472450.2017.1291351 arXiv:https://doi.org/10.1080/15472450.2017.1291351

[32] Na Du, Jacob Haspiel, Qiaoning Zhang, Dawn Tilbury, Anuj K. Pradhan, X. Jessie Yang, and Lionel P. Robert. 2019. Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload. *Transportation Research Part C: Emerging Technologies* 104 (2019), 428–442. https://doi.org/10.1016/j.trc.2019.05.025

[33] Patrick Ebel, Pavlo Bazilinskyy, Angel Hsing-Chi Hwang, Wendy Ju, Hauke Sandhaus, Aravinda Ramakrishnan Srinivasan, Qian Yang, and Philipp Wintersberger. 2023. Breaking Barriers: Workshop on Open Data Practices in AutoUI Research. In *Adjunct Proceedings of the 15th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Ingolstadt, Germany) *(AutomotiveUI '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 227–230. https://doi.org/10.1145/3581961.3609835

[34] Fredrick Ekman, Mikael Johansson, and Jana Sochor. 2016. To See or Not to See: The Effect of Object Recognition on Users' Trust in "Automated Vehicles". In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction* (Gothenburg, Sweden) *(NordiCHI '16)*. Association for Computing Machinery, New York, NY, USA, Article 42, 4 pages. https://doi.org/10.1145/2971485.2971551

[35] Mica R. Endsley. 2019. Situation Awareness in Future Autonomous Vehicles: Beware of the Unexpected. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*, Sebastiano Bagnara, Riccardo Tartaglia, Sara Albolino, Thomas Alexander, and Yushi Fujita (Eds.). Springer International Publishing, Cham, 303–309.

[36] Andrea Ferrario and Michele Loi. 2022. How Explainability Contributes to Trust in AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (, Seoul, Republic of Korea,) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1457–1466. https://doi.org/10.1145/3531146.3533202

[37] Lukas A. Flohr, Martina Schuß, Dieter P. Wallach, Antonio Krüger, and Andreas Riener. 2024. Designing for passengers' information needs on fellow travelers: A comparison of day and night rides in shared automated vehicles. *Applied Ergonomics* 116 (2024), 104198. https://doi.org/10.1016/j.apergo.2023.104198

[38] Lukas A. Flohr, Joseph Sebastian Valiyaveettil, Antonio Krüger, and Dieter P. Wallach. 2023. Prototyping Autonomous Vehicle Windshields with AR and Real-Time Object Detection Visualization: An On-Road Wizard-of-Oz Study. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (, Pittsburgh, PA, USA,) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 2123–2137. https://doi.org/10.1145/3563657.3596051

[39] Anna-Katharina Frison, Philipp Wintersberger, Tianjia Liu, and Andreas Riener. 2019. Why do you like to drive automated? a context-dependent analysis of highly automated driving to elaborate requirements for intelligent user interfaces. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 528–537. https://doi.org/10.1145/3301275.3302331

[40] Konrad Gadzicki, Razieh Khamsehashari, and Christoph Zetzsche. 2020. Early vs Late Fusion in Multimodal Convolutional Neural Networks. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*. 1–6. https://doi.org/10.23919/FUSION45008.2020.9190246

[41] Apeksha Gaonkar, Yogya Chukkapalli, P. Jahnavi Raman, Sahana Srikanth, and Sanjeev Gurugopinath. 2021. A Comprehensive Survey on Multimodal Data Representation and Information Fusion Algorithms. In *2021 International Conference on Intelligent Technologies (CONIT)*. 1–8. https://doi.org/10.1109/CONIT51480.2021.9498415

[42] Maurizio Garbarino, Matteo Lai, Dan Bender, Rosalind W. Picard, and Simone Tognetti. 2014. Empatica E3 — A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In *2014 4th International Conference on Wireless Mobile Communication and Healthcare - Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH)*. 39–42. https://doi.org/10.1109/MOBIHEALTH.2014.7015904

[43] Michael A. Gerber, Ronald Schroeter, Li Xiaomeng, and Mohammed Elhenawy. 2020. Self-Interruptions of Non-Driving Related Tasks in Automated Vehicles: Mobile vs Head-Up Display. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (, Honolulu, HI, USA,) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3313831.3376751

[44] Stuart T Godley, Thomas J Triggs, and Brian N Fildes. 2002. Driving simulator validation for speed research. *Accident Analysis & Prevention* 34, 5 (2002), 589–600. https://doi.org/10.1016/S0001-4575(01)00056-2

[45] Don Gotterbarn, Amy Bruckman, Catherine Flick, Keith Miller, and Marty J Wolf. 2017. ACM code of ethics: a guide for positive action. , 121–128 pages.

[46] Julia Graefe, Selma Paden, Doreen Engelhardt, and Klaus Bengler. 2022. Human Centered Explainability for Intelligent Vehicles – A User Study. In *Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Seoul, Republic of Korea) *(AutomotiveUI '22)*. Association for Computing Machinery, New York, NY, USA, 297–306. https://doi.org/10.1145/3543174.3546846

[47] Alex Graves and Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks* (2012), 37–45.

[48] Taehyun Ha, Sangyeon Kim, Donghak Seo, and Sangwon Lee. 2020. Effects of explanation types and perceived risk on trust in autonomous vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour* 73 (2020), 271–280. https://doi.org/10.1016/j.trf.2020.06.021

[49] Jacob Haspiel, Na Du, Jill Meyerson, Lionel P. Robert Jr., Dawn Tilbury, X. Jessie Yang, and Anuj K. Pradhan. 2018. Explanations and Expectations: Trust Building in Automated Vehicles. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago, IL, USA) *(HRI '18)*. Association for Computing Machinery, New York, NY, USA, 119–120. https://doi.org/10.1145/3173386.3177057

[50] Hendrick, Surfa Yondri, Rahmat Hidayat, Albar Albar, Hanifa Fitri, and Ivan Finiel Bagariang. 2020. Face Detection Model for Thermal Images. In *2020 International Conference on Applied Science and Technology (iCAST)*. 166–169. https://doi.org/10.1109/iCAST51016.2020.9557718

[51] Charlie Hewitt, Ioannis Politis, Theocharis Amanatidis, and Advait Sarkar. 2019. Assessing Public Perception of Self-Driving Cars: The Autonomous Vehicle Acceptance Model. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 518–527. https://doi.org/10.1145/3301275.3302268

[52] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors* 57, 3 (2015), 407–434. https://doi.org/10.1177/0018720814547570 arXiv:https://doi.org/10.1177/0018720814547570 PMID: 25875432.

[53] Pascal Jansen, Mark Colley, and Enrico Rukzio. 2022. A Design Space for Human Sensor and Actuator Focused In-Vehicle Interaction Based on a Systematic Literature Review. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 56 (jul 2022), 51 pages. https://doi.org/10.1145/3534617

[54] Marie Jaussein, Lucie Lévêque, Jonathan Deniel, Thierry Bellet, Hélène Tattegrain, and Claude Marin-Lamellet. 2021. How Do Non-driving-related Tasks Affect Engagement Under Highly Automated Driving Situations? A Literature Review. *Frontiers in Future Transportation* 2 (2021). https://doi.org/10.3389/ffutr.2021.687602

[55] Shuo Jia, Fei Hui, Shining Li, Xiangmo Zhao, and Asad J Khattak. 2020. Long short-term memory and convolutional neural network for abnormal driving behaviour recognition. *IET Intelligent Transport Systems* 14, 5 (2020), 306–312.

[56] Landu Jiang, Xinye Lin, Xue Liu, Chongguang Bi, and Guoliang Xing. 2018. SafeDrive: Detecting Distracted Driving Behaviors Using Wrist-Worn Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 144 (jan 2018), 22 pages. https://doi.org/10.1145/3161179

[57] Bu Jin, Xinyu Liu, Yupeng Zheng, Pengfei Li, Hao Zhao, Tong Zhang, Yuhang Zheng, Guyue Zhou, and Jingjing Liu. 2023. ADAPT: Action-aware Driving Caption Transformer. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. 7554–7561. https://doi.org/10.1109/ICRA48891.2023.10160326

[58] Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Shun Chen. 2018. LSTM Fully Convolutional Networks for Time Series Classification. *IEEE Access* 6 (2018), 1662–1669. https://doi.org/10.1109/ACCESS.2017.2779939

[59] Siddartha Khastgir, Stewart Birrell, Gunwant Dhadyalla, and Paul Jennings. 2018. Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles. *Transportation Research Part C: Emerging Technologies* 96 (2018), 290–303. https://doi.org/10.1016/j.trc.2018.07.001

[60] Dargahi Nobari Khazar, Alexander Hugenroth, and Torsten Bertram. 2022. Position Classification and In-Vehicle Activity Detection Using Seat-Pressure-Sensor in Automated Driving. In *AmE 2022 - Automotive meets Electronics; 13. GMM-Symposium*. 1–6.

[61] Auk Kim, Woohyeok Choi, Jungmi Park, Kyeyoon Kim, and Uichin Lee. 2018. Interrupting Drivers for Interactions: Predicting Opportune Moments for In-Vehicle Proactive Auditory-Verbal Tasks. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 175 (dec 2018), 28 pages. https://doi.org/10.1145/3287053

[62] Auk Kim, Jung-Mi Park, and Uichin Lee. 2020. Interruptibility for In-Vehicle Multitasking: Influence of Voice Task Demands and Adaptive Behaviors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 14 (mar 2020), 22 pages. https://doi.org/10.1145/3381009

[63] Gwangbin Kim, Jieun Lee, Dohyeon Yeo, Eunsol An, and SeungJun Kim. 2023. Physiological Indices to Predict Driver Situation Awareness in VR. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing* (Cancun, Quintana Roo, Mexico) *(UbiComp/ISWC '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 40–45. https://doi.org/10.1145/3594739.3610687

[64] Gwangbin Kim, Dohyeon Yeo, Taewoo Jo, Daniela Rus, and SeungJun Kim. 2023. What and When to Explain? On-road Evaluation of Explanations in Highly Automated Vehicles. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 3, Article 104 (sep 2023), 26 pages. https://doi.org/10.1145/3610886

[65] Jinkyu Kim and John Canny. 2017. Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2961–2969. https://doi.org/10.1109/ICCV.2017.320

[66] Jungsook Kim, Woojin Kim, Hyun-Suk Kim, Seung-Jun Lee, Oh-Cheon Kwon, and Daesub Yoon. 2022. A novel study on subjective driver readiness in terms of non-driving related tasks and take-over performance. *ICT Express* 8, 1 (2022), 91–96. https://doi.org/10.1016/j.icte.2021.04.008

[67] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. 2018. Textual Explanations for Self-Driving Vehicles. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II* (Munich, Germany). Springer-Verlag, Berlin, Heidelberg, 577–593. https://doi.org/10.1007/978-3-030-01216-8_35

[68] Keunwoo Kim, Minjung Park, and Youn-kyung Lim. 2021. Guiding Preferred Driving Style Using Voice in Autonomous Vehicles: An On-Road Wizard-of-Oz Study. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 352–364. https://doi.org/10.1145/3461778.3462056

[69] SeungJun Kim, Jaemin Chun, and Anind K. Dey. 2015. Sensors Know When to Interrupt You in the Car: Detecting Driver Interruptibility Through Monitoring of Peripheral Interactions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 487–496. https://doi.org/10.1145/2702123.2702409

[70] Seungju Kim, Jungseok Oh, Minwoo Seong, Eunki Jeon, Yeon-Kug Moon, and Seungjun Kim. 2023. Assessing the Impact of AR HUDs and Risk Level on User Experience in Self-Driving Cars: Results from a Realistic Driving Simulation. *Applied Sciences* 13, 8 (2023). https://doi.org/10.3390/app13084952

[71] Won Kim, Eunki Jeon, Gwangbin Kim, Dohyeon Yeo, and SeungJun Kim. 2022. Take-Over Requests after Waking in Autonomous Vehicles. *Applied Sciences* 12, 3 (2022). https://doi.org/10.3390/app12031438

[72] Kevin Koch, Varun Mishra, Shu Liu, Thomas Berger, Elgar Fleisch, David Kotz, and Felix Wortmann. 2021. When Do Drivers Interact with In-Vehicle Well-being Interventions? An Exploratory Analysis of a Longitudinal Study on Public Roads. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 19 (mar 2021), 30 pages. https://doi.org/10.1145/3448116

[73] Thomas Köhn, Matthias Gottlieb, Michael Schermann, and Helmut Krcmar. 2019. Improving Take-over Quality in Automated Driving by Interrupting Non-Driving Tasks. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 510–517. https://doi.org/10.1145/3301275.3302323

[74] Jeamin Koo, Jungsuk Kwac, Wendy Ju, Martin Steinert, Larry Leifer, and Clifford Nass. 2014. Why did my car just do that? explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *International Journal on Interactive Design and Manufacturing (IJIDeM)* 9, 4 (2014), 269–275. https://doi.org/10.1007/s12008-014-0227-2

[75] Philip Koopman and Michael Wagner. 2018. *Toward a framework for highly automated vehicle safety validation.* Technical Report. SAE Technical Paper.

[76] Johannes Kraus, David Scholz, Dina Stiegemeier, and Martin Baumann. 2020. The More You Know: Trust Dynamics and Calibration in Highly Automated Driving and the Effects of Take-Overs, System Malfunction, and System Transparency. *Human Factors* 62, 5 (2020), 718–736. https://doi.org/10.1177/0018720819853686 arXiv:https://doi.org/10.1177/0018720819853686 PMID: 31233695.

[77] Marc Alexander Kühn, Daniel Omeiza, and Lars Kunze. 2023. Textual Explanations for Automated Commentary Driving. *arXiv preprint arXiv:2304.08178* (2023).

[78] Askat Kuzdeuov, Darina Koishigarina, Dana Aubakirova, Saniya Abushakimova, and Huseyin Atakan Varol. 2022. SF-TL54: A Thermal Facial Landmark Dataset with Visual Pairs. In *2022 IEEE/SICE International Symposium on System Integration (SII)*. 748–753. https://doi.org/10.1109/SII52469.2022.9708901

[79] M. Kyriakidis, R. Happee, and J.C.F. de Winter. 2015. Public opinion on automated driving: Results of an international questionnaire among 5000 respondents. *Transportation Research Part F: Traffic Psychology and Behaviour* 32 (2015), 127–140. https://doi.org/10.1016/j.trf.2015.04.014

[80] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392 arXiv:https://doi.org/10.1518/hfes.46.1.50_30392 PMID: 15151155.

[81] Mengyao Li, Brittany E. Holthausen, Rachel E. Stuck, and Bruce N. Walker. 2019. No Risk No Trust: Investigating Perceived Risk in Highly Automated Driving. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Utrecht, Netherlands) *(AutomotiveUI '19)*. Association for Computing Machinery, New York, NY, USA, 177–185. https://doi.org/10.1145/3342197.3344525

[82] Penghui Li, Yibing Li, Yao Yao, Changxu Wu, Bingbing Nie, and Shengbo Eben Li. 2022. Sensitivity of Electrodermal Activity Features for Driver Arousal Measurement in Cognitive Load: The Application in Automated Driving Systems. *IEEE Transactions on Intelligent Transportation Systems* 23, 9 (2022), 14954–14967. https://doi.org/10.1109/TITS.2021.3135266

[83] Chengwu Liao, Chao Chen, Suiming Guo, Zhu Wang, Yaxiao Liu, Ke Xu, and Daqing Zhang. 2022. Wheels Know Why You Travel: Predicting Trip Purpose via a Dual-Attention Graph Embedding Network. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 1, Article 22 (mar 2022), 22 pages. https://doi.org/10.1145/3517239

[84] Todd Litman. 2020. Autonomous vehicle implementation predictions: Implications for transport planning. (2020).

[85] Shu Liu, Kevin Koch, Zimu Zhou, Simon Föll, Xiaoxi He, Tina Menke, Elgar Fleisch, and Felix Wortmann. 2021. The Empathetic Car: Exploring Emotion Inference via Driver Behaviour and Traffic Context. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3, Article 117 (sep 2021), 34 pages. https://doi.org/10.1145/3478078

[86] Monika Lohani, Brennan R. Payne, and David L. Strayer. 2019. A Review of Psychophysiological Measures to Assess Cognitive States in Real-World Driving. *Frontiers in Human Neuroscience* 13 (2019). https://doi.org/10.3389/fnhum.2019.00057

[87] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Building Perception Pipelines. arXiv:1906.08172 [cs.DC]

[88] Yiyue Luo, Yunzhu Li, Pratyusha Sharma, Wan Shou, Kui Wu, Michael Foshey, Beichen Li, Tomás Palacios, Antonio Torralba, and Wojciech Matusik. 2021. Learning human–environment interactions using conformal tactile textiles. *Nature Electronics* 4, 3 (2021), 193–201.

[89] Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. 2023. DRAMA: Joint Risk Localization and Captioning in Driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 1043–1052.

[90] Carina Manger, Jakob Peintner, Marion Hoffmann, Mirella Probst, Raphael Wennmacher, and Andreas Riener. 2023. Providing Explainability in Safety-Critical Automated Driving Situations through Augmented Reality Windshield HMIs. In *Adjunct Proceedings of the 15th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Ingolstadt, Germany) *(AutomotiveUI '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 174–179. https://doi.org/10.1145/3581961.3609874

[91] Carina Manger, Anna Preiwisch, Chiara Gambirasio, Simon Golks, Martina Schuß, and Andreas Riener. 2023. We're in This Together: Exploring Explanation Needs and Methods in Shared Automated Shuttle Buses. In *Proceedings of the 22nd International Conference on Mobile and Ubiquitous Multimedia* (, Vienna, Austria,) *(MUM '23)*. Association for Computing Machinery, New York, NY, USA, 145–151. https://doi.org/10.1145/3626705.3627798

[92] Carina Manger, Florian Pusch, Manuel Thöne, Marco Wenger, Andreas Löcken, and Andreas Riener. 2023. Explainability in Automated Parking: The Effect of Augmented Reality Visualizations on User Experience and Situation Awareness. In *Proceedings of the 22nd International Conference on Mobile and Ubiquitous Multimedia* (, Vienna, Austria,) *(MUM '23)*. Association for Computing Machinery, New York, NY, USA, 152–158. https://doi.org/10.1145/3626705.3627796

[93] Nikolas Martelaro and Wendy Ju. 2017. WoZ Way: Enabling Real-time Remote Interaction Prototyping & Observation in On-road Vehicles. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 169–182. https://doi.org/10.1145/2998181.2998293

[94] Jim McCambridge, John Witton, and Diana R. Elbourne. 2014. Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. *Journal of Clinical Epidemiology* 67, 3 (2014), 267–277. https://doi.org/10.1016/j.jclinepi.2013.08.015

[95] Angus McKerral, Kristen Pammer, and Cassandra Gauld. 2023. Supervising the self-driving car: Situation awareness and fatigue during highly automated driving. *Accident Analysis & Prevention* 187 (2023), 107068. https://doi.org/10.1016/j.aap.2023.107068

[96] Johanna Meurer, Christina Pakusch, Gunnar Stevens, Dave Randall, and Volker Wulf. 2020. A Wizard of Oz Study on Passengers' Experiences of a Robo-Taxi Service in Real-Life Settings. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (, Eindhoven, Netherlands,) *(DIS '20)*. Association for Computing Machinery, New York, NY, USA, 1365–1377. https://doi.org/10.1145/3357236.3395465

[97] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*. 404–411.

[98] Daniel C Molden. 2014. Understanding priming effects in social psychology: What is "social priming" and how does it occur? *Social cognition* 32, Supplement (2014), 1–11.

[99] Ananya Mondal, Martin Kaushal, and Suchetana Chakraborty. 2023. Sense as You Go: A Context-Aware Adaptive Sensing Framework for on-Road Driver Profiling. In *Proceedings of the 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation* (Istanbul, Turkey) *(BuildSys '23)*. Association for Computing Machinery, New York, NY, USA, 89–98. https://doi.org/10.1145/3600100.3623734

[100] Shokoufeh Monjezi Kouchak and Ashraf Gaffar. 2019. Using Bidirectional Long Short Term Memory with Attention Layer to Estimate Driver Behavior. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. 315–320. https://doi.org/10.1109/ICMLA.2019.00059

[101] Shokoufeh Monjezi Kouchak and Ashraf Gaffar. 2021. Detecting Driver Behavior Using Stacked Long Short Term Memory Network With Attention Layer. *IEEE Transactions on Intelligent Transportation Systems* 22, 6 (2021), 3420–3429. https://doi.org/10.1109/TITS.2020.2986697

[102] Luntian Mou, Yiyuan Zhao, Chao Zhou, Bahareh Nakisa, Mohammad Naim Rastgoo, Lei Ma, Tiejun Huang, Baocai Yin, Ramesh Jain, and Wen Gao. 2023. Driver Emotion Recognition With a Hybrid Attentional Multimodal Fusion Framework. *IEEE Transactions on Affective Computing* 14, 4 (2023), 2970–2981. https://doi.org/10.1109/TAFFC.2023.3250460

[103] Andrea Isabell Müller, Veronika Weinbeer, and Klaus Bengler. 2019. Using the wizard of Oz paradigm to prototype automated vehicles: methodological challenges. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular*

Applications: Adjunct Proceedings (Utrecht, Netherlands) *(AutomotiveUI '19)*. Association for Computing Machinery, New York, NY, USA, 181–186. https://doi.org/10.1145/3349263.3351526

[104] Richa Nahata, Daniel Omeiza, Rhys Howard, and Lars Kunze. 2021. Assessing and Explaining Collision Risk in Dynamic Environments for Autonomous Driving Safety. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. 223–230. https://doi.org/10.1109/ITSC48978.2021.9564966

[105] Mohammad Naiseh, Dena Al-Thani, Nan Jiang, and Raian Ali. 2021. Explainable recommendation: when design meets trust calibration. *World Wide Web* 24, 5 (2021), 1857–1884.

[106] Frederik Naujoks, Dennis Befelein, Katharina Wiedemann, and Alexandra Neukum. 2018. A Review of Non-driving-related Tasks Used in Studies on Automated Driving. In *Advances in Human Aspects of Transportation*, Neville A Stanton (Ed.). Springer International Publishing, Cham, 525–537.

[107] Frederik Naujoks, Katharina Wiedemann, and Nadja Schömig. 2017. The Importance of Interruption Management for Usefulness and Acceptance of Automated Driving. In *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Oldenburg, Germany) *(AutomotiveUI '17)*. Association for Computing Machinery, New York, NY, USA, 254–263. https://doi.org/10.1145/3122986.3123000

[108] Frederik Naujoks, Katharina Wiedemann, Nadja Schömig, Sebastian Hergeth, and Andreas Keinath. 2019. Towards guidelines and verification methods for automated vehicle HMIs. *Transportation Research Part F: Traffic Psychology and Behaviour* 60 (2019), 121–136. https://doi.org/10.1016/j.trf.2018.10.012

[109] Anton J Nederhof. 1985. Methods of coping with social desirability bias: A review. *European journal of social psychology* 15, 3 (1985), 263–280.

[110] Austin Lee Nichols and Jon K Maner. 2008. The good-subject effect: Investigating participant demand characteristics. *The Journal of general psychology* 135, 2 (2008), 151–166.

[111] Donald A Norman. 1990. The 'problem'with automation: inappropriate feedback and interaction, not 'over-automation'. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 327, 1241 (1990), 585–593.

[112] Kenta Okabe, Keiichi Watanuki, Kazunori Kaede, and Keiichi Muramatsu. 2018. Study on Estimation of Driver's State During Automatic Driving Using Seat Pressure. In *Intelligent Human Systems Integration*, Waldemar Karwowski and Tareq Ahram (Eds.). Springer International Publishing, Cham, 35–41.

[113] Daniel Omeiza, Sule Anjomshoae, Helena Webb, Marina Jirotka, and Lars Kunze. 2022. From Spoken Thoughts to Automated Driving Commentary: Predicting and Explaining Intelligent Vehicles' Actions. In *2022 IEEE Intelligent Vehicles Symposium (IV)*. 1040–1047. https://doi.org/10.1109/IV51971.2022.9827345

[114] Daniel Omeiza, Raunak Bhattacharyya, Nick Hawes, Marina Jirotka, and Lars Kunze. 2023. Effects of Explanation Specificity on Passengers in Autonomous Driving. arXiv:2307.00633 [cs.RO]

[115] Daniel Omeiza, Helena Web, Marina Jirotka, and Lars Kunze. 2021. Towards Accountability: Providing Intelligible Explanations in Autonomous Driving. In *2021 IEEE Intelligent Vehicles Symposium (IV)*. 231–237. https://doi.org/10.1109/IV48863.2021.9575917

[116] Daniel Omeiza, Helena Webb, Marina Jirotka, and Lars Kunze. 2022. Explanations in Autonomous Driving: A Survey. *IEEE Transactions on Intelligent Transportation Systems* 23, 8 (2022), 10142–10162. https://doi.org/10.1109/TITS.2021.3122865

[117] Ganesh Pai, Sarah Widrow, Jaydeep Radadiya, Cole D Fitzpatrick, Michael Knodler, and Anuj K Pradhan. 2020. A Wizard-of-Oz experimental approach to study the human factors of automated vehicles: Platform and methods evaluation. *Traffic injury prevention* 21, sup1 (2020), S140–S144.

[118] SungYeon Park, MinJae Lee, JiHyuk Kang, Hahyeon Choi, Yoonah Park, Juhwan Cho, Adam Lee, and DongKyu Kim. 2024. VLAAD: Vision and Language Assistant for Autonomous Driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. 980–987.

[119] Bastian Pfleging, Maurice Rang, and Nora Broy. 2016. Investigating user needs for non-driving-related activities during automated driving. In *Proceedings of the 15th International Conference on Mobile and Ubiquitous Multimedia* (Rovaniemi, Finland) *(MUM '16)*. Association for Computing Machinery, New York, NY, USA, 91–99. https://doi.org/10.1145/3012709.3012735

[120] Wolter Pieters. 2011. Explanation and trust: what to tell the user in security and AI? *Ethics and information technology* 13 (2011), 53–64.

[121] Jonas Radlmayr, Fabian Marco Fischer, and Klaus Bengler. 2019. The Influence of Non-driving Related Tasks on Driver Availability in the Context of Conditionally Automated Driving. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*, Sebastiano Bagnara, Riccardo Tartaglia, Sara Albolino, Thomas Alexander, and Yushi Fujita (Eds.). Springer International Publishing, Cham, 295–304.

[122] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. 2018. Toward Driving Scene Understanding: A Dataset for Learning Driver Behavior and Causal Reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[123] Dirk Rothenbücher, Jamy Li, David Sirkin, Brian Mok, and Wendy Ju. 2016. Ghost driver: A field study investigating the interaction between pedestrians and driverless vehicles. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 795–802. https://doi.org/10.1109/ROMAN.2016.7745210

[124] Carina Röckel and Heiko Hecht. 2023. Regular looks out the window do not maintain situation awareness in highly automated driving. *Transportation Research Part F: Traffic Psychology and Behaviour* 98 (2023), 368–381. https://doi.org/10.1016/j.trf.2023.09.015

[125] Enna Sachdeva, Nakul Agarwal, Suhas Chundi, Sean Roelofs, Jiachen Li, Mykel Kochenderfer, Chiho Choi, and Behzad Dariush. 2024. Rank2Tell: A Multimodal Driving Dataset for Joint Importance Ranking and Reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 7513–7522.

[126] Stanisław Saganowski, Joanna Komoszyńska, Maciej Behnke, Bartosz Perz, Dominika Kunc, Bartłomiej Klich, Łukasz D Kaczmarek, and Przemysław Kazienko. 2022. Emognition dataset: emotion recognition with self-reports, facial expressions, and physiology using wearables. *Scientific data* 9, 1 (2022), 158.

[127] Tobias Schneider, Sabiha Ghellal, Steve Love, and Ansgar R.S. Gerlicher. 2021. Increasing the User Experience in Autonomous Driving through Different Feedback Modalities. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 7–10. https://doi.org/10.1145/3397481.3450687

[128] Tobias Schneider, Joana Hois, Alischa Rosenstein, Sabiha Ghellal, Dimitra Theofanou-Fülbier, and Ansgar R.S. Gerlicher. 2021. ExplAIn Yourself! Transparency for Positive UX in Autonomous Driving. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 161, 12 pages. https://doi.org/10.1145/3411764.3446647

[129] Tobias Schneider, Joana Hois, Alischa Rosenstein, Sandra Metzl, Ansgar R.S. Gerlicher, Sabiha Ghellal, and Steve Love. 2023. Don't Fail Me! The Level 5 Autonomous Driving Information Dilemma Regarding Transparency and User Experience. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 540–552. https://doi.org/10.1145/3581641.3584085

[130] Rob Semmens, Nikolas Martelaro, Pushyami Kaveti, Simon Stent, and Wendy Ju. 2019. Is Now A Good Time? An Empirical Study of Vehicle-Driver Communication Timing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300867

[131] Minwoo Seong, Gwangbin Kim, Dohyeon Yeo, Yumin Kang, Heesan Yang, Joseph DelPreto, Wojciech Matusik, Daniela Rus, and SeungJun Kim. 2024. MultiSenseBadminton: Wearable Sensor–Based Biomechanical Dataset for Evaluation of Badminton Performance. *Scientific Data* 11, 1 (2024), 343.

[132] Yuan Shen, Shanduojiao Jiang, Yanlin Chen, and Katie Driggs Campbell. 2022. To Explain or Not to Explain: A Study on the Necessity of Explanations for Autonomous Vehicles. arXiv:2006.11684 [cs.AI]

[133] Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. 2019. The Performance of LSTM and BiLSTM in Forecasting Time Series. In *2019 IEEE International Conference on Big Data (Big Data)*. 3285–3292. https://doi.org/10.1109/BigData47090.2019.9005997

[134] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. 2023. DriveLM: Driving with Graph Visual Question Answering. arXiv:2312.14150 [cs.CV]

[135] Daniele Sportillo, Alexis Paljic, and Luciano Ojeda. 2020. On-Road Evaluation of Autonomous Driving Training. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction* (Daegu, Republic of Korea) *(HRI '19)*. IEEE Press, 182–190.

[136] Annika Stampf, Mark Colley, and Enrico Rukzio. 2022. Towards Implicit Interaction in Highly Automated Vehicles - A Systematic Literature Review. *Proc. ACM Hum.-Comput. Interact.* 6, MHCI, Article 191 (sep 2022), 21 pages. https://doi.org/10.1145/3546726

[137] Jork Stapel, Freddy Antony Mullakkal-Babu, and Riender Happee. 2017. Driver behavior and workload in an on-road automated vehicle. In *Proceedings of the RSS2017 Conference*.

[138] Rachel E. Stuck, Brianna J. Tomlinson, and Bruce N. Walker. 2022. The importance of incorporating risk into human-automation trust. *Theoretical Issues in Ergonomics Science* 23, 4 (2022), 500–516. https://doi.org/10.1080/1463922X.2021.1975170 arXiv:https://doi.org/10.1080/1463922X.2021.1975170

[139] Amos Tversky and Daniel Kahneman. 1981. The Framing of Decisions and the Psychology of Choice. *Science* 211, 4481 (1981), 453–458. https://doi.org/10.1126/science.7455683 arXiv:https://www.science.org/doi/pdf/10.1126/science.7455683

[140] Christian Ulrich, Benjamin Frieske, Stephan A. Schmid, and Horst E. Friedrich. 2022. Monitoring and Forecasting of Key Functions and Technologies for Automated Driving. *Forecasting* 4, 2 (2022), 477–500. https://doi.org/10.3390/forecast4020027

[141] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[142] Francesco Walker, Yannick Forster, Sebastian Hergeth, Johannes Kraus, William Payre, Philipp Wintersberger, and Marieke Martens. 2023. Trust in automated vehicles: constructs, psychological processes, and assessment. *Frontiers in Psychology* 14 (2023). https://doi.org/10.3389/fpsyg.2023.1279271

[143] Heather E.K. Walker, Rachel A. Eng, and Lana M. Trick. 2021. Dual-task decrements in driving performance: The impact of task type, working memory, and the frequency of task performance. *Transportation Research Part F: Traffic Psychology and Behaviour* 79 (2021), 185–204. https://doi.org/10.1016/j.trf.2021.04.021

[144] Peter Wang, Srinath Sibi, Brian Mok, and Wendy Ju. 2017. Marionette: Enabling On-Road Wizard-of-Oz Autonomous Driving Studies. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna, Austria) *(HRI '17)*. Association for Computing Machinery, New York, NY, USA, 234–243. https://doi.org/10.1145/2909824.3020256

[145] Gesa Wiegand, Malin Eiband, Maximilian Haubelt, and Heinrich Hussmann. 2020. "I'd like an Explanation for That!"Exploring Reactions to Unexpected Autonomous Driving. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services* (Oldenburg, Germany) *(MobileHCI '20)*. Association for Computing Machinery, New York, NY, USA, Article 36, 11 pages. https://doi.org/10.1145/3379503.3403554

[146] Gesa Wiegand, Matthias Schmidmaier, Thomas Weber, Yuanting Liu, and Heinrich Hussmann. 2019. I Drive - You Trust: Explaining Driving Behavior Of Autonomous Cars. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312817

[147] Philipp Wintersberger, Hannah Nicklas, Thomas Martlbauer, Stephan Hammer, and Andreas Riener. 2020. Explainable Automation: Personalized and Adaptive UIs to Foster Trust and Understanding of Driving Automation Systems. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Virtual Event, DC, USA) *(AutomotiveUI '20)*. Association for Computing Machinery, New York, NY, USA, 252–261. https://doi.org/10.1145/3409120.3410659

[148] Tong Wu, Nikolas Martelaro, Simon Stent, Jorge Ortiz, and Wendy Ju. 2021. Learning When Agents Can Talk to Drivers Using the INAGT Dataset and Multisensor Fusion. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3, Article 133 (sep 2021), 28 pages. https://doi.org/10.1145/3478125

[149] Yiran Xu, Xiaoyin Yang, Lihang Gong, Hsuan-Chu Lin, Tz-Ying Wu, Yunsheng Li, and Nuno Vasconcelos. 2020. Explainable Object-Induced Action Decision for Autonomous Vehicles. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9520–9529. https://doi.org/10.1109/CVPR42600.2020.00954

[150] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee. K. Wong, Zhenguo Li, and Hengshuang Zhao. 2023. DriveGPT4: Interpretable End-to-end Autonomous Driving via Large Language Model. arXiv:2310.01412 [cs.CV]

[151] Shiyan Yang, Angus McKerral, Megan Dawn Mulhall, Michael Graeme Lenné, Bryan Reimer, and Pnina Gershon. 2023. Takeover Context Matters: Characterising Context of Takeovers in Naturalistic Driving using Super Cruise and Autopilot. In *Proceedings of the 15th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Ingolstadt, Germany) *(AutomotiveUI '23)*. Association for Computing Machinery, New York, NY, USA, 112–122. https://doi.org/10.1145/3580585.3606459

[152] Yi Yang, Qingwen Zhang, Ci Li, Daniel Simões Marta, Nazre Batool, and John Folkesson. 2024. Human-Centric Autonomous Systems With LLMs for User Command Reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. 988–994.

[153] Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Ella Atkins, and David Crandall. 2020. When, where, and what? A new dataset for anomaly detection in driving videos. *arXiv preprint arXiv:2004.03044* (2020).

[154] Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Yuchen Wang, Ella Atkins, and David J. Crandall. 2023. DoTA: Unsupervised Detection of Traffic Anomaly in Driving Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 1 (2023), 444–459. https://doi.org/10.1109/TPAMI.2022.3150763

[155] Dohyeon Yeo, Gwangbin Kim, and SeungJun Kim. 2019. MAXIM: Mixed-reality Automotive Driving XIMulation. In *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. 460–464. https://doi.org/10.1109/ISMAR-Adjunct.2019.00124

[156] Dohyeon Yeo, Gwangbin Kim, and Seungjun Kim. 2020. Toward Immersive Self-Driving Simulations: Reports from a User Study across Six Platforms. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376787

[157] Sol Hee Yoon, Young Woo Kim, and Yong Gu Ji. 2019. The effects of takeover request modalities on highly automated car control transitions. *Accident Analysis & Prevention* 123 (2019), 150–158. https://doi.org/10.1016/j.aap.2018.11.018

[158] Tackgeun You and Bohyung Han. 2020. Traffic Accident Benchmark for Causality Recognition. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 540–556.

[159] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. 2020. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[160] Nidzamuddin Md. Yusof, Juffrizal Karjanto, Muhammad Zahir Hassan, Jacques Terken, Frank Delbressine, and Matthias Rauterberg. 2022. Reading During Fully Automated Driving: A Study of the Effect of Peripheral Visual and Haptic Information on Situation Awareness and Mental Workload. *IEEE Transactions on Intelligent Transportation Systems* 23, 10 (2022), 19136–19144. https://doi.org/10.1109/TITS.2022.3165192

[161] Nima Zargham, Leon Reicherts, Michael Bonfert, Sarah Theres Voelkel, Johannes Schoening, Rainer Malaka, and Yvonne Rogers. 2022. Understanding Circumstances for Desirable Proactive Behaviour of Voice Assistants: The Proactivity Dilemma. In *Proceedings of the 4th Conference on Conversational User Interfaces* (Glasgow, United Kingdom) *(CUI '22)*. Association for Computing Machinery, New York, NY, USA, Article 3, 14 pages. https://doi.org/10.1145/3543829.3543834

[162] Qiaoning Zhang, Connor Esterwood, Anuj K. Pradhan, Dawn Tilbury, X. Jessie Yang, and Lionel P. Robert. 2023. The Impact of Modality, Technology Suspicion, and NDRT Engagement on the Effectiveness of AV Explanations. *IEEE Access* 11 (2023), 81981–81994.

https://doi.org/10.1109/ACCESS.2023.3302261
[163] Qiaoning Zhang, X. Jessie Yang, and Lionel P. Robert. 2021. What and When to Explain? A Survey of the Impact of Explanation on Attitudes Toward Adopting Automated Vehicles. *IEEE Access* 9 (2021), 159533–159540. https://doi.org/10.1109/ACCESS.2021.3130489
[164] Mingming Zhao, Georges Beurier, Hongyan Wang, and Xuguang Wang. 2021. Exploration of Driver Posture Monitoring Using Pressure Sensors with Lower Resolution. *Sensors* 21, 10 (2021). https://doi.org/10.3390/s21103346
[165] Meng Zhou, Diem-Trinh Le, Duy Quy Nguyen-Phuoc, P. Christopher Zegras, and Joseph Ferreira. 2021. Simulating impacts of Automated Mobility-on-Demand on accessibility and residential relocation. *Cities* 118 (2021), 103345. https://doi.org/10.1016/j.cities.2021.103345
[166] Xingcheng Zhou, Mingyu Liu, Bare Luka Zagar, Ekim Yurtsever, and Alois C Knoll. 2023. Vision language models in autonomous driving and intelligent transportation systems. *arXiv preprint arXiv:2310.14414* (2023).

Type	Sensor	Product	Data Channel	Feature Channel	Refresh Rate	Format	Availability
Exteroception	3D LiDAR	VLP-16 (Velodyne)	Point Cloud (x, y, z, reflectance)	Point Cloud (x, y, z, reflectance)	10 fps	MAT	Publicly Available
	Stereo Camera	ZED (Stereo Labs)	Stereo Image	Stereo Image	15 fps	CSV	Publicly Available
Proprioception	GPS	ZED-F9R dual-band GNSS (u-blox)	Latitude, Longitude, HeightMSL, VelN, VelE, VelD, Heading	Latitude, Longitude, HeightMSL, VelN, VelE, VelD, Heading	5 fps	CSV	Publicly Available
	OBD	ELM327	Speed, Throttle, Brake, Steering	Speed, Throttle, Brake, Steering	5 fps	CSV	Publicly Available
	IMU	HWT905 (Wit Motion)	Acc (x, y, z), Angle (x, y, z), Angular velocity (x, y, z),	Acc (x, y, z), Angle (x, y, z), Angular velocity (x, y, z),	50 fps	CSV	Publicly Available
Interoception	Physiological Response	E4 Wristband (Empatica)	Galvanic Skin Response (Electrodermal Activity)	Galvanic Skin Reponse (Electrodermal Activity)	4 fps	CSV, HDF5	Publicly Available
				Phasic EDA	4 fps	CSV	Publicly Available
				Tonic EDA	4 fps	CSV	Publicly Available
			Blood Volume Pulse	Blood Volume Pulse	64 fps	CSV, HDF5	Publicly Available
				Interbeat Interval	1 fps	CSV, HDF5	Publicly Available
			Temperature	Temperature	4 fps	CSV, HDF5	Publicly Available
			Acceleration (x, y, z)	Acceleration (x, y, z)	32 fps	CSV, HDF5	Publicly Available
	Tactile Seat	Intelligent Carpet [88]	Resistive Sensing Pressure (32×32 matrix)	Resistive Sensing Pressure (32×32 matrix)	11 fps	CSV, HDF5	Publicly Available
	Thermal Imaging	Lepton 3.5 (FLIR)	Temperature Heatmap (160×120 matrix)	Temperature Heatmap (160×120 matrix)	8.7 fps	PNG	Publicly Available
	Depth Camera	D435 (Intel RealSense)	RGB Image	Facial Mesh, Skeletal Tracking	20 fps	JSON	Publicly Available
				RGB Image	20 fps	PNG	Available Upon Request with IRB
			Depth Image	Depthmap	20 fps	PNG	Available Upon Request with IRB
	LiDAR Camera	L515 (Intel RealSense)	Depth Image	Depthmap	20 fps	PNG	Available Upon Request with IRB

Table 4. TimelyTale dataset composition: type, Product, formatting, and refresh rate of the data and features from each sensor stream.

A Explanation Coding

Type	Code	Description
Actions	The car changes lanes	Passengers request an explanation for the vehicle's lane-changing behavior. When the passenger explicitly mentions the need for an explanation about lane changes or if the request arises after the vehicle has changed lanes, and the experimenter confirms that the inquiry pertains to the action of lane-changing with the passenger.
	The car stops	Passengers request an explanation for the vehicle's stopping behavior. When the passenger explicitly mentions the need for an explanation about the stop or if the request arises after the vehicle has come to a stop, and the experimenter confirms that the inquiry pertains to the action of stopping with the passenger.
	The car turns left	Passengers request an explanation for the vehicle's left-turning behavior. When the passenger explicitly mentions the need for an explanation about the left turn or if the request arises after the vehicle has turned left, and the experimenter confirms that the inquiry pertains to the action of left-turning with the passenger.
	The car turns right	Passengers request an explanation for the vehicle's right-turning behavior. When the passenger explicitly mentions the need for an explanation about the right turn or if the request arises after the vehicle has turned right, and the experimenter confirms that the inquiry pertains to the action of right-turning with the passenger.
	The car speeds up	Passengers request an explanation for the vehicle's acceleration. When the passenger explicitly mentions the need for an explanation about speeding up or if the request arises after the vehicle has accelerated, and the experimenter confirms that the inquiry pertains to the action of speeding up with the passenger. When the vehicle starts, both speeding up and starting codes may be applied.
	The car slows down	Passengers request an explanation for the vehicle's deceleration. When the passenger explicitly mentions the need for an explanation about slowing down or if the request arises after the vehicle has decelerated, and the experimenter confirms that the inquiry pertains to the action of slowing down with the passenger. When the vehicle stops, both slowing down and stopping codes may be applied.
	The car starts	Passengers request an explanation for the vehicle's sudden start from a full stop. When the passenger explicitly mentions the need for an explanation about starting or if the request arises after the vehicle has started moving, and the experimenter confirms that the inquiry pertains to the action of starting with the passenger.
	The car enters a merging area	Passengers request an explanation for the vehicle entering a merging area. When the passenger explicitly mentions the need for an explanation about entering the merging area or if the request arises after the vehicle has entered the merging area, and the experimenter confirms that the inquiry pertains to the action of entering a merging area with the passenger.
	The car exits the highway	Passengers request an explanation for the vehicle exiting the highway. When the passenger explicitly mentions the need for an explanation about exiting the highway or if the request arises after the vehicle has exited the highway, and the experimenter confirms that the inquiry pertains to the action of exiting the highway with the passenger.
	The car merges onto the road	Passengers request an explanation for the vehicle merging onto a larger road. When the passenger explicitly mentions the need for an explanation about merging or if the request arises after the vehicle has merged onto the road, and the experimenter confirms that the inquiry pertains to the action of merging onto the road with the passenger.

Table 5. Coded explanations for the vehicle's actions

Type	Code	Description
Justifications	because the traffic light is red	Mostly related to the vehicle stopping or slowing down due to a red traffic light. This justification is used when passengers explicitly mention the red traffic light as the reason or through follow-up questions that clarify the reason for the vehicle's stop or deceleration.
	to pass the traffic light	Solely related to the vehicle speeding up to pass a traffic light before it turns red. This justification is used when passengers explicitly mention the need to pass the traffic light before it turns red as the reason or through follow-up questions that clarify the reason for the vehicle's acceleration.
	due to traffic	Mostly related to the vehicle stopping or slowing down due to general traffic conditions. This justification distinguishes between general traffic congestion and the need to maintain distance from the front car. It is used when passengers explicitly mention traffic as the reason or through follow-up questions that clarify the reason for the vehicle's stop or deceleration.
	to reach a destination	Solely related to the vehicle changing lanes to reach a specific destination. This justification is used when passengers explicitly mention the destination as the reason or through follow-up questions that clarify the reason for the vehicle's lane change.
	because the traffic light turned green	Solely related to the vehicle starting or moving due to the traffic light turning green. This justification is used when passengers explicitly mention the green traffic light as the reason or through follow-up questions that clarify the reason for the vehicle's start.
	to maintain distance from the front car	Mostly related to the vehicle slowing down to maintain a safe distance from the car in front. This justification distinguishes between the need to maintain distance from the front car and general traffic conditions. It is used when passengers explicitly mention this reason or through follow-up questions that clarify the reason for the vehicle's deceleration.
	as it enters a school zone	Solely related to the vehicle slowing down upon entering a school zone. This justification is used when passengers explicitly mention the school zone as the reason or through follow-up questions that clarify the reason for the vehicle's deceleration.
	to exit a road	Solely related to the vehicle changing lanes to exit a road. This justification is used when passengers explicitly mention the exit as the reason or through follow-up questions that clarify the reason for the vehicle's lane change.
	as a new road merges	Solely related to the vehicle changing lanes as a new road merges. This justification is used when passengers explicitly mention the merging road as the reason or through follow-up questions that clarify the reason for the vehicle's lane change.

Table 6. Coded explanations for the vehicle's action justifications

B Dataset Composition

In this section, we detail the composition and formatting of each data channel. See Figure 18 for the structure and exemplary view of the data repository. The complete list of available data is provided in Table 4. For illustration purposes, we include exemplary data from Subject 23 in the appendix.

B.1 Metadata

Our metadata includes details such as the subject's gender, age, driving experience, and the start and stop times of the experiment. Although the data is in CSV format, we offer it as a 7z compressed file to avoid Harvard Dataverse's automatic conversion to tabular form. Figure 19 shows the header of the metadata.csv file.

B.2 Exterocpetion

B.2.1 ZED Stereo Camera. Driving front view images was recorded with a stereo camera on the top of the car. These ZED images are uploaded in a split-compressed 7z format to comply with repository requirements (<2.5GB per file), under the (SubjectID\Exteroception) folder. After downloading and decompressing the files, one can

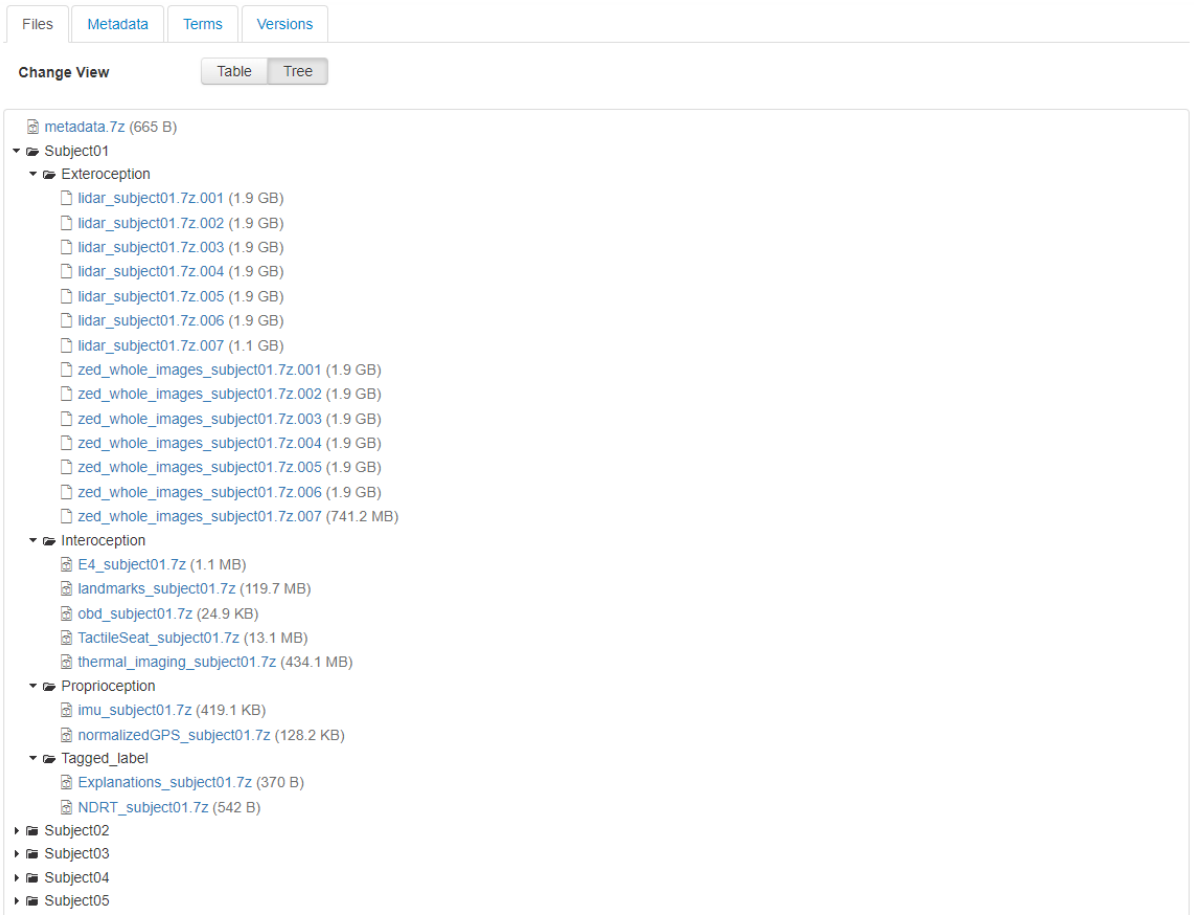


Fig. 18. TimelyTale Dataset Structure: A view available in the Harvard Dataverse public repository.

subjectID	gender	age	driving_experience	start_time	stop_time
1	female	26	3	1695307570789	1695307579602
2	male	26	3	1695349744191	1695349814289
3	female	35	10	1695365263110	1695365356098
4	male	26	1	1695378842656	1695378902861
5	male	25	2	1695382774098	1695382827755

Fig. 19. Example data and formatting of the metadata.csv file.

find a folder named zed_whole_images_subjectID. Each file contains a ZED stereo image with a resolution of 1344×376. The filenames follow the format UNIXTIME_zed (See Figure 21 for an example).

Fig. 20. Exemplary ZED image captured for participant 23 (1695616259554_zed.png under the directory zed_whole_images_subject23).

B.2.2 3D LiDAR point cloud. Point cloud data from the VLP-16 is also provided. Similar to the ZED images, these are uploaded in a split-compressed 7z format under the (SubjectID\Exteroception) folder in the TimelyTale dataset repository. Uncompressing the files yields a MAT file titled lidar_subject%d (where %d represents the subjectID). The structure of LiDAR point cloud data is detailed as follows:

- **Location**: This variable represents the spatial coordinates of each point in the point cloud, formatted as a 16×1808×3 array of doubles. Each element in this array corresponds to a point in 3D space, with the dimensions representing:
  - 16 rows, each corresponding to one of the VLP-16 LiDAR's 16 laser channels.
  - 1808 columns, representing the number of points sampled by each laser channel in one full rotation of the LiDAR sensor.
  - 3 layers in the third dimension, specifying the x, y, and z coordinates of each point relative to the LiDAR sensor's position.
- **XLimits**: This variable specifies the minimum and maximum x-coordinates observed in the entire point cloud dataset, formatted as a 2-element array $[-112.312793074318, 105.032973380692]$. These values define the horizontal span of the scanned area in the x-direction.
- **YLimits**: Similar to 'XLimits', this variable indicates the minimum and maximum y-coordinates within the point cloud, given as a 2-element array $[-93.2348408735782, 82.2257520253454]$. It outlines the extent of the scanned area in the y-direction, perpendicular to the x-axis.
- **ZLimits**: This variable provides the vertical range of the scanned area by specifying the minimum and maximum z-coordinates, noted as $[-4.57091774634623, 23.7673529117645]$ in a 2-element array. These limits indicate the lowest and highest points captured by the LiDAR in relation to its mounting position.
- **Intensity**: This array, sized 16×1808, records the intensity of the returned laser signal for each point. Intensity values are indicative of the reflective properties of the surfaces that the LiDAR beams have encountered. These values can be used to differentiate between types of materials and surfaces in the environment based on how much light they reflect back to the sensor.

B.3 Proprioception

B.3.1 GPS. The GPS file is formatted as normalizedGPS_subject%d.csv. The data is tagged with timestamps and is normalized by subtracting 35 degrees from latitude and 127 degrees from longitude. Each column of the data contains the following information.

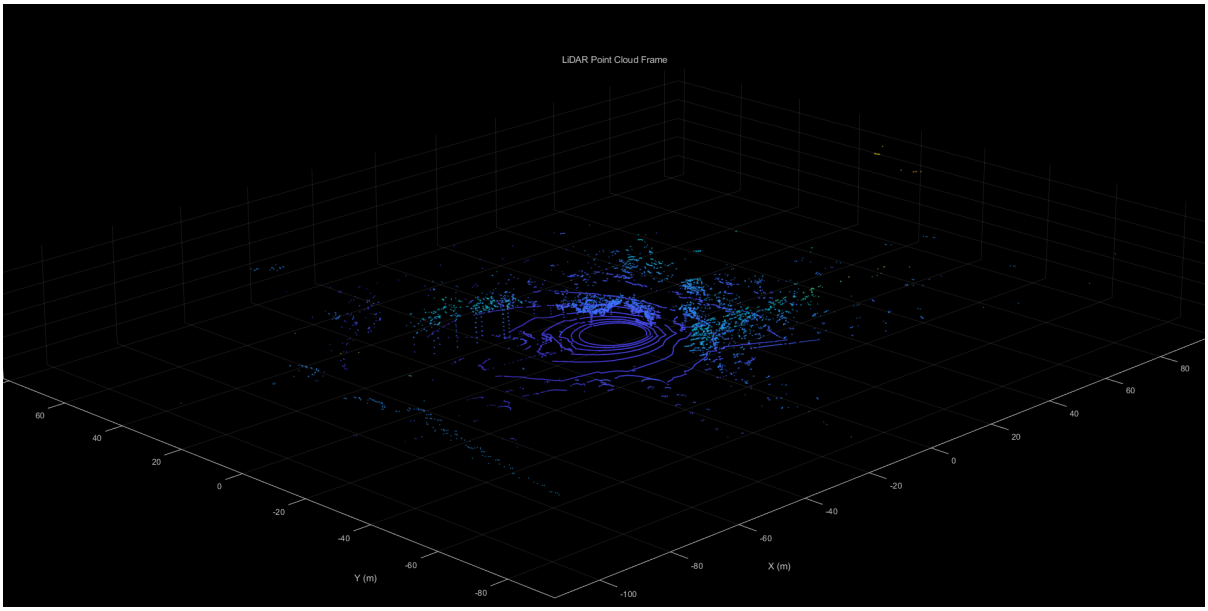


Fig. 21. Exemplary LiDAR point cloud captured for participants 23 at 1695615171.410340 UNIX time.

- **Latitude**: Geographical latitude of the GNSS module's current position, originally recorded in degrees with the Equator as 0°. The latitude has been normalized by subtracting 35 degrees from the original value.
- **Longitude**: Geographical longitude of the GNSS module's current position, originally recorded in degrees with the Prime Meridian as 0°. The latitude has been normalized by subtracting 127 degrees from the original value.
- **HeightMSL**: Height above mean sea level, indicating the altitude of the GNSS module relative to average sea level, measured in meters.
- **VeIN**: Velocity towards the north, indicating the speed and direction of the GNSS module moving northward, measured in meters per second.
- **VeIE**: Velocity towards the east, indicating the speed and direction of the GNSS module moving eastward, measured in meters per second.
- **VeID**: Velocity downwards, indicating the speed and direction of the GNSS module moving towards the Earth's center, measured in meters per second.
- **Heading**: The direction of the GNSS module's movement, measured in degrees from true north.

B.3.2 IMU. The IMU sensor data are encapsulated in files named according to the convention: im_subject%d.csv, which is uploaded in a 7z compressed format to prevent automatic conversion in the database. Each file pertains to a specific subject and contains time-stamped readings from the IMU sensor, which captures both linear acceleration and angular velocity, along with orientation angles. The data encompass the following measurements along the x, y, and z axes:

- **Acc_x**: Acceleration along the x-axis, measuring the rate of change of velocity in the direction perpendicular to the y-z plane. It is recorded in meters per second squared (m/s^2).
- **Acc_y**: Acceleration along the y-axis, measuring the rate of change of velocity in the direction perpendicular to the x-z plane. It is also recorded in meters per second squared (m/s^2).

Timestamps	Latitude	Longitude	HeightMSL	VeIN	VeIE	VeID	Heading
1695615173.0	0.2258834	-0.1596349	33.262	-0.015	-2.4e-06	-0.001	18841540.0
1695615173.0	0.2258834	-0.159635	33.263	-0.034	-6.1e-06	-0.001	18841926.0
1695615174.0	0.2258833	-0.1596352	33.264	-0.051	-9.5e-06	-0.002	18841810.0
1695615174.0	0.2258832	-0.1596354	33.265	-0.069	-1.29e-05	-0.003	18841978.0
1695615174.0	0.2258831	-0.1596357	33.266	-0.086	-1.62e-05	-0.003	18843354.0

Fig. 22. Example data and formatting of the normalizedGPS_subject05.csv file.

- **Acc_z**: Acceleration along the z-axis, measuring the rate of change of velocity in the direction perpendicular to the x-y plane. This dimension typically represents the vertical acceleration and is recorded in meters per second squared (m/s^2).
- **W_x**: Angular velocity around the x-axis, indicating the rate of rotation around the axis perpendicular to the y-z plane. It is measured in radians per second (rad/s).
- **W_y**: Angular velocity around the y-axis, indicating the rate of rotation around the axis perpendicular to the x-z plane. It is measured in radians per second (rad/s).
- **W_z**: Angular velocity around the z-axis, indicating the rate of rotation around the axis perpendicular to the x-y plane. It is measured in radians per second (rad/s).
- **Angle_x**: Orientation angle around the x-axis, representing the tilt or rotation angle relative to the horizontal plane, measured in degrees (°).
- **Angle_y**: Orientation angle around the y-axis, representing the tilt or rotation angle relative to the horizontal plane, measured in degrees (°).
- **Angle_z**: Orientation angle around the z-axis, also known as the yaw angle, representing the sensor's rotation about the vertical axis, measured in degrees (°).

Timestamps	Acc_x	Acc_y	Acc_z	W_x	W_y	W_z	Angle_x	Angle_y	Angle_z
1695615672.0	0.03125	-0.043945313	1.002441406	0.610351563	0.610351563	1.220703125	-1.922607422	2.048950195	-105.7818604
1695615672.0	0.03125	-0.043945313	1.002441406	0.610351563	0.610351563	1.220703125	-1.944580078	2.120361328	-105.7543945
1695615672.0	0.03125	-0.043945313	1.002441406	0.610351563	0.610351563	1.220703125	-1.944580078	2.120361328	-105.7543945
1695615672.0	-0.032714844	-0.042480469	0.995605469	0.610351563	0.610351563	1.220703125	-1.944580078	2.120361328	-105.7543945
1695615672.0	-0.032714844	-0.042480469	0.995605469	0.732421875	0.793457031	1.46484375	-1.944580078	2.120361328	-105.7543945

Fig. 23. Example data and formatting of the imu_subject05.csv file.

B.3.3 OBD-II. The OBD-II sensor data are in files named according to the convention: obd_subject%d.csv. Each file pertains to a specific vehicle and contains time-stamped readings from the OBD-2 system, which captures various parameters related to the vehicle's performance and state. The data include the following measurements:

- **Speed**: The vehicle's speed, measured in kilometers per hour (km/h). This value represents the instantaneous speed of the vehicle at the time of data collection.
- **Throttle**: The throttle position, indicating the degree to which the throttle valve is open and allowing air into the engine, measured as a percentage of the maximum throttle position. The range of this value is from 0 to 100%, where 0% means the throttle is completely closed and 100% means it is fully open.

- **Brake**: The brake status, indicating whether the brake is applied or not. A value of 89 denotes the brake is engaged (on), and a value of 90 indicates the brake is not pressed (unpressed). Values other than 89 or 90 are considered errors, mostly due to simultaneous readout/write-in I/O structure conflicts.
- **Steering**: The steering wheel's angle, measured in degrees. The valid range for this measurement is from -720 degrees to +720 degrees, representing the full range of left to right steering capability. Values outside this range are considered errors, attributed to the same simultaneous readout/write-in I/O structure issues as with the brake data.

Timestamps	Speed	Acc	Brake	Steering
1695615322	28	37	90	11
1695615323	28	37	90	11
1695615323	28	37	90	11
1695615323	28	33	90	11
1695615323	28	33	90	11

Fig. 24. Example data and formatting of the obd_subject05.csv file.

B.4 Interoception

B.4.1 E4 Wristband. Data from the E4 wristband were recorded for each channel and are uploaded in a single 7z zip file to prevent automatic conversion to tabular format. Upon unzipping the file, separate data streams from the E4 wristband are available, each tagged with UNIX timestamps in individual files:

- **E4_acc_subject%d.csv**: 3-axis acceleration (x, y, z)
- **E4_bvp_subject%d.csv**: Raw data of blood volume pulse received from the E4 streaming server.
- **E4_gsr_subject%d.csv**: We provide both the raw data and phasic and tonic features extracted by the software *Ledalab* [11, 12] to assist in the analysis of moment-specific and long-term changes in the GSR signal.
  - **raw_data**: Raw GSR signal received from the E4 streaming server.
  - **tonic_data**: Skin conductance level (SCL), consists of general, relatively long-term arousal.
  - **phasic_data**: Skin conductance response (SCR), consists of transient changes with spikes or peaks.
- **E4_ibi_subject%d.csv**: inter-beat interval feature calculated from the BVP signal
- **E4_tmp_subject%d.csv**: body temperature in $^{\circ}C$

Given the potential for motion artifacts in a moving vehicle, we computed the Signal-to-Noise Ratio (SNR) to validate the quality of the physiological signals recorded by the E4 wristband (see Table 7). We employed a second-order polynomial fit to the autocorrelation function data, following the method outlined by Saganowski et al. [126]. The mean SNR ranged from 22.37 dB to 45.95 dB, indicating the signal's high quality.

B.4.2 Tactile Seat. The data from the Intelligent Carpet Tactile Seat is provided in two separate files for lower and upper seat:

- Back-side pressure: TactileSeat_upper_subject%d.csv
- Seat-side pressure: TactileSeat_lower_subject%d.csv

The row data is each data frame while each column are assigned for 32×32 sensor array, using name protocol %d-%d, ranging from 0-0 to 32-32. The data is also provided together with E4 as HDF5 file format for comprehensive analysis (Figure 26), the format of which was adopted from the ActionSense framework [27].

Timestamps	acc_x	acc_y	acc_z
1695382645	-36	-8	52
1695382645	-37	-9	51
1695382645	-36	-8	52
1695382645	-36	-6	53
1695382645	-39	-9	50

(a)

Timestamps	raw_data	tonic_data	phasic_data
1695382647.0	0.9265577	0.788608961	0.138099547
1695382648.0	0.9393694	0.788608961	0.145302209
1695382648.0	1.003428	0.788608939	0.167650733
1695382648.0	1.035458	0.788608851	0.200408688
1695382648.0	0.9676244	0.788608656	0.212332795

(b)

Timestamps	tmp
1695382643.0	31.07
1695382643.0	31.07
1695382644.0	31.07
1695382644.0	31.07
1695382644.0	31.05

(c)

Timestamps	ibi
1695382657.0	0.8281634
1695382657.0	0.8125377
1695382658.0	0.7812862
1695382659.0	0.796912
1695382699.0	0.8750405

(d)

Timestamps	bvp
1695382645.0	-35.24529
1695382645.0	-32.15449
1695382645.0	-30.74363
1695382645.0	-32.82005
1695382645.0	-38.44808

(e)

Fig. 25. (a) E4_acc_subject05, (b) E4_gsr_subject05, (c) E4_tmp_subject05, (d) E4_ibi_subject05, (e) E4_bvp_subject05.

Channel	mean	std	min	q15	q25	q50	q75	q95	max
bvp	33.61	1.874	29.59	31.76	32.43	33.46	34.92	36.43	37.74
ibi	22.38	3.076	13.37	20.19	21.16	22.99	24.62	25.68	27.88
tmp	41.10	0.7778	39.92	40.53	40.68	41.03	41.32	42.16	44.14
raw_data	34.49	5.58	19.45	29.62	31.40	36.33	37.73	40.06	44.45
tonic_data	45.95	7.086	36.73	39.37	40.61	44.94	49.45	58.47	64.89
phasic_data	28.96	4.192	17.80	24.53	28.88	30.26	31.51	32.89	34.17

Table 7. Validation of the quality of the physiological responses

B.4.3 Lepton 3.5. Thermal imaging data recorded with *Lepton 3.5* images are uploaded in a 7z format, under the (SubjectID\Interoception) folder. Each file under the thermal_imaging_subjectID folder are named UNIX-TIME_heat.png. Each file contains a *fusion* color-coded normalized thermal image recording passenger's facial heat with 160×120 resolution. These images are not person-discernable and we got consent upon disclosure (See Figure 21 for an example.). While we minimized vision-related sensing which could specify one's identify, we included heat upon subject's consent, considering their direct relationship with cognitive load or attention-rleated user state.

Thermal imaging data, captured with the *Lepton 3.5* camera, are stored in 7z format within the (SubjectID\Interoception) directory. Each image file, located in the thermal_imaging_subjectID folder, is named according to the UNIXTIME_heat.png convention. These images, represented with *fusion* color-coded scheme, provide normalized thermal imaging of passengers' facial heat at a 160×120 resolution. These images are designed to be non-identifiable, ensuring privacy; consent for their use was obtained prior to data collection (for an example, see Figure 27). In our efforts to respect privacy, we specifically minimized using vision-related sensing techniques that could potentially identify individuals. However, with the subjects' consent, we included thermal imaging due to its relevance in measuring cognitive load [2] and attention-related user states [1].

We also provide the facial and landmark detection results in our dataset for further analysis, specifically focusing on the temperature distribution of the face or the temperature of specific facial landmark positions. We

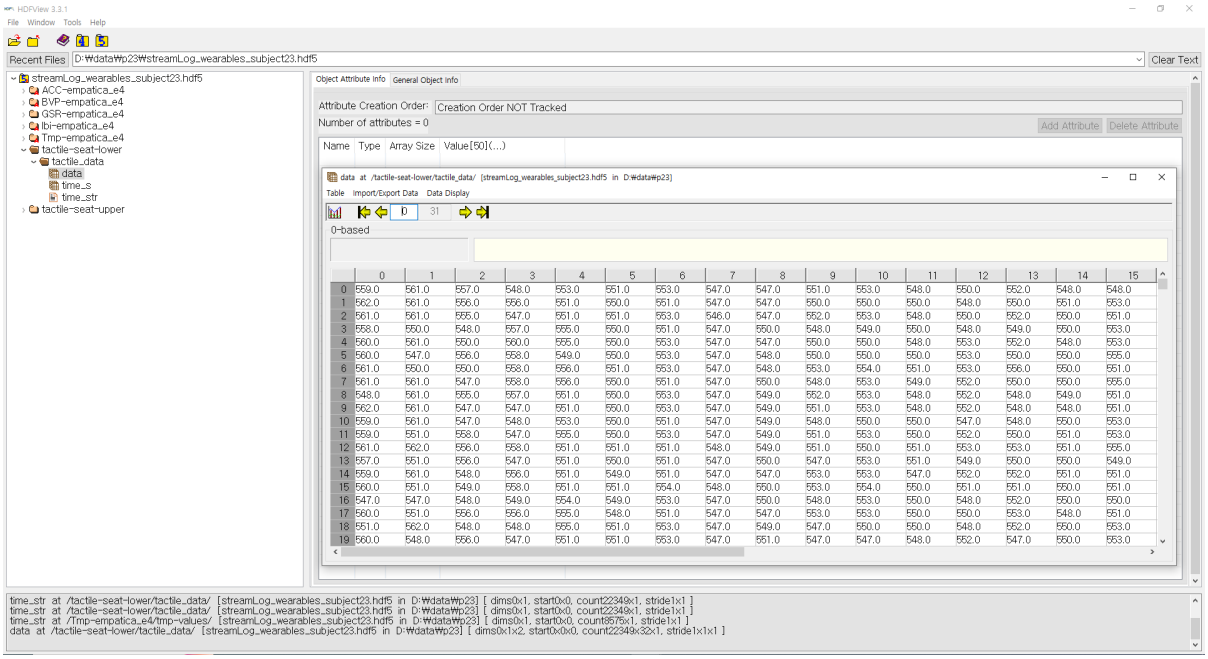


Fig. 26. Example data and formatting of the streamLog_wearables_subject05.hdf5 file.

Fig. 27. Exemplary thermal image captured for participant 23 (1695616151297_heat.png under the directory thermal_imaging_subject23).

used the pretrained models by Kuzdeuov et al. [78] and Abdrakhmanova et al. [3] to extract the facial position and landmarks (Figure 28). The extracted facial recognition results are provided in .csv files for each participant, named subjectID_thermal_heat.csv.

B.4.4 Skeletal Tracking and Facial Meshes. We provide the detected skeletal tracking and facial meshes in JSON file format, named UNIXTIME_d435.json. The joints and landmarks were extracted using MediaPipie framework and follow its data protocol [87]. For body_joints, it stores a list of dictionaries, each representing a joint with x, y, z, and visibility keys. For facial_landmarks, it contains a list of lists (one per detected face), where each list includes dictionaries for individual landmarks with x, y, and z coordinates.

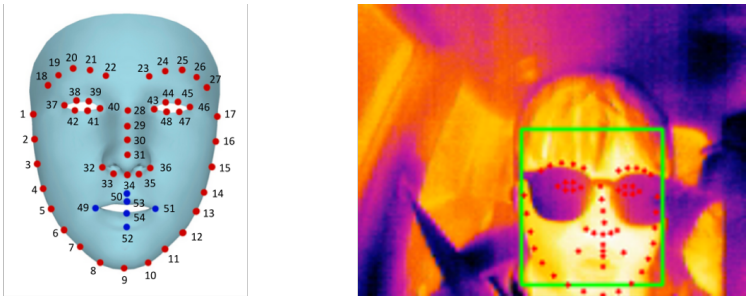


Fig. 28. Face and facial landmark detection for thermal image captured for participant 23 (1695616151297_heat.png under the directory thermal_imaging_subject23). (a) numbering convention of the landmarks [3, 78], (b) exemplary result for our dataset

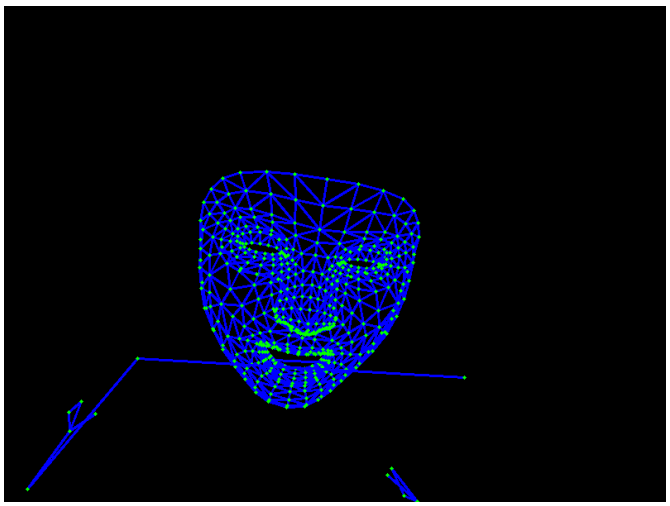


Fig. 29. Example facial landmark JSON file (1695616083780_d435.json of subject 27) plotted as meshes.

- **width**: Specifies the image's width in pixels. This parameter is used for converting the normalized x-coordinates of landmarks back to their absolute pixel positions on the image.
- **height**: Specifies the height of the image in pixels. Similar to the width, this parameter is crucial for converting normalized y-coordinates of landmarks to their absolute pixel positions. The height ensures that landmarks can be accurately plotted on the vertical axis of the image, taking into account its original size.
- **body_joints**: Contains an array of detected body joints, where each element is a dictionary detailing the normalized coordinates (x, y, and optionally z) and the visibility score. The 'x' and 'y' coordinates are fractions of the total 'width' and 'height', respectively, meaning that multiplying these normalized values by the image's actual dimensions (width for 'x', height for 'y') will yield the landmark's pixel position on the image.

  - **x**: Normalized horizontal position of the joint, ranging from 0 (left edge) to 1 (right edge).
  - **y**: Normalized vertical position of the joint, ranging from 0 (top edge) to 1 (bottom edge).
  - **z**: Depth information of the joint, relative to the plane of the camera.
  - **visibility**: Likelihood of the joint being visible, with 1 indicating high visibility.
- **facial_landmarks**: MediaPipe face mesh provides 468 3d facial landmarks, covering facial features such as the eyes, eyebrows, nose, ears, mouth, and the overall facial silhouette.
  - **x**, **y**, **z**: An array of arrays, each corresponding to a detected face, with sub-arrays containing dictionaries for each landmark. These landmarks are also specified in normalized coordinates, which can be scaled to pixel positions using the ‘width‘ and ‘height‘ of the image.

B.5 Annotation Tags

B.5.1 Explanation demands. Labeled by the protocol described in section 3, the UNIX timing at which people demanded explanations and the corresponding demanded explanations are provided in CSV file format. We also provided the encoding used to classify the explanation types (in terms of ‘what’ and ‘what+why’ framework, and content-relevant encoding for the Sankey diagram). Figure 31 shows exemplary data and formatting of the Explanation_subject23.csv file.

Timestamps	Explanation
1695615423	The car starts.
1695615487	The car slows down to maintain distance from the front car.
1695615524	The car enters the highway.
1695615596	The car merges onto the road.
1695615667	The car changes lanes to exit the highway.

Fig. 30. Example data and formatting of the obd_subject05.csv file.

B.5.2 NDRT. We included Non-Driving Related Tasks (NDRT) annotations with UNIX timestamps in CSV format. This complementary tags are intended to capture scenarios where explanations might compete for a passenger’s attention, potentially disrupting activities or the delivery of necessary information during NDRTs.

The NDRT data annotation details passenger activities with a UNIX timestamps tagged. Passegner’s NDRT action information labels are as actions such as ‘access and manage the items in and on the dashboard compartment.’, ‘relaxing in the passenger’s seat’, ‘watching outside the window’, ‘wathching inside the car’, ‘drinking water’, ‘eating a snack’, ‘interacting with a mobile phone’, ‘reading a book’, ‘reading a magazine’, ‘reading a paper’, and ‘having a phone call’. An illustration of this data structure and format is provided in Figure 31, showcasing the NDRT_subject23.csv file.

Timestamps	NDRT
1695615265	interacting with a mobile phone
1695615271	watching outside the window
1695615273	access and manage the items in and on the dashboard compartment.
1695615285	reading a magazine
1695615323	watching outside the window

Fig. 31. Example data and formatting of the obd_subject05.csv file.

C Neural Network Layers

Model	Layer-wise Structure
LSTM Dense	LSTM(64, return_sequences=True) Dropout LSTM(64, return_sequences=False) Dense(64) Dropout Dense(1, activation='sigmoid')
Bi-LSTM Dense	Bidirectional(LSTM(32, return_sequences=True)) Dropout Bidirectional(LSTM(32, return_sequences=False)) Dropout Dense(units=128, activation='relu') Dropout Dense(units=64, activation='relu') Dropout Dense(units=1, activation='sigmoid')
Attention-LSTM Dense	LSTM(64, return_sequences=True) Lambda(Attention()) Dropout Flatten() Dense(128, activation='relu') Dropout Dense(1, activation='sigmoid')
Attention Bi-LSTM Dense	Bidirectional(LSTM(64, return_sequences=True)) Lambda(Attention()) Dropout Flatten() Dense(128, activation='relu') Dropout Dense(1, activation='sigmoid')
LSTM-CNN Dense	LSTM(64, return_sequences=True) Dropout Conv1D(filters=32, kernel_size=3, activation='relu') Conv1D(filters=32, kernel_size=3, activation='relu') MaxPooling1D(pool_size=2) Flatten() Dense(128, activation='relu') Dropout Dense(1, activation='sigmoid')
Bi-LSTM-CNN Dense	Bidirectional(LSTM(64, return_sequences=True)) Dropout Conv1D(filters=32, kernel_size=3, activation='relu') Conv1D(filters=32, kernel_size=3, activation='relu') MaxPooling1D(pool_size=2) Flatten() Dense(128, activation='relu') Dropout Dense(1, activation='sigmoid')

Table 8. The structure of the evaluated neural network models