What and When to Explain? On-road Evaluation of Explanations in Highly Automated Vehicles

GWANGBIN KIM, Gwangju Institute of Science and Technology, Republic of Korea
DOHYEON YEO, Gwangju Institute of Science and Technology, Republic of Korea
TAEWOO JO, 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 help passengers understand the vehicle's state and capabilities, leading to increased trust in the technology. Specifically, for passengers of SAE Level 4 and 5 vehicles who are not engaged in the driving process, the enhanced sense of control provided by explanations reduces potential anxieties, enabling them to fully leverage the benefits of automation. To construct explanations that enhance trust and situational awareness without disturbing passengers, we suggest testing with people who ultimately employ such explanations, ideally under real-world driving conditions. In this study, we examined the impact of various visual explanation types (perception, attention, perception+attention) and timing mechanisms (constantly provided or only under risky scenarios) on passenger experience under naturalistic driving scenarios using actual vehicles with mixed-reality support. Our findings indicate that visualizing the vehicle's perception state improves the perceived usability, trust, safety, and situational awareness without adding cognitive burden, even without explaining the underlying causes. We also demonstrate that the traffic risk probability could be used to control the timing of an explanation delivery, particularly when passengers are overwhelmed with information. Our study's on-road evaluation method offers a safe and reliable testing environment and can be easily customized for other AI models and explanation modalities.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; **User studies**; **Mixed / augmented reality**.

Additional Key Words and Phrases: automated vehicles, intelligibility, in-car extended reality

ACM Reference Format:
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 (September 2023), 26 pages. https://doi.org/10.1145/3610886

1 INTRODUCTION

In highly automated driving, drivers are no longer required to take over driving-related tasks. For example, vehicles with SAE Level 4 can operate independently under limited conditions, while those with SAE Level 5 can drive autonomously under all conditions [92]. This allows drivers to engage in non-driving-related tasks (NDRTs), such as relaxing, working, texting, or viewing multimedia content. Despite the ease and convenience of

*Corresponding author.

Authors' addresses: Gwangbin Kim, gwangbin@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; Taewoo Jo, twjioi5349@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.

2474-9567/2023/9-ART104
https://doi.org/10.1145/3610886

automation, some people, particularly first-time users, may be reluctant to embrace automated vehicles due to public anxiety [33]. Thus, for passengers to widely adopt self-driving cars and fully utilize the opportunities for resting or other activities, automated vehicles must be designed to elicit passenger trust and acceptance.

Another source of driver anxiety and lack of trust in and acceptance of automated vehicles is their unanticipated behavior [33, 89]. Regardless of their driving performance, the anxiety is intensified when vehicles behave in a way that the driver did not expect. Explanations play an important role in mitigating negative experiences in automated vehicles by giving the increased feeling of control [76] and helping passengers calibrate their trust level based on the vehicle's actual capabilities [70]. Since vehicles with SAE Level 4 and 5 do not take driver control such as take-over into account, explanations that give an increased feeling of control get particularly important for highly automated vehicles to mitigate driver's demand for take-over [77].

However, explanations that do not meet human expectations can negatively impact passenger experience [29]. Explanations must be carefully crafted to convey information effectively without distracting the passengers [30]. Therefore, several studies have examined the effects of different explanation presentation methods on passenger experiences to design explanations that enhance the passenger experience, such as trust, comfort, and machine acceptance, while minimizing fear, anxiety, and cognitive load [11, 12, 46, 75]. These studies have focused on different design factors, such as information modality or quantity, and have considered various driving scenarios and NDRTs for effective information delivery.

Most previous research on in-vehicle explanations has relied on graphical simulations with fully rendered objects. Although such studies have explored the impact of explanations on passenger experience and the effect of design factors, their results can be strengthened through additional validation in more ecologically valid situations, as demonstrated by the on-road study with a wizard driver by Schneider et al. [77]. Additionally, recent advances in explainable artificial intelligence (XAI) have made it possible to produce explanations regarding the hidden intentions of automated control systems [70], helping to actualize the explanations that were conceptually presented using simulations. Yet, they were rarely tested with human subject experiments though the explanations are ultimately employed by passengers. As we strive to develop safer and more reliable automated vehicles, equal efforts should be dedicated to making these vehicles both trustable and widely accepted through a unified approach between HCI and AI, leveraging rigorous human subject research and actual implementation.

1.1 The Present Study

This study focuses on explanations in highly automated driving with SAE Levels 4 and 5. We aimed to build upon prior research on in-vehicle explanations, which was primarily conducted in laboratory settings, and to validate findings under real-road conditions. We investigated the impact of explanations, focusing on how explanation type and timing influence passenger trust, acceptance, and other experiences. To design a model for describing whether and how explanations facilitate automated vehicle acceptance, we established the following research questions to guide the direction of this research.

- RQ1: What types of explanations and when should they be provided on the road to yield a better passenger experience in highly automated vehicles?
- RQ2: How do explanation types, timing mechanisms, and the resulting passenger experience affect passenger acceptance of automated vehicles?

To answer these questions, we utilized a camera-based AR driving platform designed to simulate self-driving, operated by a human "wizard" driver. We augmented the vehicle's windshield with a windshield display (WSD) that visualized the vehicle's understanding of the driving situation. We first focused on visual explanations, as they provide passive information without requiring drivers to be constantly alert. We presented the vehicle's perception and attention state, the importance of which has been highlighted by Wiegand et al. [90], and varied the timing of the explanations provided, employing a traffic risk classification algorithm to explore the interplay

between explanation type and timing in shaping passenger trust and acceptance of automated vehicles. The contributions of the present study are summarized as follows.

- We build a TCP-based Unity-Python framework to test algorithmic explanations under in-car VR settings.
- We test the impact of perception and attention explanations in actual vehicles and report both quantitative and qualitative passenger experiences, along with empirical findings from a 30-participant experiment.
- We provide a model that illustrates how different types and timing mechanisms of explanations promote the acceptance of automated vehicles.

2 RELATED WORK

2.1 Explanation Research for Automated Vehicles

Public anxiety regarding automated vehicles has led to hesitation in their adoption [50]. Low trust in automated vehicles can also increase worry and lower acceptance [53]. In this context, explanations can help alleviate potential negative experiences and assist people in understanding the capabilities of automated systems. For example, Koo et al. [46] demonstrated that appropriate explanation content could help drivers overcome their anxiety and build trust in automated vehicles. Similarly, explanations provided during or after a ride can mitigate the negative experiences of passengers by offering a greater sense of control [76]. While previous studies highlighted the importance of explanations using simulation environments and high-level contextual information, they were less aligned with current practices used in the development of explainable driving algorithms. Nevertheless, recent studies have shown that explanations more directly tied to the driving states or decisions themselves, such as vehicle perception or path planning information [13, 18], can improve user experience and trust. These attempts suggest that established design considerations for in-vehicle explanations such as explanation content [30], timing[89], modality [46, 75], aesthetics [22], and visualization methods applied [12] can be integrated into real-world environments when combined with proper algorithms to generate explanations.

Explainable AI (XAI) describes the hidden intention behind the decision-making of a model. When applied to automated vehicles, XAI models can be used to design more transparent and trustable automated vehicles by explaining the reasons behind their driving decisions. In addition to accuracy and precision, the success of XAI models depends on various design factors, including the content of the explanation and visualization methods applied. For example, many models use heatmap-based attention visualization to show the attention regions of an image that an algorithm focuses on when making a decision [41, 58], whereas others use textual explanations [42, 43] or other graphical representations such as arrows [94]. Although most XAI models are designed to provide operational or tactical driving explanations, they can also provide other driving-related information, such as accident risk [65]. Wiegand et al. [90] also emphasized the importance of explaining the machine perception itself using mental models of the passengers, such as the state of the vehicle sensors and object detection. Although AI algorithms offer methods for delivering explanations in automated vehicles, relatively limited research has been conducted to assess their effectiveness. Only a few studies have tested these algorithms with human participants and only using videos played on a monitor [68, 69], which is far from an actual riding experience. In addition, most of these algorithms have rarely been implemented on physical platforms. Because human drivers and passengers ultimately use such algorithms, it is important to examine the actual impact of these explanations on the passenger experience in the development of truly helpful explainable AI models.

Based on prior works, we aim to evaluate visual explanations for the perception and attention state of SAE Level 4 and 5 automated vehicles. We first considered visual explanations, as explanations on WSD may not interfere with or alarm passengers when they do not watch it, which may be an important feature for Level 4 and 5 automated vehicles where passengers do not have to maintain the full situational awareness required for driving. Among visual explanations, we tested perception information to further validate its effectiveness under laboratory conditions [11, 12], and considering the mental model presented by Wiegand et al. [90]. Additionally,

we included saliency-based attention information given that their impact was rarely tested with humans despite its direct relation with the driving decisions made by neural network-based algorithms. Among different types of attention, we presented the vehicle's attention in predicting traffic risk as explaining traffic risk is important in reducing discomfort in automated vehicles [28].

2.2 Driving Simulators for Automotive UI/UX Research

Automotive user interfaces for manually driven vehicles have traditionally focused on promoting safe driving behavior. Driving simulators, which demonstrate behavioral validity by reproducing similar driver performance patterns to real-world conditions, such as speed maintenance [26, 95] and lane-keeping behavior [72], have been widely used for designing and testing these interfaces. However, as the automotive industry transitions to automated vehicles, particularly those with Society of Automotive Engineers (SAE) levels 4 and 5, the emphasis has shifted towards building a satisfactory passenger experience in terms of trust and acceptance [84]. Since driving simulators in laboratories are inherently safe, concerns have been raised that these simulators may not perfectly mimic the experience of being in automated vehicles. Hock et al. [35] specifically highlight the potential impact on trust measurement, noting: 'the inherently safe environment may influence measurements of trust in automation [21]', and 'participants who are more immersed may experience a more realistic feeling of trust [74]'.

One possible solution to this is conducting an outdoor experiment on actual roads. While simulators provide immersive and reproducible testing environments, evaluating interfaces in real-road settings can yield more ecologically valid results, as all automotive interfaces are ultimately integrated with actual vehicles on the roads. By applying the Wizard-of-Oz paradigm [15] and hiding the wizard driver under the seat [73] or behind a partition [2, 83, 88], actual vehicles can be transformed into automated driving simulators to test the experiences of passengers and pedestrians without safety or ethical concerns. The wizard driver in on-road simulators can also be hidden by connecting the physical system with a virtual [27, 34, 61, 63] or augmented reality environment [63, 96, 97], where participants cannot see the driver. These platforms are particularly effective for testing advanced interfaces, as they allow the augmentation of automotive UI/UX services in real-road testing environments.

On-road environments, despite their inability to simulate accident-critical scenarios, can provide a more realistic experience of traffic risks than inherently safe indoor environments. Meanwhile, risk, significantly impacts automated driving experiences and attitudes towards explanations. As the risk level increases, reliance on automation requires a higher degree of driver trust, particularly during initial interactions [85]. Consequently, passengers' experiences with explanations in automated vehicles can vary depending on traffic risk levels [30, 55]. Recent research by Goldman and Bustin [28] even emphasizes the importance of explaining the risk scenario itself in reducing passenger discomfort.

In this study, we tested visual explanations using actual vehicles under real-road conditions, thereby leveraging the two benefits of on-road experimentation: a naturalistic driving scenario and a realistic experience of traffic risk. In particular, we explored the presentation of vehicle-interpreted risky areas as part of attention information and used the vehicle's interpretation of traffic risk as a means to determine the timing of explanation delivery.

3 ON-ROAD EXPLANATION TEST METHOD

3.1 In-car Extended Reality for Explanation Visualization

Our system extends the on-road platform MAXIM [96, 97] and adopts the Wizard-of-Oz method [15] for exploring self-driving scenarios without safety issues or ethical concerns. The vehicle was driven by a "wizard" driver placed in the driver's seat, and the study participant sat in the front passenger seat while wearing a VR head-mounted display (HMD) (Figure 1). We used a Varjo VR-2 device for our system (1440x1600 per-eye resolution, 87° horizontal FoV, 90Hz). The participant is shown to be sitting in the driver's seat in an extended reality environment, developed using the Unity 3D framework, in which the driver is removed, and the 360° streaming

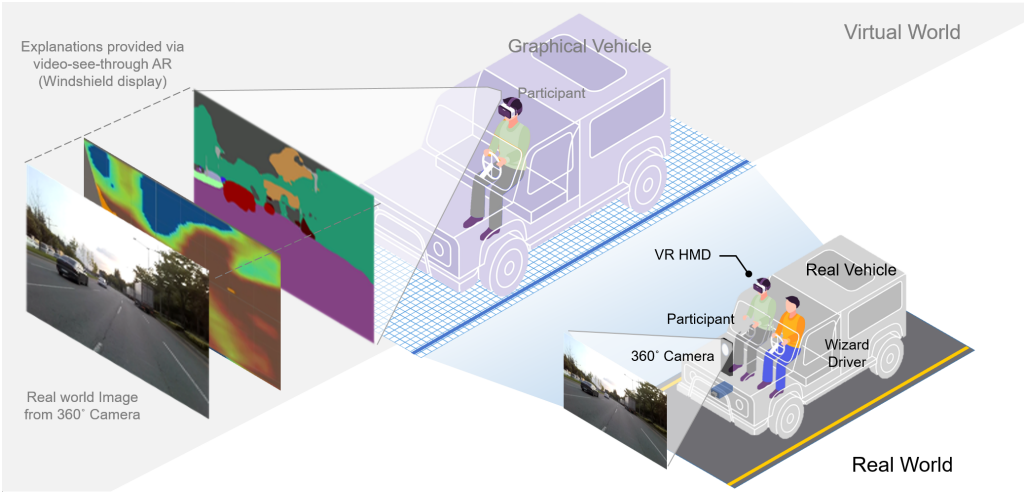


Fig. 1. Schematic overview of on-road explanation assessment framework.

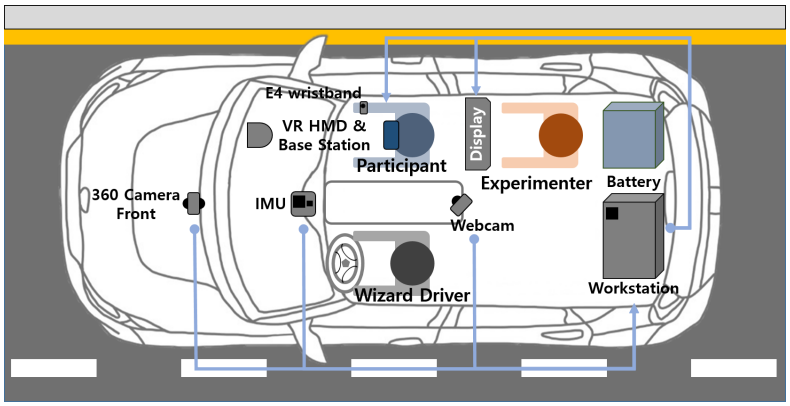
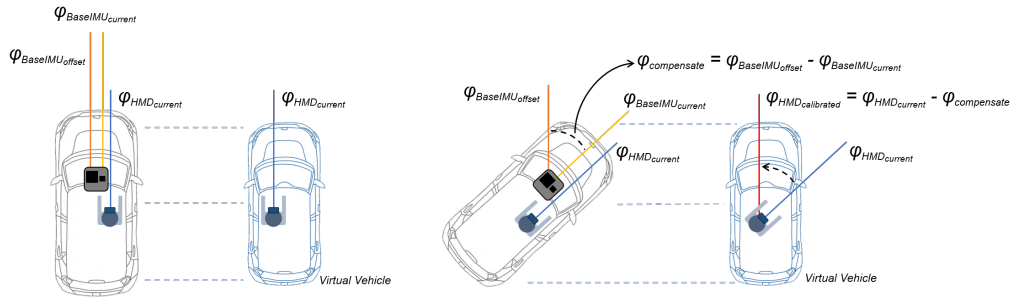


Fig. 2. Implementation diagram for on-road explanation assessment platform.

camera image surrounds the graphical model of a vehicle to form the Wizard-of-Oz-based self-driving experience (Figure 2). Because the participant sees a graphically rendered vehicle, the user interfaces of the vehicle can be easily augmented through extended reality. Based on previous studies' reports on the strengths of WSDs for information delivery in automated vehicles [10–12], we set a simulated WSD as a method for providing visual explanations in vehicles. Since the surrounding image and WSD are streamed independently, any delay from the explanation algorithms does not impact the overall simulator experience. Because the video see-through MR environment viewed by the participant is identical to the video being fed into the machine-learning algorithms generating the explanations, the platform enables a contact analog registration [64].

The use of extended reality technologies in moving vehicles poses a tracking challenge for HMDs. The base station used to track the HMD is incompatible with a moving platform, and the IMU embedded in the HMD does not distinguish between the motions generated by the user and those of the vehicle [60, 62]. To address this issue, we constrained the translational movement of the HMD and calibrated its horizontal rotation to reflect

the rotation of the user in the reference frame of the vehicle only. Specifically, we set a base IMU to track the orientation of the vehicle to distinguish its rotation from that of the HMD in a global reference frame. Before the experiment began, the horizontal orientation of the vehicle was captured by an IMU sensor placed in the vehicle to set the offset for calibration (Figure 3 (a)). During the drive, the HMD of the user was calibrated using a compensation angle, which is the difference of the current orientation of the vehicle from the IMU offset (Figure 3 (b)). To correct for the accumulated IMU drift, the base station of the vehicle recalibrated the orientation of the HMD when the vehicle stopped for a particular amount of time, such as waiting at a traffic light (Algorithm 1).



(a) Registration of the offset for calibration (b) Calibration during the drive

Fig. 3. Calibration process for horizontal rotation of the HMD of the user in a moving vehicle

Algorithm 1 Calibration Process for Horizontal Rotation of the HMD

$\varphi_{BaseIMU_{offset}} \leftarrow \varphi_{BaseIMU_{current}}$
while the car is driving **do**
  $\varphi_{compensate} \leftarrow \varphi_{BaseIMU_{offset}} - \varphi_{BaseIMU_{current}}$
  $\varphi_{HMD_{calibrated}} \leftarrow \varphi_{HMD_{current}} - \varphi_{compensate}$
  if $\sum_t acc < acc_{threshold}$ **then**
    $\varphi_{compensate} \leftarrow 0$
  end if
end while

3.2 Explanation Algorithms

Figure 4 provides an overview of the algorithms used to generate the *in situ* explanations. The front view of the vehicle is captured in Unity 3D and then sent to a Python environment where machine-learning algorithms generate visual explanations. The outcomes are returned back to the Unity 3D to be visualized in a mixed-reality environment. In the Python part, the streamed front view undergoes two parallel processes: 1) semantic segmentation, which is a part of the perception state of the vehicle and 2) a 3D CNN for traffic risk prediction with Grad-CAM, which is used both for a visual explanation for attention and a means to modulate the explanation timing. Depending on the experimental conditions, the two types of explanations were provided separately or together to form three explanation conditions to be tested (perception, attention, and perception+attention).

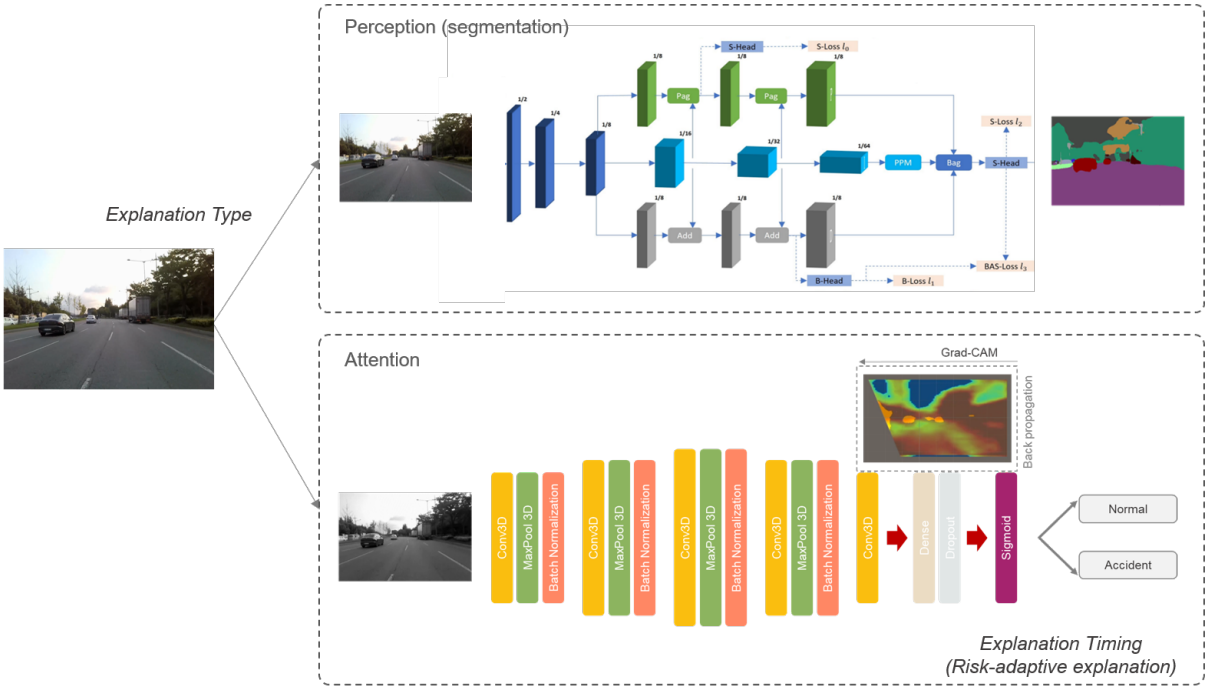
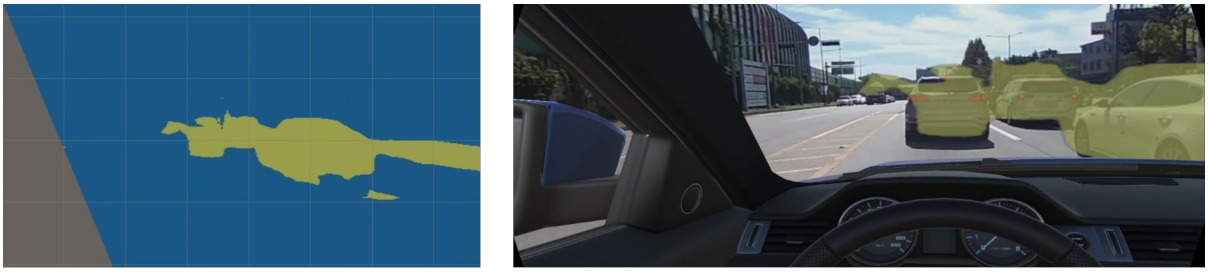


Fig. 4. Overview of algorithms used in our study to provide an explanation.

3.2.1 Perception (Segmentation). As Wiegand et al. [90] noted regarding the significance of explaining machine perception, we presented passengers with a semantic segmentation map as part of the vehicle's perception state. In a video-based experiment, the segmentation information provided in automated vehicles increased driver situation awareness [11]. To offer semantic segmentation by superimposing segmented objects on the WSD, we incorporated PDINet [93] to represent the state of the machine perception of the vehicle. A PIDNet-S model trained on the Cityscapes dataset with a small number of parameters [14] demonstrated an mIOU of 78.6% and FPS of 93.2. We assigned the yellow color to the labels car, truck, bus, motorcycle, bicycle, caravan, trailer, person, and rider while omitting other classes such as the sky, road, sidewalk, building, vegetation, parking, traffic signs, and traffic lights (see Figure 5). The color coding and removal of classes were intended to provide an adequate amount of information, preventing passengers from being visually overloaded. Also, we aimed to avoid the need for the passengers to decipher the meaning of each color, which may cause additional cognitive load.

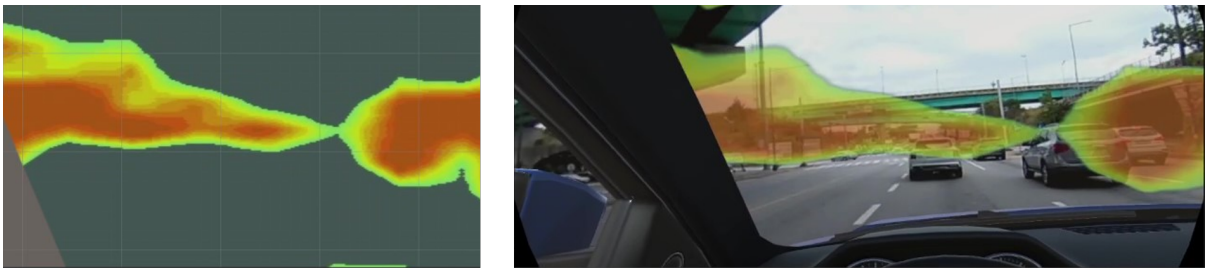
3.2.2 Traffic Risk Prediction (Attention/Explanation Timing). We developed a custom 3D CNN model to predict the traffic risk probability (Figure 4, lower), which was used to control the explanation timing. The 3D CNN model was designed to take the image volume with a width of eight and classify whether a video contained an accident. We trained the model with the Car Crash Dataset [3], resulting in a 93.74% validation accuracy. We considered the sigmoid output of the classification model as the vehicle's interpretation of the in-situ traffic risk, providing explanations when the probability was greater than .5 for conditions with risk-adaptive explanations.

Also, the Grad-CAM [80] was used to visualize the vehicle's attention, showing what prompted it to determine whether a given driving scenario was hazardous. The Grad-CAM was designed to compute the back-propagated



(a) (b)

Fig. 5. (a) Color-coded semantic segmentation map and (b) segmentation map overlaid on WSD.



(a) (b)

Fig. 6. (a) Color-coded Grad-CAM attention map and (b) attention map overlaid on the augmented reality head-up display.

weight up to the final CNN and generate the class activation map. The Grad-CAM was color-coded, omitting low saliency regions (<.8) to prevent information overload due to heatmap covering the entire WSD (Figure 6).

3.3 Sanity Test for ML-predicted Risk as a Predictor of Passenger Arousal

We conducted a pilot study to investigate the relationship between traffic risk probability and passenger arousal, aiming to explore the potential of traffic risk as an unobtrusive, indirect predictor of passenger arousal. Physiological signals, such as electrodermal activity (EDA) or pupil diameter, change in response to the cognitive demands or arousal that a task might induce (e.g., event-related EDA [16, 56], task-evoked pupillary responses (TEPR) [20]). Our study focused on these physiological responses to varying levels of traffic risks while in automated vehicles rather than responses to specific tasks. We did not differentiate between sources of arousal, which may be cognitive load, fear, anxiety, or demand for situational awareness, but assessed how external risks, as quantified by the risk prediction algorithm, influenced the arousal of the passenger watching the environment. The Granger-causality test was applied to determine whether traffic risk probability could significantly predict the EDA and pupillary responses, indicators of passenger arousal. The subsequent subsections delve into further details of the pilot experiment.

3.3.1 Pilot Experiment Settings. We recruited five participants with an average age of 24.6 years (SD = 0.89, 2 Females) for our study. The participants were seated in front of a screen with Shimmer3 sensors attached to their index and middle fingers, measuring their electrodermal activity at a frequency of 16Hz. A Tobii Pro X2-60 eye-tracker was also set below the screen to capture the participants' gaze activities and pupil dilation at 60Hz.

While factors such as lighting conditions can influence pupil dilation, the experiment was conducted indoors, ensuring consistent illumination.

Initially, the participants observed nine dot stimuli for eye-tracking calibration. They then watched a ten-minute nature relaxation video to stabilize the EDA signal and establish a baseline for pupil dilation. Following this, participants were asked to watch a 15-minute recording of naturalistic driving. Participants were tasked to view the video as if they were passengers in automated vehicles, focusing on the overall experience rather than annotating each specific traffic event.

3.3.2 Result and Analysis. We conducted a Granger-causality test with the predicted traffic risk and the recorded physiological responses. The Granger-causality test evaluates whether the 'effect' variable is influenced by past and present values of the 'cause' variable. Similar methodologies have been used in studies by Lavanuru et al. [51] and Ghouali et al., [25], investigating the causality between physiological response and perceived workload, and between cardiorespiratory and myogalvanic signals during driving tasks, respectively. As we considered the task of watching the video, which simulated the experience of riding in automated vehicles, to be a holistic experience rather than tasks with time-specific events, we analyzed the physiological responses recorded over the entire 15-minute duration of the naturalistic drive.

The results of our Granger-causality analysis indicate that traffic risk probability can significantly predict the EDA signal (Table 1). Although traffic risk probability was not a significant predictor of pupil dilation for some participants, it consistently served as a significant predictor of the passengers' EDA signal. Since an increase in EDA is indicative of heightened arousal states—including stress, workload, and anxiety [56, 57]—our findings suggest that traffic risk probability could potentially be used to predict moments of passenger arousal due to external risks when riding automated vehicles.

Table 1. Results of the Granger-causality test for risk probability relative to the EDA signal and pupil dilation.

Participant ID	Electrodermal activity		Pupil dilation	
	F	p	F	p
#1	4.920	0.0266*	0.622	0.430
#2	4.378	0.0364*	1.747	0.186
#3	3.407	0.0086**	2.685	0.0676
#4	3.883	0.0488*	8.847	0.0029**
#5	4.413	0.0121*	0.283	0.595

3.4 Experimental Conditions

The conditions comprise seven distinct conditions, including the default condition without an explanation (condition 1), as well as three explanation types (perception, attention, and perception + attention) and two explanation timings (continuously and only when it is risky) (see Figure 7).

3.5 Implementation Note

We implemented the proposed system on a computer with an Intel® i9@2.50 GHz CPU, 128 GB of RAM, and an RTX 3090. We used the Unity 3D environment to provide the participants with an extended reality environment guaranteed to render at least 30 frames per second (fps). The Python side computed the segmentation and attention map with a framerate of greater than 15 fps. Because both sides transfer images through TCP socket communication, any visual explanation can be added to our system with an appropriate communication configuration. The detailed system implementation, including the TCP-based communication framework, is available at https://github.com/GWANGBIN/WW2E.

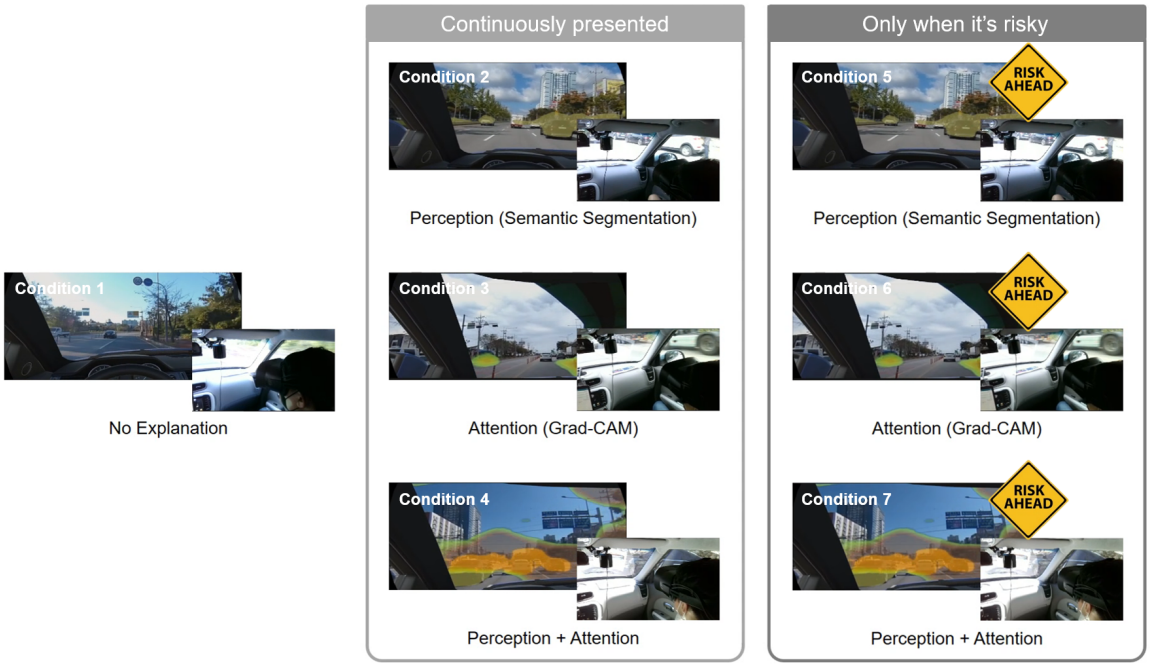


Fig. 7. Seven experimental conditions for the user study. Three explanation types (perception, attention, and perception + attention) and two explanation timings (continuously presented (always) and only when it is risky (if risky)) were tested.

4 USER STUDY

We conducted a user study to compare the passenger experience when algorithmic explanations of automated vehicles' perception and attention state were provided. To ensure ecologically valid experimental settings, the study was conducted on actual roads with a wizard experimenter under a naturalistic driving scenario. We investigated usability, trust, perceived safety, situational awareness, cognitive load, preference, and other factors associated with the acceptance of automated vehicles by exposing participants to a variety of explanation settings. In addition, as indications of arousal, we assessed the physiological responses of participants during the ride.

4.1 Participants

We recruited 30 participants (8 females) with an average age of 28.4 (SD = 8.34, min = 19, max = 50). 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, 23 had driver's licenses, with an average of 6.72 years of driving experience (SD = 7.43, min = 1, max = 30). All participants were Korean nationals, and the user study experiment was approved by the Institutional Review Board.

4.2 Procedure

The user study was conducted with the following experimental protocol and driving scenarios.

4.2.1 Protocol. Initially, the participants were instructed about the experiment and wore an E4 wristband. We opted for the E4 over the Shimmer3, used in our preliminary test, as it offered a firmer body attachment and allowed for multi-modal physiological response measurements (PPG sampled at 64Hz and EDA sampled at 4Hz).

Participants then filled out questionnaires regarding their age, driving experience, and trust propensity. Following this, they experienced 8–12 min of naturalistic driving in ascending order along the route shown in Figure 8.

During the ride, participants wore a Varjo VR-2 HMD and were instructed to behave like they were passengers in highly automated vehicles without needing to control the vehicle. Though a wizard driver controlled the vehicle during the experiment, we informed participants that this driver was present primarily for safety and regulatory reasons and would only intervene with the vehicle's operations at the beginning or end of the experiment or to handle specific experimental scenarios. This was done to prevent the participants from perceiving the vehicle as non-automated during the experiment, despite any subtle auditory cues that the wizard driver might have produced. We view that the presence of the 360° camera and the machine-generated explanations provided to the passengers also helped the deception that they were in an automated vehicle.

Participants were exposed to seven different explanation conditions. These included the default condition without an explanation and three explanation types (perception, attention, and both) for each of the two explanation timings (continuously provided and provided only when conditions are evaluated as risky). Using a balanced Latin square, the explanation condition was counterbalanced to ensure that the driving route and order of the experimental conditions did not influence the results. After each condition, participants provided their responses. The experiment concluded with a semi-structured interview in which participants numerically rated their preferences, explained their reasoning, and suggested improvements. On average, the entire experiment took approximately 2 hours and 30 minutes per participant. We informed participants they could halt the experiment if they experienced discomfort from motion sickness, yet no such requests were made during the study.

4.2.2 Driving Scenarios. To ensure naturalistic experiments and maintain external validity concerning road types, we diversified the types of roads within the given experimental site. These were counterbalanced over experimental conditions. Routes #1 and #7 are urban roads, each with a length of 2.4km, a speed limit of 60km/h, and consisting of 12 crosswalks (10 equipped with traffic lights). Routes #2(#6) and #5 are arterial roads, each 3.1km long with a speed limit of 70km/h and 6 crosswalks with traffic lights. Routes #3 and #4 are local highways, 2.9km long, with a speed limit of 80km/h and 10 traffic-light controlled crosswalks. The experiments were conducted from 9 am to 6 pm to account for varying traffic volumes while ensuring ample light for the 360° camera.

We evaluated the average percentage of situations that were classified as 'risky' by our algorithm in each driving scenario, using post-experiment 360° video recordings. The average proportion of risky situations in each scenario was as follows: Route1: M=3.79%, SD=2.07, Route2: M=2.63, SD=2.04, Route3: M=4.98, SD=2.13, Route4: ?M=3.08, SD=1.95, Route5: M=3.28, SD=1.87, Route6: M=2.87, SD=1.86, Route7: M=3.72, SD=1.97. Statistical analysis revealed that the difference in proportion among driving routes was not significant, $F(6) = 0.912$, $p = .305$.

4.2.3 Automation Wizard. Rather than providing an experimental protocol to various drivers, one of the authors (a 30-year-old male with 10 years of driving experience) served as the automation wizard, fully understanding the study objectives (we referred to [17]). The wizard driver was instructed to cautiously follow the designated route, maintain 50-80% of the speed limit and avoid abrupt lane changes, sudden acceleration, or deceleration. However, responses to unpredictable road events, such as reducing speed for an inappropriately overtaking vehicle, were acknowledged as inevitable.

4.3 Measurement

We collected questionnaires, interviews, and physiological responses to triangulate each method's results.

4.3.1 Questionnaire. Usability was tested based on the system usability scale (SUS) [7]. SUS evaluates the usability of a system with 10 questionnaire items using a 1–5 Likert scale, transformed into a 0–4 scale for a total of 100 points. A system is considered to have acceptable (above average) usability when its SUS score is greater than 68 [54]. Passenger *trust* towards automated vehicles was assessed using the scale of trust in automated

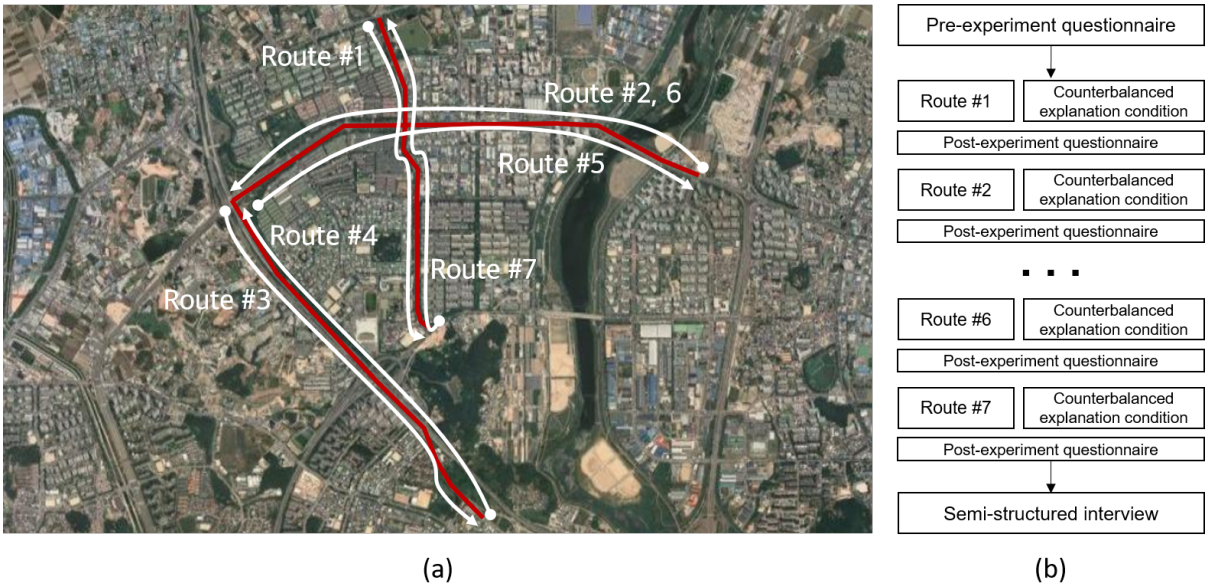


Fig. 8. (a) Driving routes and (b) experimental protocol of our user study experiment.

systems, which comprises 12 questionnaire items [37]. *Situational awareness* was assessed using the situational awareness rating technique (SART) with a 7-point Likert scale [86], and *cognitive load* was measured using the mental demand item of NASA-TLX [31] based on a 0–10 point scale. While most of the measures were adapted from questionnaires with confirmed reliability and validity, the mental demand item is a single-item subscale of NASA-TLX. We intended to capture the immediate mental demand with a minimized item right after the experiment, before participants responded to the detailed experience, to exclude the effect of the cognitive load of answering the survey itself. However, the limited reliability of a single-item questionnaire should also be noted.

We also measured the dependence, understandability, familiarity, and propensity to trust as a way to model the acceptance of automated vehicles in terms of the Reliability/Competence, Understanding/Predictability, Familiarity, and Propensity to Trust subscales based on Q1–6, Q7–10, Q11–12, and Q15–17 of the trust in automation scale provided by Körber [47], respectively. Attitudes towards technology, self-efficacy, anxiety, willingness (behavioral intention), and *perceived safety*, each of which was also used to form our acceptance model, were measured using the Attitude Towards (Using) Technology, Self-Efficacy, Anxiety, Behavioral Intention (to use the Vehicle), and Perceived Safety subscales of the AVAM questionnaire [33], i.e., Q13–15, Q16–18, Q19–21, Q22–23, and Q24–26, respectively.

4.3.2 Physiological Response. Using the E4 wristband, we measured the participants' physiological responses to triangulate the results of the questionnaires and interviews. The measurements included body temperature, heart rate (HR), and electrodermal activity (EDA). The participant's heart rates were analyzed as they were calculated from the PPG signal. The EDA signal was preprocessed by omitting data with values of less than .05, smoothed with a Gaussian window having a width of 8, leaving repeatedly measured sample frames of 23 participants. We then categorized the EDA signal into phasic and tonic components using MATLAB-based ledalab software [5, 6].

4.4 Results

The subsections below describe the results for each aspect of the passenger experience. Descriptive statistics for the survey result are given in Table 2. All measurements underwent skewness and Kurtosis normality check and were compared between conditions using a two-way repeated analysis of variance and Holm *post-hoc* analysis (Figure 9). We also checked the internal reliability for each questionnaire; all questionnaires showed valid internal consistency with Cronbach's alpha higher than the acceptable range of 0.7 [78].

Table 2. Descriptive statistics for the survey results from study participants are shown. Bold highlights the best case, while underline indicates conditions higher than the baseline. For most measures, a higher value represents a better experience, but for cognitive load and preference, a lower value denotes a better result.

Question	Cond.1	Cond.2	Cond.3	Cond.4	Cond.5	Cond.6	Cond.7	Cronbach's α
	Baseline No exp	Perception Always	Attention Always	Per+Att Always	Perception if risky	Attention if risky	Per+Att if risky	
Value	*Mean SD*	*Mean SD*	*Mean SD*	*Mean SD*	*Mean SD*	*Mean SD*	*Mean SD*	α
SUS	71.0 15.8	**78.7** 9.93	66.7 15.3	67.8 15.8	74.3 11.6	66.7 15.5	71.4 12.2	0.785
Trust	3.28 1.10	**3.60** 0.80	2.95 1.13	3.28 1.07	3.18 1.04	2.83 1.00	3.52 1.02	0.874
SART	5.08 .075	5.24 0.69	4.72 0.87	4.94 1.05	**5.27** 0.76	4.89 0.86	5.17 0.75	0.775
SART-U	5.32 1.15	5.00 1.25	4.44 1.39	4.91 1.29	**5.44** 0.94	5.06 1.41	4.86 1.13	
SART-S	5.30 0.66	**5.46** 0.83	4.99 0.91	5.08 1.26	5.37 0.96	5.19 0.95	5.41 0.84	
SART-D	4.56 1.14	**5.20** 0.76	4.64 0.96	4.78 1.15	4.97 0.89	4.33 0.90	5.17 0.99	
Depend	3.29 0.94	**3.67** 0.66	2.97 0.89	3.19 0.97	3.22 0.86	2.94 0.88	3.66 0.87	0.924
Understand	3.37 1.13	**3.84** 0.80	3.21 0.97	3.44 0.87	3.39 0.85	3.23 1.00	3.66 0.94	0.809
Familiar	**2.67** 1.09	2.58 0.98	2.40 1.00	2.30 0.99	2.40 0.96	2.48 1.04	2.37 1.00	0.728
Attitude	3.22 1.00	**3.44** 0.89	2.94 1.14	3.30 1.03	3.02 1.03	2.92 1.15	3.43 0.99	0.892
Efficacy	3.80 0.97	**3.97** 0.66	3.47 1.03	3.71 0.90	3.78 0.78	3.62 0.71	3.89 0.84	0.840
Anxiety	3.36 1.09	**3.58** 0.91	2.88 0.95	3.19 0.91	3.24 0.87	2.98 1.01	3.46 0.96	0.846
Intention	3.45 1.06	**3.57** 1.07	2.88 1.19	3.18 1.16	3.10 1.21	2.93 1.15	3.35 1.03	0.920
Safety	3.41 0.94	**3.68** 0.81	3.12 0.98	3.33 1.03	3.28 0.91	3.02 0.85	3.53 0.83	0.899
Cog. Load	6.00 2.73	**5.13** 2.16	6.60 2.19	5.97 2.33	5.50 2.32	6.20 2.28	5.73 2.15	-
Pref.Rank	4.50 2.17	**2.97** 1.88	4.77 1.79	3.67 1.71	3.57 2.03	4.50 1.55	4.03 2.34	-

4.4.1 Usability. The SUS scores varied significantly depending on the explanation type $F(2) = 8.755$, $p < .001$. Perception information provided higher usability than attention information or the combination of both, i.e., $t_{Holm} = 4.075$, $p < 0.001$ and $t_{Holm} = 2.861$, $p = .012$, respectively. Specifically, the perception information being continuously presented (condition 2) was perceived to have the highest usability of 76.7 ($SD = 9.93$), which was significantly higher than those of conditions 3 (attention, Always), 4 (perception + attention, Always), and 6 (attention, if risky), with $t_{Holm} = 3.960$, $p = .002$; $t_{Holm} = 3.602$, $p = .006$; and $t_{Holm} = 3.999$, $p = .002$, respectively.

4.4.2 Trust. Perception and perception + attention information also yielded higher trust over attention information, i.e., $F(2) = 8.487$, $p < .001$; $t_{Holm} = 3.538$, $p = .002$; and $t_{Holm} = 3.597$, $p = .002$, respectively. The interaction effect between the explanation type and timing indicates that modulating the explanation timing with the risk prediction algorithm is particularly effective in promoting trust towards perception+attention explanation type ($F(2) = 3.443$, $p = .039$).

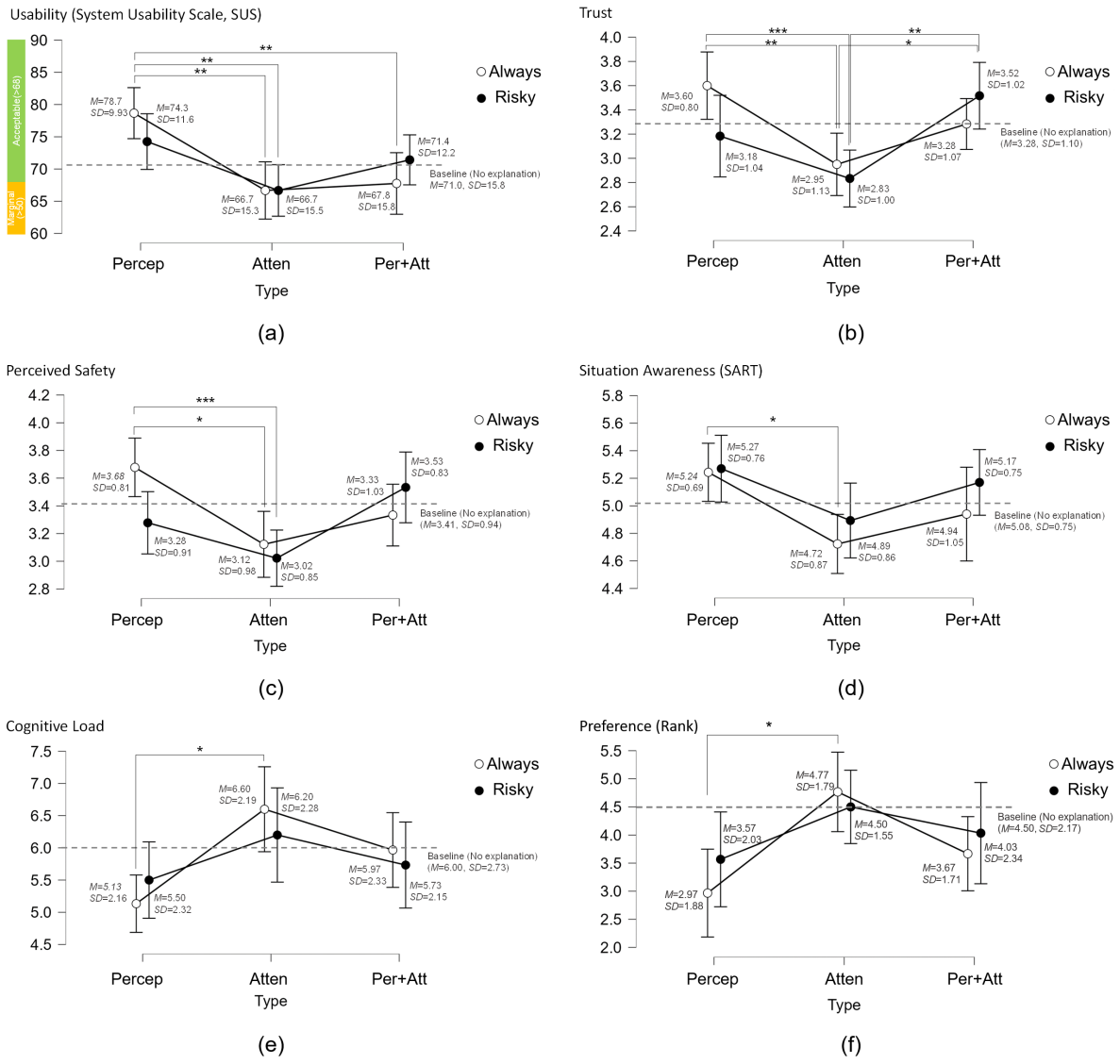


Fig. 9. Results of the user study comparing different types and timings of explanations. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Upon comparing each condition, it was found that condition 2 (perception, Always) also yielded the highest trust, which was significantly higher than that of conditions 3 (attention, Always) and 6 (attention, if risky), i.e., $t_{Holm} = 3.458$, $p = .010$ and $t_{Holm} = 4.078$, $p = .001$, respectively. Condition 7 (perception + attention, if risky) ranked second-highest and was trusted significantly more than conditions 3 (attention, continuously) and 6 (attention, if risky), i.e., $t_{Holm} = 3.014$, $p = .038$ and $t_{Holm} = 3.635$, $p = .006$, respectively.

Some participants trusted the vehicle more without explanations, as failure cases had a greater impact on their trust than the vehicle's abilities. P25 and P29 mentioned that imperfect information led to distrust, while P7 and P21 felt anxious or distrusted the vehicle when its perception was not perfect. Participants also noted specific failure cases beyond the experimental vehicle's design capabilities, such as not looking at traffic lights (P11) or not checking the left and right sides of the car (P15).

4.4.3 Perceived Safety. Perception and perception + attention information were preferred over attention information in terms of perceived safety, i.e., $F(2) = 6.819$, $p = .002$, with $t_{Holm} = 3.367$, $p = .004$ and $t_{Holm} = 2.998$, $p = .008$. The interaction effect shows that the risk-adaptive explanations enhance the perceived safety only when combined with the perception + attention explanation, i.e., $F(2) = 3.951$, $p = .025$.

Specifically, explanation condition 2 (perception, Always) was perceived to be the safest, significantly more so than conditions 3 (attention, Always) and 6 (attention, if risky), with $t_{Holm} = 3.452$, $p = .011$ and $t_{Holm} = 4.173$, $p < .001$, respectively. Also, explanation condition 7 (perception + attention, if risky) was rated the second safest, and only these two conditions ranked higher than the baseline condition without an explanation.

In most cases, risk-adaptive explanations resulted in adverse effects on perceived safety. This is because the moment individuals experience a driving hazard does not necessarily correspond with the algorithmic decisions. For example, participants expressed concerns when the vehicle's judgment of a traffic hazard differed from their perspective as drivers *"The vehicle's judgment of a traffic hazard differed from my perspective as a driver,"* (P16, P25, P29) and when explanation timings were irrelevant *"Some of the explanation timings were irrelevant; they were offered notwithstanding the actual risk."* (P19, P30). They also felt less safe when the vehicle did not provide an explanation despite the imminent danger, fearing it wouldn't handle the issue appropriately.

4.4.4 Situational Awareness (SART). Providing perception information resulted in higher situational awareness than attention information, i.e., $F(2) = 5.885$, $p = .005$ with $t_{Holm} = 3.425$, $p = .003$. Explanation condition 5 (perception, if risky) promoted the highest situational awareness, followed by condition 2 (perception, Always). Condition 7 (perception + attention, if risky) supported the third highest situational awareness, and the other three conditions were not superior to the baseline condition without an explanation. Overall, risk-adaptive explanations supported higher situational awareness compared to continuous presentation. Participants appreciated selective information delivery in high-risk scenarios (P20, P22, P24), but some found the abrupt appearance of information disruptive (P14). A mismatch between perceptions of risk and risky driving conditions contributed to negative experiences with risk-adaptive explanations.

Regarding the SART subscales, the demand subscale varied significantly depending on the explanation type, $F(2) = 8.237$, $p < .001$. Conditions with perception information scored significantly higher than those with attention information, with $t_{Holm} = 3.817$, $p < .001$.

4.4.5 Cognitive Load. Providing perception information resulted in the lowest cognitive load, i.e., $F(2) = 5.120$, $p = .009$. By contrast, attention information exhibited the highest cognitive load, significantly higher than perception information, with $t_{Holm} = 3.328$ and $p = .018$, and higher than the default condition, albeit without a statistically significant difference.

The implicit nature of attentional information, which required a deliberate interpretation process, led to increased cognitive performance, as observed in the interviews. Participants found it difficult to understand the attention information (P10) and noted that perception (segmentation) information was more direct, while attention information needed interpretation (P18, P26). They also had to interpret why the vehicle paid attention to specific areas (P30).

4.4.6 Preference (Rank). The passenger experience also created a different preference among the explanation options, i.e., $F(2) = 5.607$, $p = .006$. Provisioning of the perception information was preferred, i.e., $t_{Holm} = 3.337$, $p = .004$ (the measure is the preference rank, and thus the rank is high for low values). Explanation condition

2 (Perception, Always) ranked the highest (M = 2.97, SD = 1.88) and was significantly higher than condition 3 (attention, Always), i.e., t_{Holm} = 3.284, p = .020. Conditions 4, 5, and 7 were less favored than the default condition without an explanation.

4.4.7 Physiological Responses. Although most of the signal was statistically insignificant, the phasic EDA was significantly different among the conditions, i.e., $F(6)$ = 2.232, p = .044 (Figure 10). Specifically, the Holm *post-hoc* analysis results showed that the phasic EDA for condition 4 in which the perception and attention were continuously displayed (M = .164, SD = .198) was significantly higher than that for the condition 1, default without any explanations (M = .071, SD = .110), with t_{Holm} = 3.104, p = .049.

Li et al. [56] have reported that phasic EDA, which is also referred to as the skin conductance response, is most significantly responsive to cognitive load. Our results indicate that condition 4, which presents perception + attention information constantly induces the highest cognitive load, consistent with the highest self-reported cognitive load. In addition, the insignificance of the physiological responses between conditions 1 (default) and 7 (perception + attention, if risky) also suggests that such arousal can be abbreviated by delivering the explanation only under a driving scenario evaluated as hazardous.

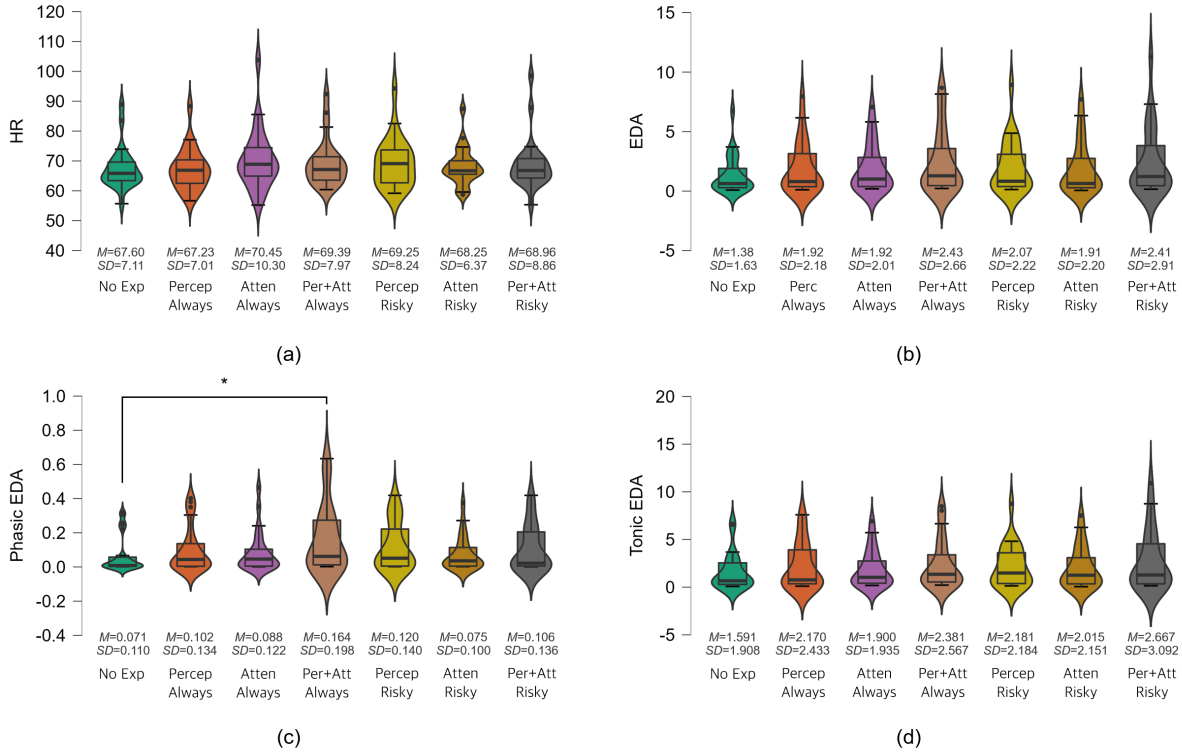


Fig. 10. Physiological responses of participants who experienced each explanation condition: (a) Heart rate, (b) EDA, (c) Phasic EDA, and (d) Tonic EDA. $^*p < .05$.

4.4.8 Lessons From Study Participants. During interviews, participants suggested ways to enhance in-vehicle explanation visualization. Participant P4 recommended sharpening the contour of the segmentation map and removing color of the map to avoid interfering with visibility, while P21 advocated for user-customizable explanations with diverse visualization options. Also, P30 suggested on-demand explanations, P5 advised continuous display of the segmentation map coupled with attention map presentation only under hazardous conditions, and P29 proposed employing a distinct color code on the segmentation map to denote hazardous situations. These suggestions provide intriguing prospects for designing in-vehicle explanations that are less visually and cognitively demanding.

We also observed that people are insensitive to minute alignment errors of their viewpoints in physical/virtual vehicles. Although our system was designed to properly track the position and orientation of an HMD in a moving vehicle, the drift of the IMU sensor and intrinsic inaccuracies in the HMD system caused an angular deviation from the participant's standard viewpoint. Most participants instinctively corrected these by adjusting their head orientation. However, they remained oblivious to their physical head orientation until these were rectified using a VR base station. This observation is consistent with findings from VR-redirected walking experiments, which manipulated undetectable gains [67].

5 DISCUSSION

5.1 What Type of Explanation? Sharing the Perception State of Automated Vehicles Was Favored Over Attention Information in Most Passenger Experience Measures (RQ1).

The provision of perception information through WSD was deemed to have the highest usability, trust, and perceived safety among the explanation conditions tested. It fostered greater situational awareness without increasing cognitive burden. This result is consistent with prior research [11, 12] that segmentation visualization promotes passenger trust and situational awareness while reducing cognitive load. On the other hand, despite being the most widely employed among AI engineers and dataset experts, the saliency-based attention map (Grad-CAM) was less effective in promoting end-user passenger experience than the perception state itself, or in some measures, the condition without explanations. The most frequently mentioned problem regarding the attention heatmap was its indirectness. One must interpret why the vehicle is paying attention to a given object in terms of the object's behavior and the potential consequences of the situation. Since the driving scene changes rapidly, individuals may be unable to accept and analyze implicit information quickly. Hence, explanations should be sufficiently clear, either by providing direct and obvious information or by applying additional algorithms to translate indirect explanations into a more human-centered format.

5.2 When to Explain? Traffic Risk-adaptive Explanations Can Be Effective When the Amount of Information Is Overwhelming (RQ1).

Explanation timing had no main effect, but it had an interaction effect with the explanation type on trust and perceived safety. Specifically, risk-adaptive explanations improved the levels of trust and perceived safety when combined with the perception + attention explanation type, whereas it had adverse effects for the perception-only or attention-only explanation conditions. Such enhanced passenger experiences also lead to greater user preferences. Participant responses indicate that providing explanations based on predicted traffic risk can prevent information overload, particularly when excessive amounts of visual information are presented. The lower arousal of participants measured based on phasic EDA in condition 7 (perception + attention, if risky) than in condition 4 (perception + attention, always) supports the idea that risk-adaptive explanations can be an effective strategy for reducing passenger burden in automated vehicles. Moreover, risk-adaptive explanations support higher situational awareness for all types of explanations, despite the reduced amount of information delivered.

5.3 How Do Explanations Help Acceptance? Explanations Foster Acceptance by Promoting Understandability, Perceived Safety, and Trust (RQ2).

While numerous studies have reported the role of explanations in fostering trust and acceptance for automated vehicles, the specific aspects of passenger experience affected by explanations and how they translate to acceptance have not been actively modeled. In our research, we focused on the provision of perception and attention state information, as well as the timing of these provisions, and how they can lead to automated vehicle acceptance, mediated by passenger experience and other UX-related measures. Referring to Körber [47] and Choi and Ji [9], we established a latent growth model to understand how provisioning explanations affected passenger experience and perceived capabilities of the vehicles (Figure 11). The model fits well with the comparative fit index (CFI) at .961 (>.9), the Tucker–Lewis index (TLI) at .957 (>.9), and Bollen's relative fit index (RFI) at .916 (>.9). The provision of perception and attention information positively affects the perceived capabilities of vehicles and passenger experience with perception information having greater impact than the attention information. In addition, the model describes that situation awareness and familiarity with automated vehicles are the most important factors in determining the perceived capabilities of automated vehicles and passenger experience.

We connected the latent growth model to a structural equation model designed to represent the relationship between perceived capabilities, passenger experience, and acceptance. Referring to Hewitt et al. [33], the model views acceptance in three ways: willingness (behavioral intention to use the vehicle), self-efficacy, and attitudes toward technology. The model fits well with the CFI at .953 (>.9), TLI at .920 (>.9), and Bollen's RFI at .953 (>.9). The structural equation model reveals that under automated vehicle explanation scenarios, perceived safety, trust, the user's propensity to trust, and understandability are the most crucial factors in facilitating user acceptance of automated vehicles. Since propensity to trust is an individual factor, explanations should be designed to enhance automated vehicle acceptance by promoting perceived safety and trust among passenger experience factors and understandability among user-perceived capability factors.

Nonetheless, it's important to note that the model, though fitted with 210 data samples, reflects the experiences of just 30 participants. While this model aptly represents the study participants' experiences, it may not universally represent all passengers and the results should be interpreted in the context of these limitations.

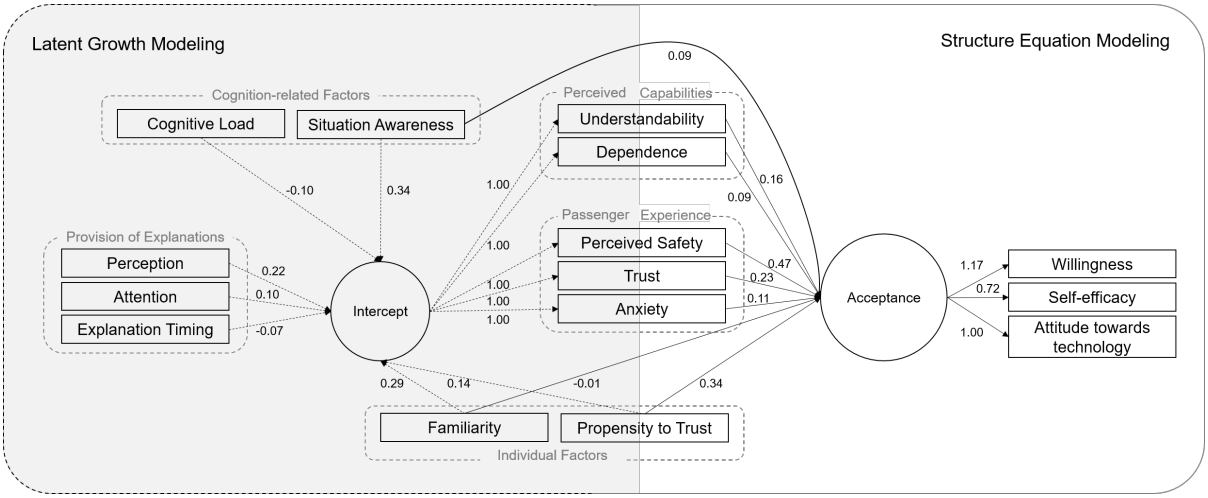


Fig. 11. Modeling acceptance of automated vehicles mediated by passenger experience with explanation provisioning.

5.4 Understanding the Scope of Explanations and Managing Error Types are Crucial for Building Trust in Automated Vehicles

Ensuring that passengers understand the scope of explanations and managing different types of errors are essential for building trust in automated vehicles. Some participants pointed out that the car did not look at the left and right sides of the vehicles, which was outside the scope of our WSD explanation. Participants' reactions to different types of errors in explanations were also noteworthy. When the vehicle did not appear to perceive a particular object, participants questioned its capabilities. However, they were not as concerned when the vehicle's segmentation was erroneously superimposed on roads, trees, or traffic lights. Participants either did not detect the error, believed the vehicle segmented the image due to exceptional circumstances, or did not care about the error, interpreting it as a possible precautionary behavior. Thus, they were more accepting of false-positive errors (perceiving vehicles or pedestrians when none were present) than false-negative errors (failing to perceive vehicles or pedestrians that were present), regarding the "what to explain".

Regarding "when to explain," participants preferred false negatives (not explaining when a situation was dangerous). Those who did not prefer risk-adaptive explanations questioned the vehicle's criteria for judging a situation as risky. Their complaints primarily focused on false positives, where the vehicle provided an explanation in situations they did not perceive as risky. Conversely, they were not as concerned about false negatives, where the vehicle did not provide an explanation despite a perceived danger. They believed the vehicle coped with the situation safely and did not perceive it as a hazard. While explanations should be designed to minimize errors, a rigorous investigation into how individuals perceive different types of errors can be leveraged to enhance the passenger experience, which may vary depending on how individuals are engaged in monitoring tasks.

6 LIMITATIONS & FUTURE WORK

6.1 On-road Simulators Can Be Complemented by Indoor Experiments and Actual Implementation

Our system, while enhancing ecological validity by allowing experiments in actual vehicles on roads, still exhibits limitations compared to genuine self-driving cars. For instance, it constricts the participant's view with an HMD that suffers from latency, reduced field-of-view, and a lower framerate compared to human vision. Moreover, our system cannot distinguish between the translational movement of the HMD and the vehicle, thus restricting the HMD's movement. Prolonged use also leads to rotational disorientation of the HMD due to IMU drift, as discussed in section 4.4.8. To create on-road simulators with higher fidelity, additional sensors such as OBD2, GPS, and GNSS could be used to track the motion of both the HMD and the vehicle [63].

Additionally, our method may be less suitable for automated vehicles of SAE Level 3 or lower since the car is controlled by a wizard driver and the passenger passively experiences the drive without any possibility of intervention. As these vehicles allow for driver take-over, safety measures related to driving, such as performance during and after take-over [8, 45], or critical scenarios that can lead to potential crashes, should be tested using indoor simulators. Thus, the choice of a simulator platform should depend on the experiment type, considering the trade-off between naturalistic scenarios and fully controllable settings. Indoor simulators and on-road testing methods can complement each other in designing explanations for automated vehicles.

Also, the limitations of wizard-of-oz automated driving simulations should be noted. Detjen et al. [17] expressed concerns that the discovery of the wizard could influence passengers' perceptions and behaviors. In contrast, Schneider et al. [77] found in their WoZ study, conducted on actual roads, that informing passengers about the wizard driver did not affect their experience of the automated vehicle simulation. In our study, we did not ask whether each participant had noticed the non-automated nature of the experiment. Consequently, there is a possibility that the passenger behavior observed may not precisely reflect behavior in real automated vehicles. Furthermore, the use of a Head-Mounted Display (HMD) in our mixed-reality-based wizard-of-oz automated vehicle study may have prevented participants from actively engaging in certain Non-Driving Related Tasks

(NDRTs), such as eating or drinking. Therefore, implementing visual explanations using an actual Windshield Display (WSD) should be considered to further validate results from wizard-of-oz studies.

6.2 Effect of NDRTs, Information Quantity, and Explanation Modality Should Be Further Explored.

Since working memory is a finite resource, the task of understanding in-vehicle explanations can compete with ongoing NDRTs. Our study has a limitation in not incorporating complex NDRTs like eating, drinking, reading, or interacting with multimedia due to experimental constraints such as HMD. However, the quantity of information and the modality of explanations should be carefully designed, depending on the NDRT types, to effectively deliver the explanation to the passenger without imposing additional cognitive load.

Passengers in automated vehicles engage in NDRTs that require visual, auditory, and motor skills [82]. While motor tasks, such as those involving handheld devices, were key considerations for designing take-over requests [87], they become less significant in highly automated vehicles that don't require human intervention. Instead, the impact of NDRTs on human memory resources and the quantity of information conveyed through different channels become crucial. For instance, auditory channels can effectively relay information in vehicles when passengers are engaged in NDRTs [4] as they don't distract from the visual attention necessary for tasks like reading and watching [32]. Therefore, further exploration into the use of textual [43], auditory [19], or sonic explanations [23] could improve the passenger experience in automated vehicles. Given that our platform is based on the Unity 3D engine, incorporating audio-based applications such as speech recognition [49], text-to-speech generation [4], and natural language understanding [38, 48] would establish an environment suitable for testing verbal explanations or natural interactions [1] with automated vehicles.

6.3 Passenger Perceived Risk and Demand for Explanation is More than Binary.

We evaluated the explanation condition using a risk-prediction model that was trained on the Car Crash Dataset [3] to classify driving scenarios into two categories based on the probability of a risk. However, driving scenarios and passenger responses are often more complex and nuanced than such a binary framework can capture. For instance, Li et al. [55] derive situational risk from three scenarios: speed, traffic, and abnormal behaviors, while Wiegand et al. [89] identify six categories to describe unexpected driving behaviors. Drawing from these systematic approaches to situational classification and established research on driving situation analysis [71, 91], self and external interruptions [24, 36, 81], and driver interruptibility [39, 40, 44], more accurately moments when passengers require explanations can be detected.

6.4 Passenger Experience with Vehicle Attention can be Enhanced by More Accurate Saliency Map.

Despite its popularity in describing the behavior of deep-learning models, displaying an attention map to the passengers did not enhance the passenger experience and was not preferred over displaying the perception information of the vehicle. However, we also stress that the outcome should be taken in light of our particular saliency map configuration, which can be enhanced by employing alternative algorithms. Whereas the saliency map applied in the current study is generated using a CNN model that predicts the risk probability, driving a car requires a more comprehensive visual analysis than predicting the likelihood of a traffic accident. Consequently, the saliency maps generated by algorithms that cover more complex activities of automated vehicles, such as end-to-end driving [41], may improve the passenger experience. In addition, CNN-based solutions may demonstrate a poor saliency map for the driving decisions, i.e., they highlight irrelevant regions such as bushes over the road horizon and along the edges of the road [52]. Incorporating more accurate saliency maps can enhance the passenger experience by showing that the vehicle is focused on the regions crucial for driving decisions.

The current study tested the WSD-based explanations, and the potential of additional visualization approaches should be further investigated. For example, the Tesla ADAS features visualization of the surrounding objects, and

its effectiveness should be further investigated. Colley et al. [11] have discovered that although AR visualization is preferable to a tablet placed on the fascial center, a tablet-based system similar to that developed by Tesla can potentially enhance the passenger experience [10]. An on-road comparison of visualization methods, such as the ADAS features of the Tesla Autopilot and the visualization methods suggested by the study participants, can inspire the development of explainable algorithms that can generate effective explanations.

6.5 The Platform can be Extended to Test the Interaction Between Passengers and Automated Vehicles with Parallel Autonomy.

The current study investigated the passenger experience with explanations in a highly automated car with a wizard driver. In our experiment, passengers only viewed the Wizard-of-Oz-based automated ride of the vehicle. However, safety and regulatory considerations make shared automated vehicles a more likely and immediate future, as they do not totally eliminate the role of drivers in cars. Collaborative autonomy provides safer driving because humans and artificial intelligence can cross-check or assist one another [59]. Among the different varieties of shared autonomy, cars with parallel autonomy operate as "guardian angels" that avoid potential accidents by adjusting human driving [79]. The platform used in this study can be expanded to test vehicles with parallel autonomy using natural language services or displays that are less distracting, such as HUDs, center fascias, or optical see-through MR applications.

Future studies may include advancing the platform to test the "guardian angel" feature for automated vehicles with parallel autonomy. By integrating a drive-by-wire system, sensors, and algorithms for self-driving, the platform can test various types of feedback and explanation methods for parallel automated vehicles (e.g., a parallel autonomy research platform [66]). More detailed descriptions of the vehicle state and decisions with expanded modalities, such as verbal or textual explanations during or after a driving adjustment, can be tested to enhance the passenger experience for automated vehicles with parallel autonomy. Such a guardian system can also be applied with implicit interactions [84] to promote safe control of the driver and minimize the discrepancies between the self-driving algorithm and human driver in an unobtrusive manner.

7 CONCLUSION

In this study, we examined the impact of explanation type and timing mechanisms provided in automated vehicles on passenger experience using a mixed-reality Wizard-of-Oz self-driving simulator. We compared three types of windshield displays for explanations: perception, attention, and a combination of both perception and attention. Through a human-subject experiment conducted on actual roads, we validated previous indoor study results, confirming that sharing perception state itself enhanced perceived usability, trust, safety, and situation awareness. In addition, we leveraged the benefits of outdoor experiments, which can provide a more realistic sense of risk when testing explanations. Specifically, we utilized Grad-CAM attention to highlight risky regions under naturalistic driving scenarios and provide explanations selectively depending on traffic risk levels. Although attention information alone was not highly favored, the risk-adaptive strategy for explanation delivery was effective in the perception + attention condition, where passengers were provided with extensive information. In our study, we emphasized the importance of suitable explanations for fostering understandability, safety, and trust, consequently promoting the acceptance of automated vehicles. However, our findings also suggest a nuanced perception among participants regarding the "what" and "when" aspects of explanations, which can be leveraged in tailoring in-vehicle explanations.

ACKNOWLEDGMENTS

This work was supported by the GIST-MIT Research Collaboration grant funded by the GIST in 2023.

REFERENCES

[1] 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

[2] 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

[3] Wentao Bao, Qi Yu, and Yu Kong. 2020. Uncertainty-Based Traffic Accident Anticipation with Spatio-Temporal Relational Learning. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 2682–2690. https://doi.org/10.1145/3394171.3413827

[4] Pavlo Bazilinskyy and Joost de Winter. 2015. Auditory interfaces in automated driving: An international survey. *PeerJ Computer Science* 1 (Aug. 2015), e13. https://doi.org/10.7717/peerj-cs.13

[5] 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

[6] 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

[7] John Brooke. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[8] Damee Choi, Toshihisa Sato, Takafumi Ando, Takashi Abe, Motoyuki Akamatsu, and Satoshi Kitazaki. 2020. Effects of cognitive and visual loads on driving performance after take-over request (TOR) in automated driving. *Applied Ergonomics* 85 (2020), 103074. https://doi.org/10.1016/j.apergo.2020.103074

[9] 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

[10] Mark Colley, Christian Bräuner, Mirjam Lanzer, Marcel Walch, Martin Baumann, and Enrico Rukzio. 2020. Effect of Visualization of Pedestrian Intention Recognition on Trust and Cognitive Load. 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, 181–191. https://doi.org/10.1145/3409120.3410648

[11] 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

[12] 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

[13] 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

[14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3213–3223. https://doi.org/10.1109/CVPR.2016.350

[15] 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

[16] 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.

[17] 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

[18] Henrik Detjen, Maurizio Salini, Jan Kronenberger, Stefan Geisler, and Stefan Schneegass. 2021. Towards Transparent Behavior of Automated Vehicles: Design and Evaluation of HUD Concepts to Support System Predictability Through Motion Intent Communication. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction* (Toulouse Virtual, France) *(MobileHCI '21)*. Association for Computing Machinery, New York, NY, USA, Article 19, 12 pages. https://doi.org/10.1145/3447526.3472041

[19] 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

[20] Andrew T. Duchowski, Krzysztof Krejtz, Izabela Krejtz, Cezary Biele, Anna Niedzielska, Peter Kiefer, Martin Raubal, and Ioannis Giannopoulos. 2018. The Index of Pupillary Activity: Measuring Cognitive Load Vis-à-Vis Task Difficulty with Pupil Oscillation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173856

[21] Donald L Fisher, Matthew Rizzo, Jeffrey Caird, and John D Lee. 2011. *Handbook of Driving Simulation for Engineering, Medicine, and Psychology*. CRC Press.

[22] Anna-Katharina Frison, Philipp Wintersberger, Andreas Riener, Clemens Schartmüller, Linda Ng Boyle, Erika Miller, and Klemens Weigl. 2019. In UX We Trust: Investigation of Aesthetics and Usability of Driver-Vehicle Interfaces and Their Impact on the Perception of Automated Driving. 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, 1–13. https://doi.org/10.1145/3290605.3300374

[23] Nick Gang, Srinath Sibi, Romain Michon, Brian Mok, Chris Chafe, and Wendy Ju. 2018. Don't Be Alarmed: Sonifying Autonomous Vehicle Perception to Increase Situation Awareness. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Toronto, ON, Canada) *(AutomotiveUI '18)*. Association for Computing Machinery, New York, NY, USA, 237–246. https://doi.org/10.1145/3239060.3265636

[24] 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

[25] Samir Ghouali, Yassine Zakarya Ghouali, and Mohammed Feham. 2017. An investigation of analytic decision during driving test. *International journal of advanced computer science and applications (IJACSA)* 8 (2017).

[26] 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

[27] David Goedicke, Jamy Li, Vanessa Evers, and Wendy Ju. 2018. VR-OOM: Virtual Reality On-ROad Driving SiMulation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3173574.3173739

[28] Claudia V. Goldman and Ronit Bustin. 2022. Trusting Explainable Autonomous Driving: Simulated Studies. In *2022 IEEE Intelligent Vehicles Symposium (IV)*. 1255–1260. https://doi.org/10.1109/IV51971.2022.9827312

[29] 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

[30] 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

[31] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[32] Ellen Hass and Judy Edworthy. 2006. *Handbook of warnings* (1st. ed.). CRC Press. 189–198 pages.

[33] 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

[34] Philipp Hock, Sebastian Benedikter, Jan Gugenheimer, and Enrico Rukzio. 2017. CarVR: Enabling In-Car Virtual Reality Entertainment. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 4034–4044. https://doi.org/10.1145/3025453.3025665

[35] Philipp Hock, Mark Colley, Ali Askari, Tobias Wagner, Martin Baumann, and Enrico Rukzio. 2022. Introducing VAMPIRE – Using Kinaesthetic Feedback in Virtual Reality for Automated Driving Experiments. 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, 204–214. https://doi.org/10.1145/3543174.3545252

[36] Christian P. Janssen, Shamsi T. Iqbal, Andrew L. Kun, and Stella F. Donker. 2019. Interrupted by my car? Implications of interruption and interleaving research for automated vehicles. *International Journal of Human-Computer Studies* 130 (2019), 221–233. https://doi.org/10.1016/j.ijhcs.2019.07.004

[37] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04 arXiv:https://doi.org/10.1207/S15327566IJCE0401$_0$4

[38] 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. *arXiv preprint arXiv:2302.00673* (2023).

[39] 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

[40] 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

[41] 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

[42] Jinkyu Kim, Suhong Moon, Anna Rohrbach, Trevor Darrell, and John Canny. 2020. Advisable Learning for Self-Driving Vehicles by Internalizing Observation-to-Action Rules. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9658–9667. https://doi.org/10.1109/CVPR42600.2020.00968

[43] 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

[44] 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

[45] 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

[46] 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

[47] Moritz Körber. 2019. Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. 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, 13–30.

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

[49] Tuomo Kujala and Hilkka Grahn. 2017. Visual Distraction Effects of In-Car Text Entry Methods: Comparing Keyboard, Handwriting and Voice Recognition. 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, 1–10. https://doi.org/10.1145/3122986.3122987

[50] 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

[51] Pravallika Lavanuru, Sawon Pratiher, Karuna P Sahoo, Mrinal Acharya, Nirmalya Ghosh, and Amit Patra. 2023. Parasympathetic-Sympathetic Causal Interactions and Perceived Workload for Varying Difficulty Affective Computing Tasks. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[52] Mathias Lechner, Ramin Hasani, Alexander Amini, Thomas A. Henzinger, Daniela Rus, and Radu Grosu. 2020. Neural circuit policies enabling auditable autonomy. *Nature Machine Intelligence* 2, 10 (2020), 6420652. https://doi.org/10.1038/s42256-020-00237-3

[53] 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$_3$0392 PMID: 15151155.

[54] James R. Lewis and Jeff Sauro. 2018. Item Benchmarks for the System Usability Scale. *J. Usability Studies* 13, 3 (may 2018), 158–167.

[55] 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

[56] 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

[57] 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

[58] Harsh Mankodiya, Dhairya Jadav, Rajesh Gupta, Sudeep Tanwar, Wei-Chiang Hong, and Ravi Sharma. 2022. OD-XAI: Explainable AI-Based Semantic Object Detection for Autonomous Vehicles. *Applied Sciences* 12, 11 (2022). https://doi.org/10.3390/app12115310

[59] Steffen Maurer, Rainer Erbach, Issam Kraiem, Susanne Kuhnert, Petra Grimm, and Enrico Rukzio. 2018. Designing a Guardian Angel: Giving an Automated Vehicle the Possibility to Override Its Driver. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Toronto, ON, Canada) *(AutomotiveUI '18)*. Association for Computing Machinery, New York, NY, USA, 341–350. https://doi.org/10.1145/3239060.3239078

[60] Mark McGill and Stephen Brewster. 2019. Virtual Reality Passenger Experiences. 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, 434–441. https://doi.org/10.1145/3349263.3351330

[61] Mark McGill, Alexander Ng, and Stephen Brewster. 2017. I Am The Passenger: How Visual Motion Cues Can Influence Sickness For In-Car VR. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5655–5668. https://doi.org/10.1145/3025453.3026046

[62] Mark McGill, Julie Williamson, Alexander Ng, Frank Pollick, and Stephen Brewster. 2019. Challenges in passenger use of mixed reality headsets in cars and other transportation. *Virtual Reality* 24, 4 (2019), 583–603. https://doi.org/10.1007/s10055-019-00420-x

[63] Mark McGill, Graham Wilson, Daniel Medeiros, and Stephen Anthony Brewster. 2022. PassengXR: A Low Cost Platform for Any-Car, Multi-User, Motion-Based Passenger XR Experiences. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 2, 15 pages. https://doi.org/10.1145/3526113.3545657

[64] Tobias Müller, Mark Colley, Gülsemin Dogru, and Enrico Rukzio. 2022. AR4CAD: Creation and Exploration of a Taxonomy of Augmented Reality Visualization for Connected Automated Driving. *Proc. ACM Hum.-Comput. Interact.* 6, MHCI, Article 177 (sep 2022), 27 pages. https://doi.org/10.1145/3546712

[65] 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

[66] Felix Naser, David Dorhout, Stephen Proulx, Scott Drew Pendleton, Hans Andersen, Wilko Schwarting, Liam Paull, Javier Alonso-Mora, Marcelo H. Ang, Sertac Karaman, Russ Tedrake, John Leonard, and Daniela Rus. 2017. A parallel autonomy research platform. In *2017 IEEE Intelligent Vehicles Symposium (IV)*. 933–940. https://doi.org/10.1109/IVS.2017.7995835

[67] Niels Christian Nilsson, Tabitha Peck, Gerd Bruder, Eri Hodgson, Stefania Serafin, Mary Whitton, Frank Steinicke, and Evan Suma Rosenberg. 2018. 15 Years of Research on Redirected Walking in Immersive Virtual Environments. *IEEE Computer Graphics and Applications* 38, 2 (2018), 44–56. https://doi.org/10.1109/MCG.2018.111125628

[68] 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

[69] 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

[70] 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

[71] Erfan Pakdamanian, Shili Sheng, Sonia Baee, Seongkook Heo, Sarit Kraus, and Lu Feng. 2021. DeepTake: Prediction of Driver Takeover Behavior Using Multimodal Data. 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 103, 14 pages. https://doi.org/10.1145/3411764.3445563

[72] Florence Rosey, Jean-Michel Auberlet, Olivier Moisan, and Guy Dupré. 2009. Impact of Narrower Lane Width: Comparison Between Fixed-Base Simulator and Real Data. *Transportation Research Record* 2138, 1 (2009), 112–119. https://doi.org/10.3141/2138-15 arXiv:https://doi.org/10.3141/2138-15

[73] 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

[74] Davide Salanitri, Glyn Lawson, and Brian Waterfield. 2016. The Relationship Between Presence and Trust in Virtual Reality. In *Proceedings of the European Conference on Cognitive Ergonomics* (Nottingham, United Kingdom) *(ECCE '16)*. Association for Computing Machinery, New York, NY, USA, Article 16, 4 pages. https://doi.org/10.1145/2970930.2970947

[75] 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

[76] 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

[77] 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

[78] Martin Schrepp. 2020. On the Usage of Cronbach's Alpha to Measure Reliability of UX Scales. *Journal of Usability Studies* 15, 4 (2020).

[79] Wilko Schwarting, Javier Alonso-Mora, and Daniela Rus. 2018. Planning and Decision-Making for Autonomous Vehicles. *Annual Review of Control, Robotics, and Autonomous Systems* 1, 1 (2018), 187–210. https://doi.org/10.1146/annurev-control-060117-105157 arXiv:https://doi.org/10.1146/annurev-control-060117-105157

[80] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 618–626. https://doi.org/10.1109/ICCV.2017.74

[81] 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

[82] Farzaneh Shahini and Maryam Zahabi. 2022. Effects of levels of automation and non-driving related tasks on driver performance and workload: A review of literature and meta-analysis. *Applied Ergonomics* 104 (2022), 103824. https://doi.org/10.1016/j.apergo.2022.103824

[83] 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.

[84] 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

[85] 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

[86] Richard M. Taylor. 2011. *Situational Awareness* (1st. ed.). CRC Press. 111–128 pages.

[87] Bernhard Wandtner, Nadja Schömig, and Gerald Schmidt. 2018. Effects of Non-Driving Related Task Modalities on Takeover Performance in Highly Automated Driving. *Human Factors* 60, 6 (2018), 870–881. https://doi.org/10.1177/0018720818768199 arXiv:https://doi.org/10.1177/0018720818768199 PMID: 29617161.

[88] 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

[89] 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

[90] 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

[91] Frederik Wiehr, Baris Cakar, Florian Daiber, and Antonio Krüger. 2021. The Effect of Surrounding Scenery Complexity on the Transfer of Control Time in Highly Automated Driving. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 92–97. https://doi.org/10.1145/3397481.3450677

[92] Bob Williams. 2021. *Automated Driving Levels*. 19–41. https://doi.org/10.1002/9781119765394.ch2

[93] JJiacong Xu, Zixiang Xiong, and Shankar P. Bhattacharyya. 2022. PIDNet: A Real-time Semantic Segmentation Network Inspired from PID Controller. arXiv:2206.02066 [cs.DL]

[94] 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

[95] Xuedong Yan, Mohamed Abdel-Aty, Essam Radwan, Xuesong Wang, and Praveen Chilakapati. 2008. Validating a driving simulator using surrogate safety measures. *Accident Analysis Prevention* 40, 1 (2008), 274–288. https://doi.org/10.1016/j.aap.2007.06.007

[96] 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

[97] 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