MUST: Smartwatch-based Multimodal Framework for Predicting Driver State and Takeover Performance

Seokyong Sheem
Korea University
Seoul, South Korea
sheemsy@korea.ac.kr

Yujin Cho
Korea University
Seoul, South Korea
jinyu317@korea.ac.kr

In Kyung Lee
Inha University
Incheon, South Korea
9ruddls3@inha.edu

Hanjun Cho
Hyundai Motor Company
Seoul, South Korea
successjo@hyundai.com

Taegeun Kim
Hyundai Motor Company
Seoul, South Korea
taegeun.kim@hyundai.com

Byung Hyung Kim*
Inha University
Incheon, South Korea
bhyung@inha.ac.kr

Daekyum Kim*
Korea University
Seoul, South Korea
daekyum@korea.ac.kr

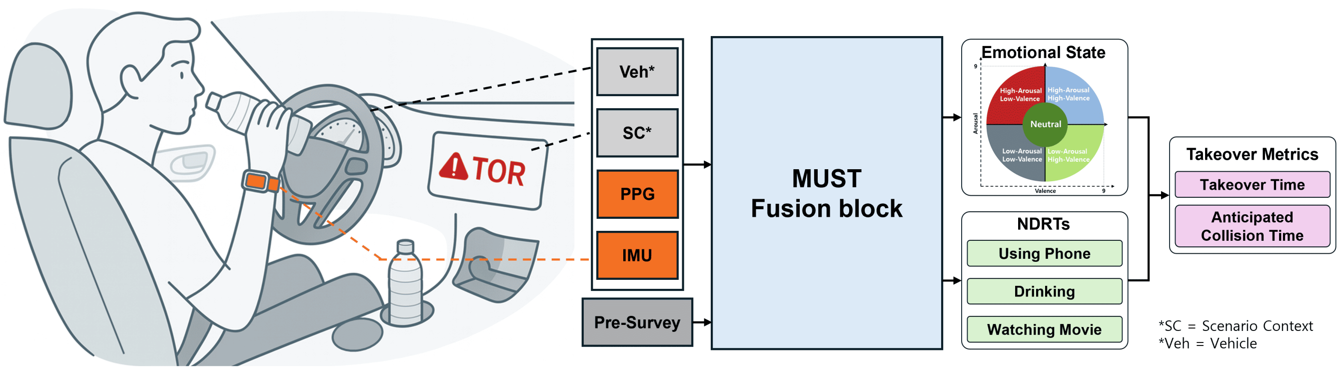

Figure 1: MUST framework integrates multimodal signals and processes them through a fusion block. Extracted features are then analyzed with deep learning models to estimate driver states, non-driving related tasks (NDRTs), and arousal–valence, enabling accurate prediction of takeover-related metrics.

Abstract

Ensuring timely takeover in conditionally autonomous vehicles presents a significant challenge, especially when drivers are distracted by non-driving-related tasks or are in suboptimal emotional states. Existing driver monitoring systems struggle with a trade-off between practicality and reliability. Physiological sensors are intrusive, vision-based methods are sensitive to occlusions and variable lighting, and current multimodal learning approaches often rely on simple fusion strategies that fail to reconcile heterogeneous data. We introduce MUST (Multimodal Unified Smartwatch-based Takeover), a framework that predicts driver state and takeover performance using unobtrusive smartwatch signals. MUST employs an asymmetric causal fusion mechanism to model the interplay between driver behavior and emotion. The performance of the architecture was validated in diverse simulator environments reflecting real-world driving conditions, demonstrating robust driver state estimation and takeover prediction. This work establishes the smartwatch as a practical tool for adaptive takeover support, enabling reliable readiness assessment without intrusive hardware or fragile vision systems.

*Corresponding author.

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3791703

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**.

Keywords

Automated Driving, Takeover behavior, Smartwatch, Multimodal Fusion, Multi-task learning

ACM Reference Format:
Seokyong Sheem, Yujin Cho, In Kyung Lee, Hanjun Cho, Taegeun Kim, Byung Hyung Kim, and Daekyum Kim. 2026. MUST: Smartwatch-based Multimodal Framework for Predicting Driver State and Takeover Performance. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3772318.3791703

1 Introduction

The transition to SAE Level 3 automation fundamentally shifts the driver's role from an active controller to a supervisory monitor [75]. While providing convenience, it also introduces a serious safety challenge. Drivers are expected to safely resume control when the system triggers a Takeover Request (TOR) due to legal or technical limitations.

Even with strong Automated Driving System (ADS), TOR handling remains a necessary fallback due to operational design domain (ODD) limits and edge cases. Experimental evidence demonstrates that they often struggle with this transition, exhibiting poor situational awareness and slow reaction time to stabilize control [9, 35, 85]. Real-world safety reports support these findings, highlighting human supervisory failures as a recurring factor in incidents. Notably, several ADS crash analyses attribute accidents to the driver's inability to re-engage effectively [58].

Such re-engagement failures often occur as timely responses to TORs are disrupted by drivers' engagement in non-driving-related tasks (NDRTs) or by strong emotional states [25, 38, 45]. Specifically, when emotional arousal levels fluctuate, hazard perception and reaction times are compromised, directly impairing takeover readiness [46, 79]. To capture and evaluate these risks in a structured manner, researchers have examined takeover performance, primarily through two key constructs: takeover time (TOT) and takeover quality (TOQ) [4, 15, 25].

While prior research has extensively investigated factors such as NDRTs and road conditions influencing takeover performance, these studies are primarily descriptive, lacking real-time predictive capabilities [45, 82, 84]. Recent work has applied machine learning to predict takeover outcomes, showing the potential to improve predictive accuracy [25]. However, major challenges still persist: practical limitations of sensing modalities, methodological difficulties in modeling multimodal data, and insufficient validation under diverse scenarios that realistically mirror real-world traffic.

Regarding sensing modalities, physiological signals such as electroencephalography (EEG) and galvanic skin response (GSR) provide high-fidelity measurements of driver states [6, 20, 83]. However, their traditional implementations typically require cumbersome setups, including EEG caps or skin electrode patches, making them impractical for everyday use. Vision-based methods often suffer from occlusion, poor illumination, and sensitivity to camera placement, limiting their reliability in dynamic driving conditions [2, 34, 36, 54].

Methodologically, many multimodal deep learning approaches still rely on relatively simple fusion strategies, limiting the effective synthesis of heterogeneous data [21, 26, 32, 60]. Finally, most studies have validated their methods under controlled simulation environments, questioning their generalizability to real-world driving scenarios [28, 76].

To overcome these challenges, we introduce the **Multimodal Unified Smartwatch-based Takeover (MUST)** framework, an integrated architecture inferring both driver states and takeover-related metrics in real time using a wrist-worn sensor. MUST leverages Photoplethysmography (PPG) and inertial measurement unit (IMU) signals from the smartwatch, alongside vehicle telemetry and pre-survey data, suggesting potential feasibility for scalable deployment on commercially available devices [24]. This sensing strategy preserves usability by remaining unobtrusive and integrating seamlessly into drivers' daily routines.

The core of the framework is a hierarchical modeling approach tailored to capture the multifaceted nature of driver state. Stage 1 trains modality-specific expert models that separately encode broader psycho-physiological context alongside immediate behavioral determinants. Stage 2 then integrates these expert representations through an asymmetric causal fusion mechanism, allowing complementary knowledge to be shared while respecting temporal and semantic differences. This hierarchical principle of *specialization followed by integration* is well established in ML [53, 70], and here it is adapted to an HCI challenge: linking heterogeneous driver states into a coherent predictive model that can inform adaptive interface design.

To evaluate the framework under conditions that elicit diverse affective and contextual dynamics, we implemented a TOR interface in the CARLA simulator. We followed the protocols in existing studies by combining visual and auditory cues [60], enabling us to systematically induce emotional variability while ensuring experimental control.

We validated MUST in 13 hazardous driving scenarios in CARLA, including roadwork zone, sudden decelerations, and pedestrian incursions. The results show that the framework can accurately infer driver states and takeover metrics under diverse conditions while remaining practical for real-world deployment [72].

The main contributions of this paper are as follows:

- **Smartwatch-based driver monitoring:** An end-to-end framework that leverages unobtrusive wearable sensors with vehicle telemetry to infer driver states in real time, mitigating limitations of intrusive physiological sensing and fragile vision systems.
- **Asymmetric causal fusion:** A novel mechanism where motion cues inform emotional inference via delayed cross-attention, while emotional states provide global context for motion prediction, enabling robust multi-task learning.
- **Takeover performance prediction:** Integration of behavioral and affective state inferences to enhance predictions of TOT and ACT.
- **Validation in realistic scenarios:** Evaluation in 13 CARLA-based takeover scenarios involving 48 participants, demonstrating accurate state inference and reliable takeover prediction under dynamic hazards.

2 Related Work

Sensing Modalities. Balancing practicality and fidelity remains essential to ensure that sensing technologies are applicable in real-world driving contexts. Physiological sensors like EEG or GSR directly measure nervous system activity and are highly sensitive to subtle changes in cognitive load [6, 20, 33, 37]. However, they often require intrusive devices such as EEG caps or adhesive electrodes [73]. These devices are uncomfortable for drivers, impractical for daily use, and may alter driving behavior [31, 71]. As a contact-free alternative, vision-based systems are widely adopted to analyze features like gaze direction, head pose, and facial expressions [27, 38, 77]. While large-scale vision datasets have enabled the development of high-capacity models, their effectiveness can be compromised by the unpredictable real-world driving conditions [60]. For instance, Janveja et al. [43] noted that a common RGB camera pipeline failed to detect the driver's face in over 90% of nighttime frames. Accuracy also declines sharply with occlusions and changes in driver's head pose. Gaze estimation error can increase by over 25 percentage points when the head is rotated beyond 45 degrees to check mirrors [77], and sunglasses can cause emotion misclassification exceeding 30% [59]. These challenges illustrate the difficulty of developing sensing approaches that are both practical and robust in real-world driving. In response to these limitations, consumer-grade wearable devices have emerged as a compelling compromise for in-vehicle driver monitoring [5, 14, 57]. Compared to vision-based systems, wearable sensors are more robust to lighting conditions or occlusions, and they are significantly less intrusive than traditional physiological setups like EEG caps [80, 86]. Moreover, wrist-worn devices benefit from everyday familiarity and high user acceptance, making them particularly suitable for integration into naturalistic driving contexts [29, 65]. While susceptible to motion artifacts during intense physical activity, wrist-worn PPG has been identified as a feasible and promising modality for real-time driver monitoring [5]. Supporting this feasibility, Costantini et al. validated wrist-worn PPG for heart-rate–variability estimation in stressful driving scenarios and reported acceptable agreement with reference ECG measures [17].

Modeling and Fusion Strategies. Recent work has applied ML and DL methods to predict driver takeover performance [7, 26, 87]. Early efforts typically relied on feature-level concatenation to merge data from various sources into a single deep neural network, like Deeptake [61]. While superior to unimodal baselines, this approach fails to explicitly model interdependencies and remains sensitive to noise in a single data stream [13]. To overcome these limitations, later work has incorporated more advanced mechanisms, such as transformer-based architectures [28] and cross-modal attention [67], enabling richer integration across modalities. Building on these advances, EmoTake [36] incorporated emotional cues into takeover prediction, and its successor Multi-TBP [28] extended this line of work by framing takeover outcomes as a multi-task learning problem. These studies represent important progress, but driver states such as emotion are still treated primarily as auxiliary inputs rather than explicit predictive objectives, leaving their interdependence with behavior insufficiently explored. Empirical studies also support this interdependence: Mesken et al. [56] found that heightened anxiety is associated with increased perceived risk, while anger is linked to higher driving speeds. These findings underscores the need for predictive approaches that explicitly incorporate the interdependence between affective states and driving behavior.

Validation Practice. Takeover performance has generally been characterized through two dimensions: takeover time (TOT) and takeover quality (TOQ) [4, 15, 25]. While TOT captures the latency of driver response, TOQ is typically assessed using metrics such as lane offset and time-to-collision (TTC) [42, 61]. As an alternative to speed-based TTC, anticipated collision time (ACT) has been proposed to provide a more robust indicator of takeover safety [78]. ACT has since been adopted in several studies [50, 69], but its use has often been restricted to simplified contexts such as lane changes or intersection approaches, where only the minimum value (minACT) is considered as a scalar proxy for risk. However, Venthuruthiyil et al. [78] emphasize that in realistic and heterogeneous traffic environments, takeover stability cannot be captured by a single extreme value. Instead, the full temporal trajectory of ACT is needed to distinguish between short-lived transients and sustained periods of elevated risk. Motivated by this perspective, we adopted ACT as a continuous trajectory-level metric and validated it across thirteen diverse and hazard-rich scenarios. This design enables a more comprehensive analysis of how scenario type influences takeover stability. At the same time, for statistical analyses where trajectory-level data are less tractable, we also report minACT as a supplementary summary measure. In summary, existing work faces persistent limitations in sensing, modeling, and validation: physiological and vision-based modalities lack practicality or robustness, predictive models often miss the causal link between emotion and behavior, and evaluations remain confined to simplified scenarios. MUST addresses these gaps by combining smartwatch-driven signals with a fusion strategy that models affect–behavior interactions and by validating its effectiveness across diverse scenarios, advancing both practicality and predictive power in takeover research.

3 User Study

3.1 Participants

Figure 2: User study setup. The custom driving simulator consists of a 38-inch curved monitor, a Logitech G29 steering wheel and pedal set, and a Galaxy Tab S6 Lite tablet used to present video-based NDRTs. Participants received TORs through a Sennheiser Accentum wired headset while driving in CARLA. A custom smartwatch worn on the right wrist provided IMU and PPG signals for biometric acquisition.

Forty-eight licensed drivers were recruited across the study (19 pilot, 29 main study). The main cohort (20 male, 9 female; mean age 25.1, SD=2.82) had at least one year of driving experience (Mean-3.42, SD=2.19, normal vision and no history of simulator sickness. All provided informed consent before participation.

3.2 Simulator Environment

All sessions were conducted on a desktop-based CARLA simulator running on an NVIDIA RTX 4090 GPU, and all experiments were recorded using OBS Studio for subsequent analysis(Figure 2). Physiological and motion data were collected from a custom smartwatch worn on the right wrist, integrating an LSM9DS1 IMU and a SEN0203 PPG sensor, both sampled at 100 Hz.

To verify signal integrity under active driving, we validated the wrist-worn PPG sensor against a gold-standard ECG ($N = 15$). The device demonstrated high HRV consistency and robustness under motion artifacts, with full quantitative results reported in Appendix A. Smartwatch streams, simulation telemetry, and scenario logs

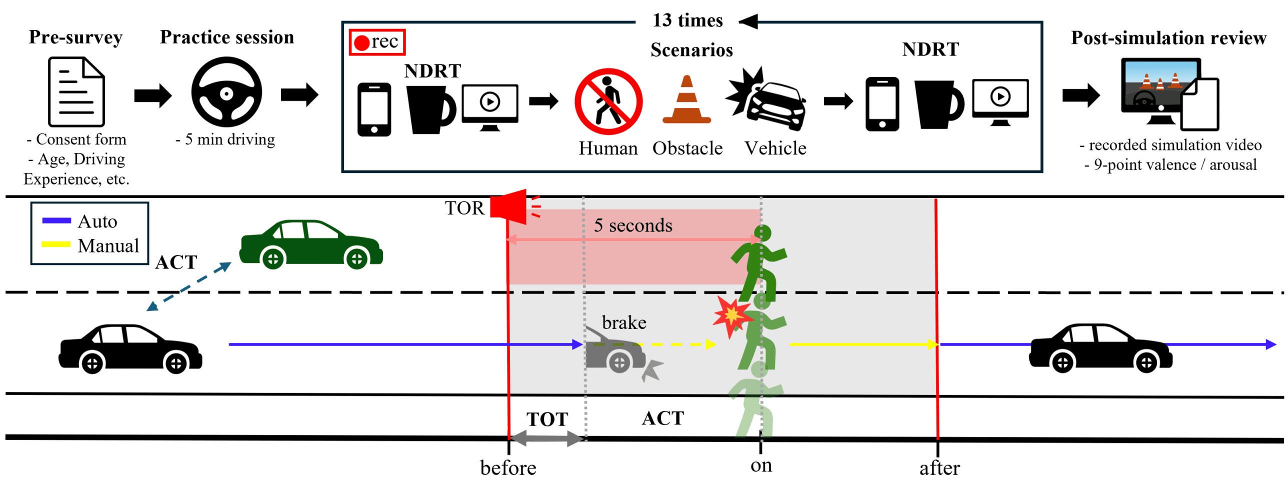

Figure 3: Experimental protocol. Participants completed a pre-survey and practice session before driving 13 takeover scenarios under three NDRTs (smartphone use, drinking, and video watching). TORs were issued in scenarios involving vehicles, pedestrians, or static obstacles, requiring drivers to assume manual control, perform evasive maneuvers, and then return to automation. All sessions were recorded, and post-surveys used the videos for emotion self-report.

were synchronized with sub-millisecond precision using the Lab Streaming Layer (LSL) over a wired Ethernet connection.

The experimental protocol comprised 13 hazard-driven takeover scenarios grouped into three categories: (1) vehicle-related conflicts, (2) unpredictable pedestrian actions, and (3) road obstacles.

3.3 Data Collection Procedure

The study comprised four stages: pre-survey, practice session, driving simulation, and post-simulation review (Figure 3).

Pre-survey. Prior to the experiment, participants completed questionnaires on demographics, driving history, prior simulator experience, and self-rated ability to handle urgent events, along with the Perceived Stress Scale-10 (PSS-10) [16]. The survey captured both cumulative background factors (e.g., confidence, experience, familiarity) and immediate condition variables (stress, fatigue, drowsiness) to support subsequent individual-difference analyses.

Practice session. Participants then completed a 5-minute familiarization drive on a simple road without hazards. This session allowed them to adapt to the simulator environment, practice manual/automatic mode switching via the steering wheel button (R2), and calibrate steering/pedal control for natural driving.

Driving simulation. The main experiment consisted of 13 fixed takeover scenarios, distributed across three NDRTs: phone use (4), drinking (4), and video watching (5). Each trial began with IMU calibration, after which participants resumed their assigned NDRT. TORs were triggered by vehicles, pedestrians, or road obstacles. Participants were required to assume manual control, perform the necessary evasive maneuver, and re-engage automation at their discretion before returning to the NDRT.

Post-simulation review. To ensure continuity, we adopted a retrospective cued-recall method. Immediately after the simulation, participants reviewed the videos to rate valence and arousal at three phases of each event: (1) before takeover, (2) during manual control, and (3) after automation resumed [8, 11].

3.4 Driving Scenarios and Takeover Events

The experimental design followed a two-phase process, beginning with a pilot study involving 19 drivers. The pilot phase tested simple obstacle avoidance scenarios on straight or curved roads and a "no alarm" baseline. Results revealed that (i) the no-alarm condition caused inevitable collisions since all participants remained fully engaged in their NDRTs, rendering TOT and ACT unusable as performance metrics, and (ii) static-obstacle-only scenarios produced limited variance in takeover-related metrics and emotional responses. These results guided the main study design, enabling the selection of takeover conditions that reflect real-world hazards and produce measurable performance differences.

Based on these insights, we developed a broader set of scenarios for the main study. Thirteen unique, high-fidelity automated driving scenarios were constructed to simulate more hazards, categorized into three types: vehicle-related conflicts, unpredictable pedestrian actions, and road obstacles. Across all scenarios, the vehicle operated in automated mode at a constant speed of 50 km/h, reflecting the typical urban limit in Korean traffic policy. Table 1 summarizes the scenarios and following NDRTs, which were balanced across phone use, drinking, and video watching.

In all scenarios, TORs were delivered through a multimodal alert combining (i) a 2s beep via headset and (ii) a textual message ("Takeover right now") displayed at the bottom of the monitor. The auditory–visual cues were issued 5 seconds before the hazard encounter, ensuring a consistent warning interval across scenarios. Takeover events were segmented into three phases: Before (the 5-second before hazard appearance), On (hazard onset), and After (re-engagement of automation).

Table 1: Takeover Scenarios under Different NDRTs

Index	Type	NDRTs	Description
1	Vehicle	Phone	Parked car door opened suddenly
2	Vehicle	Phone	Vehicle from right ignored signal
3	Human	Phone	Pedestrian jaywalking
4	Human	Phone	Pedestrian from behind bus stop
5	Obstacle	Phone	Road blocked by construction
6	Vehicle	Drink	Non-yielding vehicles at intersection
7	Vehicle	Drink	Road blocked by accident
8	Vehicle	Drink	Sudden stop of lead vehicle
9	Human	Video	Pedestrian from behind trash bin
10	Vehicle	Video	Vehicle blocking right-turn lane
11	Vehicle	Video	Vehicle from right ignored signal
12	Obstacle	Video	Obstacle in opposing lane (2-lane road)
13	Obstacle	Video	Road blocked by construction

4 Data Processing

4.1 Preprocessing

All data streams were synchronized to a 100 Hz master timeline. Vehicle telemetry and scenario context were upsampled from 30 Hz using linear interpolation for continuous variables and forward-filling for discrete states. IMU quaternions were converted to Euler angles in the ENU convention [23, 48]. PPG signals were processed with a 0.5–3 Hz band-pass filter and polynomial detrending to extract HRV metrics (RMSSD, SDNN) [22]. Finally, streams were segmented into fixed windows anchored to takeover events.

4.2 Data Labeling

Each takeover event was labeled for (i) emotion, (ii) response time, (iii) safety margin, and (iv) active NDRT.

Emotion (Valence and Arousal). Subjective emotion labels were collected using the 9-point Self-Assessment Manikin (SAM) [8]. Following a standard Cued-Recall Debriefing (CRD) protocol [11, 62], participants reviewed time-locked video clips of each event immediately post-simulation to provide retrospective ratings. For classification, these raw scores were mapped into three levels: Low (1–3), Neutral (4–6), and High (7–9) [3].

Takeover time (TOT). Time elapsed (s) from auditory TOR to manual disengagement. TOT exhibited a right-skewed distribution and was discretized into three ordinal classes (Fast ≤0.93 s, Normal 0.93–1.93 s, Slow >1.93 s), based on global quartiles of the observed data [52, 61]. Sensitivity analysis confirmed that this quartile-based strategy yields predictive performance consistent with alternative thresholds, ensuring the robustness of our results (Appendix B).

Anticipated Collision Time (ACT). For each frame, ACT was computed relative to (a) the primary hazard in the scenario and (b) the two closest surrounding vehicles. Following Venthuruthiyil et al. [78], ACT is defined as

$$\mathrm{ACT} = \frac{\delta}{\mathrm{Rel}(v_{12}, v_{21}) + \mathrm{Rel}(a_{12}, a_{21})\, t + \mathrm{Rel}(\dot{\theta}_1, \dot{\theta}_2)\, \delta}, \quad (1)$$

where δ is the instantaneous separation distance, v speed, a longitudinal acceleration, $\dot{\theta}$ yaw rate, t time since takeover, and $\mathrm{Rel}(\cdot)$ denotes the relative component along the approach line. For each takeover event, we recorded both the full ACT trajectory and the minimum value, obtained by comparing the values between the three targets within each scenario.

Non-driving related task (NDRT). Phone/Drink/Video labels were derived from the experimenter's keystroke log and video coding: numeric keys triggered a spoken instruction for the next NDRT and generated timestamps; frame-level annotations were reconciled with these anchors.

4.3 Modeling Approach

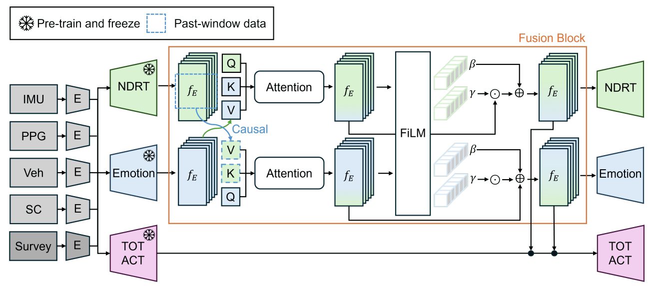

Figure 4: Overview of the MUST fusion architecture. Inputs from multiple modalities are first encoded by task-specific encoders. Within the Fusion Block, we utilize causal cross-attention (Motion→Emotion), complementary attention (Emotion→Motion), and FiLM integration [63] to produce a fused representation for robust takeover prediction.

MUST adopts a two-stage hierarchical framework for predicting takeover performance (TOT and ACT), following the principle of *specialization followed by integration*. Stage 1 develops expert encoders for distinct modalities, while Stage 2 fuses their representations through a bidirectional interaction block.

Stage 1: Expert Encoders. Each modality is first encoded by a specialized network: vehicle telemetry and IMU streams are processed through temporal convolutional encoders, physiological signals (e.g., PPG, HRV features) through a TCN-based encoder, and contextual factors (scenario and pre-survey inputs) through embedding layers. All outputs are projected into a shared latent space for cross-modality comparability. This design enables the model to capture both immediate motion determinants and broader psycho-physiological context.

Stage 2: Fusion and Prediction. Figure 4 illustrates the overall fusion architecture. Encoded features from Stage 1 are combined through a fusion block that jointly represents behavioral and affective states. The fused sequence then directed to task-specific heads: a motion head for takeover action prediction, an affect head for valence–arousal estimation, and an alignment head enforcing consistency between motion and emotion embeddings. Training proceeds in two phases: (1) warm-up updating only the fusion and prediction heads, and (2) joint fine-tuning of all encoders, stabilized by uncertainty-weighted multi-task loss.

Causal Cross-Attention. A central component of the fusion block is the bidirectional causal cross-attention mechanism (Figure 5). Behavior features provide causal context within a past window, reflecting how driver actions unfold. Emotion features query from a delayed window, accounting for the psychological lag between behavior and affective response. In parallel, emotion provides global context to behavior, yielding asymmetric but complementary fusion. To refine alignment, Feature-wise Linear Modulation (FiLM) [63] layers rescale and shift representations across modalities, enhancing coherence. Inspired by findings that urgent contexts often trigger action before conscious appraisal [66], we explicitly model behavior-to-emotion causality in this design.

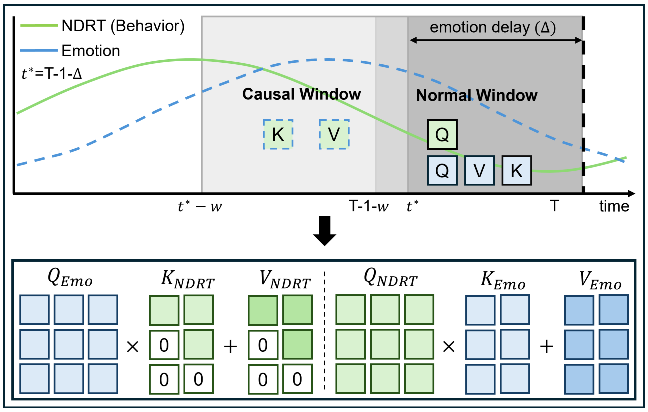

Figure 5: Bidirectional causal cross-attention mechanism. Behavior features (green, NDRT) provide causal context within a past window, while affective features (blue, emotion) query from a delayed window reflecting psychological lag. In parallel, emotion provides global context to behavior, yielding asymmetric but complementary fusion.

This architecture ensures that motion and emotion are not simply concatenated but interact in a temporally structured, causally grounded, and context-aware manner, yielding richer representations of takeover behavior.

5 Performance Evaluation

This section presents the results of our study. We first define evaluation metrics across tasks. We then benchmark MUST against representative state-of-the-art methods, followed by ablation studies that assess input modalities and fusion strategies. Finally, we present task-specific performance, including ACT regression, TOT prediction, NDRT classification, and affective state estimation.

5.1 Evaluation Metrics

We evaluate MUST across diverse tasks using task-specific metrics.

Classification. We report accuracy as the primary metric and *macro-averaged F1* for categorical (NDRT, Valence, Arousal, and TOT) outcomes to describe class imbalance. We also include confusion matrices to inspect error patterns across reaction-speed classes.

Regression (TOQ). For ACT, RMSE is computed between predicted and ground-truth trajectories over the evaluation window; for minACT, RMSE is computed on the scalar minimum value per trial. Because ACT models a full trajectory whereas minACT collapses risk to a single point, RMSE magnitudes are not directly comparable across the two targets; we therefore compare methods within targets. To complement numerical results, regression plots visualized bias and error spread.

5.2 Results and Analysis

5.2.1 Baseline Comparison. Table 2 benchmarks MUST against DeepTake [61], ACTNet [50], and Multi-TBP [28]. While datasets and metrics differ MUST attains TOT accuracy of 91.4%, a level comparable to camera and physiology-based approaches, while relying only on wearable sensing. In addition, unlike prior works that reduce takeover quality to single-point measures such as minACT or lane offset, MUST models full ACT trajectories, which makes it possible to evaluate both transient and sustained risk. This demonstrates that wearable sensing can provide reliable accuracy while also supporting richer safety assessments.

Table 2: Comparison of takeover prediction performance and modalities across representative studies.

Study	Modality	TOT	TOQ
DeepTake [61]	Physio+Glasses*	93.0%	83.0% (Lane Offset)
ACTNet [50]	Camera-based	–	1.6 s (minACT)
Multi-TBP [28]	Camera-based	90.8%	84.6% (TP score)
MUST	**Wearable**	**91.4%**	**2.3 s** (ACT)

*DeepTake uses GSR, Heart Rate, PPG, and Smart Glasses.

5.2.2 Ablation Study. Table 3 shows the contribution of input modalities. Emotion recognition performance dropped significantly without PPG (<0.4), confirming the importance of physiological signals for affective modeling. NDRT classification accuracy fell below 0.8 when IMU signals were removed, indicating the role of motion cues. These findings suggest that affective and behavioral predictions depend on distinct sensing channels. Meanwhile, adding overlapping modalities often degraded performance through redundancy and gradient conflicts, showing that prediction quality improves through careful signal selection than by including all available inputs. All results reported in Tables 3–5 averaged over 10 random seeds. Across seeds, classification metrics (accuracy and F1) varied within ±0.01–0.02, and ACT regression exhibited similarly small fluctuations with RMSE varying by ±0.10–0.15 seconds. The corresponding 95% confidence intervals were narrow and did not overlap across modalities, confirming that the reported effects are stable and not driven by a particular initialization.

Based on these observations, we selected {PPG, SC, survey} as the optimal backbone for emotion recognition and {IMU, veh} for NDRT classification. These combinations consistently outperformed larger sets that included redundant features, demonstrating that compact backbones yield stronger generalization.

Building on these backbones, fusion ablation experiments revealed trade-offs: concatenation and symmetric cross-attention boosted behavioral accuracy but degraded affective accuracy. This performance imbalance primarily stems from the inherent density

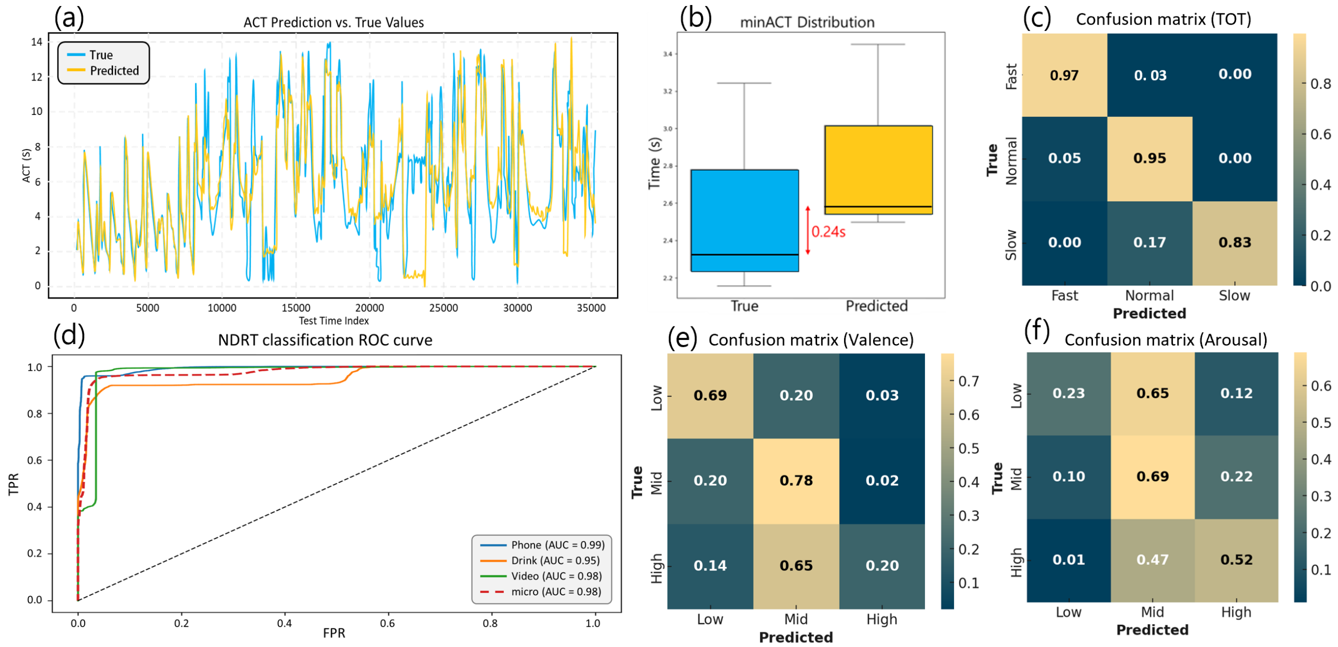

Figure 6: Task-specific performance of the MUST framework. (a) Regression of ACT trajectories against ground truth. (b) Distribution of minACT predictions (box plot). (c) Confusion matrix for TOT classification. (d) ROC curve of NDRT classification. (e) Confusion matrix for valence classification. (f) Confusion matrix for arousal classification.

Table 3: Input modality ablation during pretraining.

Modalities	Emotion (Acc)	NDRT (Acc)
{PPG}	(0.62, 0.52)	–
{PPG, survey}	(0.63, 0.27)	–
{PPG, survey, SC}	(0.63, **0.53**)	–
{PPG, veh}	(0.59, 0.52)	–
{PPG, survey, SC, veh}	(**0.67**, 0.51)	–
All	(0.62, 0.51)	–
{IMU}	–	0.82
{IMU, survey, SC}	–	0.88
{IMU, veh}	–	**0.91**
{IMU, veh, SC}	–	0.91
{IMU, veh, SC, survey}	–	0.90
All	–	0.89

Note: Emotion is reported as (Valence, Arousal). {PPG, survey, SC, veh} excludes IMU, while {IMU, survey, SC, veh} excludes PPG.

Table 4: Fusion ablation results.

Method	NDRT	Emotion	TOT	veh1_ACT
baseline (pre-train)	0.93	(0.63, 0.53)	0.67	2.7
Concat	0.89	(0.65, 0.54)	0.81	2.5
Cross-attn	0.93	(0.58, 0.49)	0.83	2.5
Bi-attn (causal)	0.94	(0.63, 0.51)	0.88	2.4
FiLM	0.91	(0.61, 0.54)	0.75	2.6
MUST	**0.95**	**(0.68, 0.54)**	**0.91**	**2.3**

Note: Emotion is reported as (Valence, Arousal).

mismatch between high-frequency NDRT labels and sparse, discontinuous emotion annotations. By explicitly constraining the data flow through causal cross-attention and modulating with FiLM, MUST achieved balanced improvements across both domains. Consequently, TOT accuracy rose from 0.67 to 0.91, and ACT error fell from 2.7s to 2.3s (Table 4).

5.2.3 Task-specific Performance. We further examine the per-task results of the MUST, focusing on four key dimensions: ACT regression, TOT prediction, NDRT classification, and emotion estimation.

ACT regression. Predicted ACT trajectories closely followed the ground-truth, capturing key trends (Figure 6(a). The model achieved an average RMSE of 2.3 s across scenarios, demonstrating temporal fidelity. Figure 6(b) further shows that predicted minACT values deviated by only 0.24 s on average, capturing reasonably.

TOT prediction. Figure 6(c) shows that TOT classification benefited most from multimodal fusion, with accuracy improving from 0.69 to 0.91 (Macro-F1: 0.90). Most errors occurred near class boundaries, indicating robust modeling of driver readiness.

NDRT classification. As illustrated in Figure 6(d), secondary tasks such as phone use, drinking, and video watching were distinguished with over 90% accuracy, confirming that smartwatch-derived motion and physiological cues provide strong discriminability for behaviors relevant to takeover readiness.

Table 5: Summary of per-task performance metrics.

Task	Metric	Score
NDRT Classification	Accuracy	0.95
Valence Estimation	Acc. / Macro-F1	0.68 / 0.62
Arousal Estimation	Acc. / Macro-F1	0.51 / 0.49
TOT Prediction	Acc. / Macro-F1	0.91 / 0.90
ACT Regression	RMSE (s)	2.3

Affective state estimation. Figures 6(e) and (f) show that valence estimation achieved balanced performance (Acc: 0.68, Macro-F1: 0.62), whereas arousal estimation was more challenging (Acc: 0.51, Macro-F1: 0.49). High and low affective states were classified more consistently than intermediate ones, reflecting the difficulty of modeling subtle dynamics from wearable signals.

5.3 Data Analysis

To examine the relationship between takeover performance and affective responses, we analyzed descriptive statistics across scenarios (Table 6) and conducted inferential tests (ANOVA, correlation).

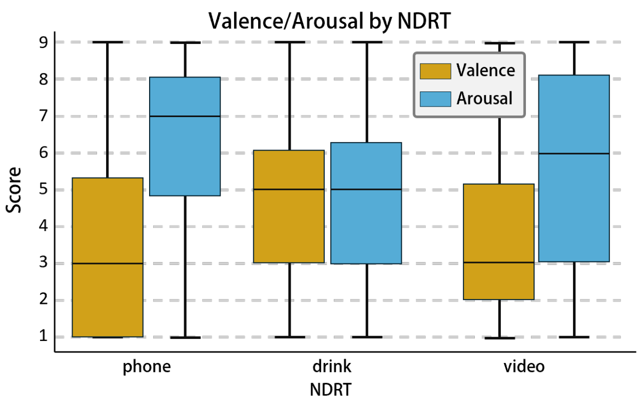

Figure 7: Effects of NDRT type on valence and arousal. Phone use decreased valence and increased arousal, while drinking produced the opposite pattern.

5.3.1 *Effects of NDRTs on Emotion.*

NDRT engagement significantly modulated participants' affective states(Figure 7). ANOVA revealed clear differences in valence ($F = 14.40, p < .001$) and arousal ($F = 12.50, p < .001$). Among conditions, *phone use* elicited the lowest valence ($M = 3.10$) and highest arousal, suggesting an emotionally taxing and disruptive effect. In contrast, *drinking* was associated with the highest valence ($M = 4.61$) and lowest arousal, indicating a comparatively calming effect, while *video watching* fell in between. Interestingly, group-level differences in TOT and minACT were not significant across NDRTs, implying that while performance metrics remained stable, the underlying emotional states were strongly affected. This dissociation highlights that traditional takeover measures may overlook subtle but meaningful affective impacts of secondary tasks.

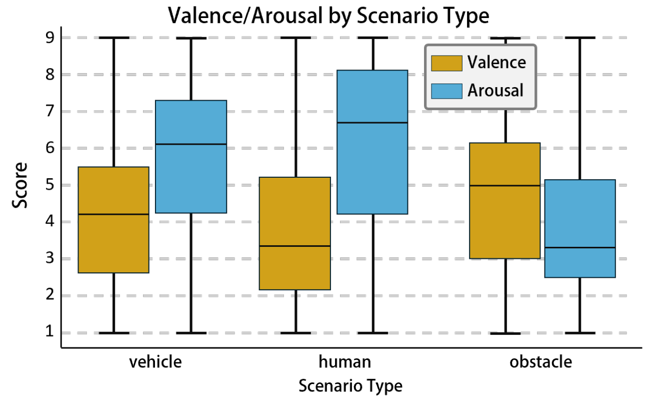

Figure 8: Effects of hazard type on valence and arousal. Pedestrian hazards triggered the most reactive patterns, while static obstacles produced calmer profiles.

5.3.2 *Effects of Event Context.*

Scenario types also shaped emotional responses (Figure 8). Pedestrian hazards elicited the strongest reactivity, with elevated arousal ($M = 4.51 \pm 2.28$) and reduced valence ($M = 4.43 \pm 2.20$), consistent with the sudden and high-salience nature of such events. Static obstacles produced lower arousal and more neutral valence, while vehicle-related hazards showed intermediate patterns. These findings suggest that hazard type shapes the risk profile while shaping drivers' affective states, with direct implications for takeover readiness and the design of context-sensitive TOR cues.

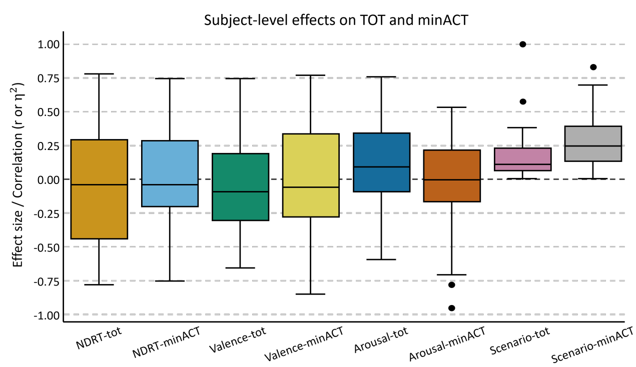

Figure 9: Distribution of subject-level correlations between NDRT type, affective states, and takeover performance metrics. Correlations vary widely across individuals, indicating that some drivers benefit from heightened arousal while others are impaired.

5.3.3 *Statistical Analysis of Takeover Metrics.*

Despite clear group-level trends in affective responses, individual variability dominated takeover performance. As shown in Figure 9, subject-level correlations ranged from $r = -.8$ to $r = +.8$. In some participants, higher arousal was associated with faster and safer responses, whereas in others it coincided with degraded performance. Such dispersion suggests that population averages mask important subgroups, necessitating personalized modeling in takeover systems.

Table 6 summarizes scenario-level descriptive statistics for TOT, minACT, valence, and arousal. According to the analysis, variation in TOT and minACT was greater between individuals than between manipulations. For example, minACT ranged from 0.73 s to 2.53 s. These distributions indicate that population averages provide limited insight on TOQ, which requires the development of personalized modeling in future systems.

Table 6: Descriptive statistics of takeover performance and affective states across hazard types.

Scenario	TOT (s)	minACT (s)	Valence	Arousal
1	2.26 ± 0.25	2.14 ± 0.83	8.02 ± 4.48	5.15 ± 2.09
2	1.71 ± 0.24	1.94 ± 1.08	4.06 ± 2.00	5.73 ± 1.96
3	2.03 ± 0.46	1.06 ± 1.16	4.03 ± 1.97	5.77 ± 2.15
4	1.85 ± 0.29	1.57 ± 1.07	4.30 ± 2.28	5.15 ± 2.41
5	1.41 ± 0.09	1.33 ± 1.02	4.47 ± 2.17	4.51 ± 2.21
6	2.17 ± 0.65	0.73 ± 0.81	4.42 ± 1.99	4.67 ± 2.12
7	2.90 ± 0.56	1.06 ± 0.90	4.94 ± 2.08	4.13 ± 2.10
8	1.73 ± 0.31	1.09 ± 1.09	4.75 ± 2.17	4.30 ± 2.24
9	2.16 ± 0.54	2.02 ± 1.12	4.43 ± 2.20	4.51 ± 2.28
10	1.41 ± 0.24	2.32 ± 0.89	4.36 ± 2.18	4.72 ± 2.32
11	1.59 ± 0.21	2.53 ± 1.00	4.21 ± 2.04	5.06 ± 2.35
12	3.02 ± 0.41	1.76 ± 1.08	4.49 ± 2.21	4.20 ± 2.24
13	1.59 ± 0.06	1.49 ± 1.02	4.73 ± 2.28	3.76 ± 2.38

5.4 Real-time Inference Performance

Inference speed was evaluated on an NVIDIA RTX 4090 (PyTorch 2.1, CUDA 12.1) with batch size 1 (Table 7). The model achieved ∼10.3 ms latency (97 FPS) under FP32, comfortably exceeding real-time requirements for driver monitoring (10–30 Hz). Latency was stable across sequence lengths, confirming efficient scaling of temporal encoders and context pooling. FP16 mixed precision did not yield improvements, showing slightly higher latency due to limited Tensor Core utilization at small batch sizes.

Table 7: Inference latency and throughput across precisions and sequence lengths.

Precision	Seq. Length	Latency (ms)	Throughput (FPS)
FP32	100	10.46 ± 3.00	95.60
FP16	100	14.70 ± 3.79	68.00
FP32	200	10.10 ± 3.57	99.00
FP16	200	14.74 ± 3.94	67.80
FP32	500	10.33 ± 3.56	96.80
FP16	500	14.72 ± 3.76	68.00
FP32	1000	10.35 ± 3.39	96.60
FP16	1000	14.29 ± 3.39	70.00

6 Discussion

This study introduced MUST, which integrates behavioral and affective states into takeover prediction. Whereas prior work treated driver states as secondary inputs, our approach positions behavioral and affective context as central predictors [25, 76, 81]. The discussion elaborates methodological insights from ACT trajectories, HMI considerations, ethical issues, and remaining limitations.

6.1 Methodological Validity: Retrospective Emotion Labeling

A continuous driving flow is essential for capturing authentic affective responses during takeover. Artificially “pausing” the simulation to query driver states would shatter immersion and interrupt the emotional process. This requirement is paramount in our design, which features 13 events within a 15-minute session. In this high-frequency setting, in-situ reporting would cause cognitive fragmentation and artificial reactivity, compromising the ecological validity of the responses [68]. To maintain the integrity of the high-arousal driving experience, we adopted a CRD protocol [11, 62].

While the physiological surge is transient, the model posits that emotion is driven by a cognitive appraisal process accessible through memory reconstruction [74]. To minimize the reconstruction bias inherent in retrospective reporting, we strictly time-locked video cues to each event. These cues serve as a scaffold to rebuild the original context, facilitating accurate episodic recall.

Furthermore, empirical evidence addresses concerns regarding the rapid decay of physiological sensations in high-arousal states. Studies demonstrate that video-stimulated re-immersion effectively reactivates the necessary physiological and contextual memory traces required for valid reporting [12]. By combining immediate post-simulation review with these retrieval cues, our protocol balances the rigorous demands of affective labeling with the need for undisturbed driving behavior.

As additional empirical support, we conducted a validation study comparing CRD-based ratings with in-situ emotion reports collected immediately after each event. The two methods exhibited strong consistency, with a 3-class agreement of 90%, MAE of 0.8, and no non-adjacent class mismatches. Detailed procedures and statistical results are provided in Appendix C. Nevertheless, retrospective self-reports may remain susceptible to cognitive biases such as the peak-end effect, emphasizing salient moments over continuous experience. Thus, while CRD is a pragmatic compromise, it does not fully eliminate the memory–experience gap, and our claims are limited to event-level affective tendencies.

6.2 Beyond Single-Point Metrics: Affective and Behavioral Dynamics of Takeover

Figure 10 illustrates how two drivers, faced with the same hazard, adopted markedly different strategies: one stopped and reversed, while the other smoothly changed lanes. Although their minACT values were nearly identical, this single-point measure collapsed qualitatively different maneuvers into the same outcome [78]. By contrast, ACT trajectories captured whether risk unfolded through unstable fluctuations or gradual decline, providing a richer picture of how drivers experience hazards and when safety interventions are most needed. This trajectory-level view is methodologically more informative and more relevant for HCI, as it suggests opportunities for designing interfaces that respond to the dynamics of unfolding risk rather than isolated moments [10].

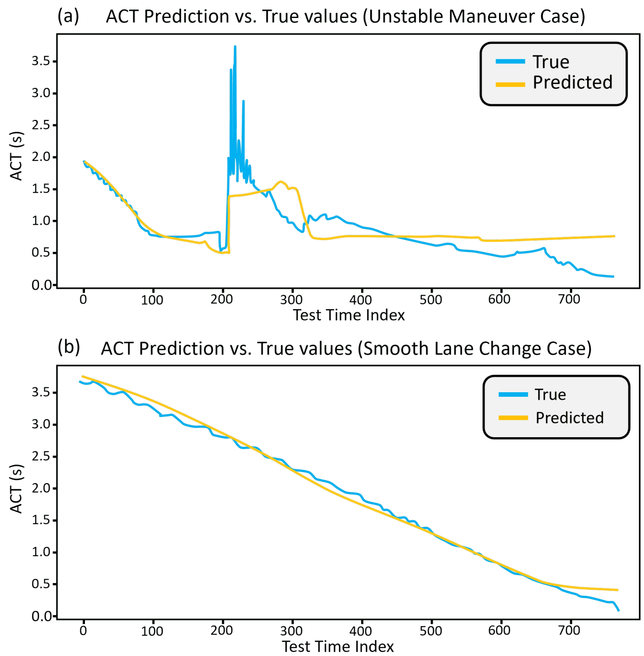

Figure 10: Comparison of ACT and minACT under identical takeover scenarios. (a) Unstable maneuver with oscillatory fluctuations, where minACT fails to distinguish transient instability. (b) Smooth lane change with gradual risk decline, where ACT trajectories capture sustained stabilization.

Yet trajectory metrics alone remain insufficient to fully characterize takeover quality. Much prior work has emphasized performance indicators such as lane offset, TTC, or minACT [25, 45, 50, 61], while giving limited attention to drivers' affective and contextual states. Our findings show that these states are not secondary but central: phone use consistently lowered valence and heightened arousal, whereas drinking was associated with calmer affective profiles. Hazard types further modulated responses, with pedestrians eliciting sharp arousal surges and static obstacles producing steadier but less activated patterns [44]. Moreover, although arousal typically rose and valence declined at TOR onset, recovery trajectories varied considerably across individuals.

From HCI standpoint, this heterogeneity carries direct safety implications. Rapid stabilization correlated with seamless handovers, whereas delayed recovery often manifested as unstable maneuvers and extended risk exposure. These observations point to a concrete design principle: TOR systems should move beyond optimizing immediate reaction times and instead monitor recovery dynamics, using affect-sensitive signals to detect delayed stabilization and adapt support accordingly [44].

6.3 Implications for Adaptive and Personalized HMI

Our case analysis illustrates accurate and problematic predictions (Figure 11). In typical situations involving gradual steering or repeated hazards, MUST reproduced ACT trajectories with high fidelity. By contrast, atypical cases with oscillatory control, such as repeated braking or confusion switching between manual and automated modes, produced unstable predictions. These findings indicate that inference accuracy alone is insufficient for capturing individual variability, reflecting both algorithmic limitations and driver-specific habits.

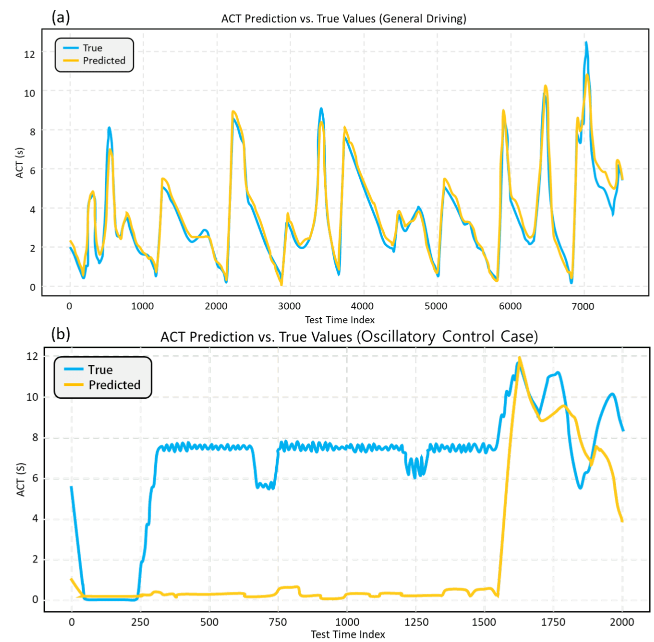

Figure 11: Representative ACT prediction cases. (a) Smooth trajectories across repeated hazard events, where predictions closely track ground truth. (b) Oscillatory control patterns, such as repeated braking or mode-switch toggling, where predictions diverge from actual dynamics.

To probe such variability further, we examined self-reported pre-survey variables (Figure 12). Contrary to expectations, accumulated factors such as confidence, driving experience, or familiarity did not explain takeover alignment. Instead, immediate condition variables, including fatigue and drowsiness, emerged as stronger predictors. Stress also showed a nuanced role: it did not directly correlate with alignment but amplified misalignment when coupled with high confidence, suggesting a moderating effect. We also conducted a demographic subgroup analysis. Takeover alignment showed no meaningful gender differences (female: Spearman $r = 0.09 \pm 0.43$; male: 0.02 ± 0.48, all $p > 0.1$). Regarding age, results were unclear. Participants under 30 showed no correlation ($r = -0.01 \pm 0.47$), whereas the small older group ($n = 4$) showed a nonsignificant trend ($r = 0.36 \pm 0.26$). Together, these findings argue that takeover performance depends more on drivers' real-time cognitive and affective resources than on static traits, underscoring the importance of adaptive HMI that responds dynamically to condition-based variability [64].

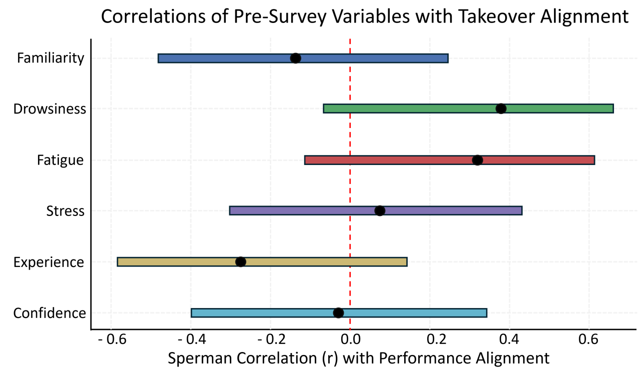

Figure 12: Correlations between self-reported survey variables and takeover alignment. While long-term traits such as confidence, driving experience, and familiarity showed little explanatory power, immediate condition variables including fatigue, drowsiness, and stress were more strongly associated with alignment outcomes.

Building on these findings, we outline practical implications for autonomous vehicle stakeholders. These recommendations aim to bridge the gap between predictive models and real-world safety protocols for automakers, HCI designers, and policymakers.

Automakers: Recovery–informed graded automation

Current binary takeover protocols often cause unstable recovery [51]. To mitigate this, manufacturers should adopt graded automation retention policies rather than instantaneous disengagement. Our trajectory-based ACT prediction facilitates this by distinguishing between smooth stabilization and oscillatory risk patterns in real time. When an unstable trajectory is forecast, the system can temporarily retain partial control (e.g., lane centering), providing a safety buffer during the critical transition window [1, 30].

HCI Designers: State-dependent multimodal alerts

Driver alerts must evolve beyond one-size-fits-all designs [47, 55]. By leveraging real-time physiological cues, TOR strategies can dynamically adapt to the driver's momentary readiness:

- **High-arousal:** Use softer, non-intrusive auditory cues to prevent startle responses and over-correction.
- **Fatigue/Low-vigilance:** Use more salient haptic or visual signals to ensure prompt engagement.

This state-adaptive approach ensures that interfaces respond not only to the external traffic situation but also to the driver's internal condition, reducing mismatches between TOR modality and cognitive state. However, such affective intelligence demands high reliability; for instance, misinterpreting focus as stress could trigger unnecessary interventions, leading to driver frustration. To prevent automation complacency and overreliance, the HMI must act as a transparent collaborator. By providing clear rationales for support and ensuring seamless manual overrides, the system maintains calibrated trust, keeping drivers as active participants even amid uncertain predictions.

Policymakers: Standardization Beyond Reaction Time

Existing regulatory frameworks typically assess takeover safety based on static reaction-time thresholds [41, 76]. However, our empirical results demonstrate that fast reactions (TOT) do not guarantee safe control; indeed, some drivers with fast TOTs exhibited highly unstable maneuvers. Policymakers should evolve safety standards to include stabilization metrics, such as trajectory smoothness and oscillation decay rates, as criteria for ADS certification. Continuous risk-assessment methods, as demonstrated in this work, offer a pathway to validate these rigorous safety standards.

6.4 Ethical and Human Factors Considerations

Integrating affective intelligence into automated driving systems transcends technical feasibility and requires a rigorous ethical foundation. As systems like MUST gain access to sensitive driver states, broader implications involving algorithmic fairness, privacy sovereignty, and the preservation of human agency must be addressed. State-aware HMI carries an inherent risk of reproducing biases embedded in training data; in safety-critical contexts, such disparities do not merely degrade performance but can translate into unequal safety benefits, where underrepresented groups may receive less effective assistance [40, 49]. Ethical deployment therefore demands systematic auditing of algorithmic performance across diverse demographic groups, ethnicities, and neurotypes to ensure equitable and human-centric operation.

At the same time, continuous monitoring of physiological and behavioral signals introduces privacy concerns, raising the possibility that automated vehicles could become surveillance instruments [18, 19]. To safeguard autonomy, adaptive architectures should adopt privacy-by-design principles in which sensitive biometric information functions only as ephemeral inputs for real-time estimation rather than persistent records. Processing data exclusively at the edge without cloud retention embodies strong data minimization, ensuring that drivers retain sovereignty over their biometric information and reinforcing trust as a prerequisite for real-world adoption.

Finally, effective automation introduces the paradox that support intended to enhance safety may inadvertently encourage behavioral adaptation, resulting in overreliance or automation complacency—an issue well documented in longstanding HCI discussions on adaptive systems and the need to preserve user agency [51]. To mitigate this, MUST functions as a selective collaborator rather than a continuous autopilot, intervening only during high-risk recovery to keep the driver actively engaged. If the system misinterprets a driver's state, such as mistaking focus for stress, a seamless manual override allows the driver to instantly dismiss unnecessary assistance. This transparency prevents the system from taking over unnecessarily, ensuring that drivers maintain agency and that human responsibility remains central to the driving task.

Addressing this requires systems that operate not as opaque or paternalistic correctors but as transparent collaborators that actively support situational awareness. Calibrating user trust and preventing misinterpretation requires both explaining the rationale for interventions, like stabilization assistance for oscillatory steering, and explicitly stating system confidence and limitations [39]. Taken together, these considerations highlight that the ethical deployment of affect-aware automation must protect data autonomy, ensure fairness, and foster transparent collaboration that reinforces, rather than diminishes, human responsibility.

Given these ethical concerns, takeover-supportive models must reinforce rather than replace driver responsibility. In this regard, MUST is a state-aware assistant that offers targeted support when unstable, high-risk recovery patterns emerge. By intervening selectively and communicating the basis of its assistance, the system maintains driver engagement while reducing the safety impact of sensing failures or miscalibrated trust.

6.5 Limitations and Future Work

The interpretation of our results requires consideration of several experimental constraints. First, the absence of inertial forces in a static simulator alters vehicle dynamics and likely induces lower physiological arousal than real-world emergencies, limiting ecological validity. Second, despite robust validation (Appendix A), PPG sensors remain susceptible to artifacts from contact pressure and skin pigmentation. As our cohort was limited to East Asian participants, further empirical confirmation across the full Fitzpatrick scale is required to ensure sensing fairness. Third, the model showed reduced robustness in rare scenarios like oscillatory mode switching, necessitating architectures more resilient to non-standard dynamics.

A significant limitation also lies in the demographic homogeneity of the participant pool. The study cohort was predominantly composed of young, East Asian male adults (20M/9F; $M_{age} = 25.1$), which may not fully capture the physiological and behavioral variance found in the broader population. Although initial subgroup analysis showed no gender bias, the small sample of older participants and lack of ethnic diversity leave the potential for algorithmic bias unaddressed. Therefore, these limitations necessitate cautious interpretation of our findings, particularly with respect to safety-critical deployment scenarios.

Future work will address these gaps by extending MUST to on-road settings and expanding the participant spectrum across diverse ages, genders, and ethnicities. We aim to develop robust model architectures capable of handling edge cases and to implement privacy-by-design edge processing for reliable, real-world adoption. By incorporating more diverse neurotypes and physical traits, we seek to refine MUST into a truly equitable and inclusive driver-monitoring system.

7 Conclusion

This paper presents MUST, a smartwatch-based framework that predicts takeover outcomes and driver states in real-time through unobtrusive multimodal fusion. Unlike prior approaches that rely on specialized sensors or treat driver states as ancillary, MUST demonstrates that wearable devices already available in everyday life can capture both behavioral and affective dynamics with sufficient fidelity.

The results show that effective takeover depends on both affective regulation and maneuver stability, as well as rapid reactions. By linking these dimensions to trajectory based safety metrics, the study contributes methodological advances and design insights for HCI. Wearable driven sensing offers a pathway toward affect-aware TOR interfaces that personalize feedback, support recovery, and preserve driver agency and trust.

As an exploratory step, this work establishes feasibility while pointing to clear directions for future research. Key next steps include validation in naturalistic contexts, recruitment of more diverse participants, and refinement of adaptive strategies that balance personalization with transparency and ethical safeguards. Such developments can move takeover systems beyond prediction toward truly supportive and trustworthy interaction in real driving environments.

Acknowledgments

This work was primarily supported by Hyundai Motor Company (No. T104600124020045) and in part by the Korean ARPA-H Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (No. RS-2025-25455839). It was also supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2023-00229074). Finally, we are grateful to all the study participants for their invaluable time and contribution to this research.

References

[1] David A. Abbink, Mark Mulder, and Erwin R. Boer. 2012. Haptic shared control: smoothly shifting control authority? *Cognition, Technology & Work* 14, 1 (2012), 19–28. doi:10.1007/s10111-011-0192-5

[2] Shabnam Abtahi, Behnoosh Hariri, and Shervin Shirmohammadi. 2011. Driver drowsiness monitoring based on yawning detection. In *Proceedings of the 2011 IEEE International Instrumentation and Measurement Technology Conference (I2MTC '11)*. IEEE, Binjiang, China, 1–4. doi:10.1109/IMTC.2011.5944101

[3] Feryal A. Alskafi, Ahsan H. Khandoker, and Herbert F. Jelinek. 2021. A Comparative Study of Arousal and Valence Dimensional Variations for Emotion Recognition Using Peripheral Physiological Signals Acquired from Wearable Sensors. In *Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC '21)*. IEEE, Mexico City, Mexico, 1104–1107. doi:10.1109/EMBC46164.2021.9630759

[4] Jackie Ayoub, Na Du, X. Jessie Yang, and Feng Zhou. 2022. Predicting Driver Takeover Time in Conditionally Automated Driving. *IEEE Transactions on Intelligent Transportation Systems* 23, 7 (2022), 9580–9589. doi:10.1109/TITS.2022.3154329

[5] Roza Eleni Barka and Ioannis Politis. 2024. Driving into the Future: A Scoping Review of Smartwatch Use for Real-Time Driver Monitoring. *Transportation Research Interdisciplinary Perspectives* 25 (2024), 101098. doi:10.1016/j.trip.2024.101098

[6] Aayush Bhardwaj, Ankit Gupta, Pallav Jain, Asha Rani, and Jyoti Yadav. 2015. Classification of human emotions from EEG signals using SVM and LDA Classifiers. In *Proceedings of the 2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN '15)*. IEEE, Noida, India, 180–185. doi:10.1109/SPIN.2015.7095376

[7] U. Bilotti, C. Bisogni, M. De Marsico, and S. Tramonte. 2024. Multimodal Emotion Recognition via Convolutional Neural Networks: Comparison of different strategies on two multimodal datasets. *Engineering Applications of Artificial Intelligence* 130 (2024), 107708. doi:10.1016/j.engappai.2023.107708

[8] Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25, 1 (1994), 49–59. doi:10.1016/0005-7916(94)90063-9

[9] Stefan Brandenburg and Lewis Chuang. 2019. Take-over requests during highly automated driving: How should they be presented and under what conditions? *Transportation Research Part F: Traffic Psychology and Behaviour* 66 (2019), 214–225. doi:10.1016/j.trf.2019.08.023

[10] Michael Braun, Bastian Pfleging, and Florian Alt. 2018. A Survey to Understand Emotional Situations on the Road and What They Mean for Affective Automotive UIs. *Multimodal Technologies and Interaction* 2, 4 (2018), 75. doi:10.3390/mti2040075

[11] Anders Bruun, Effie Lai-Chong Law, Matthias Heintz, and Poul Svante Eriksen. 2016. Asserting Real-Time Emotions through Cued-Recall: Is it Valid?. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction (NordiCHI '16)*. ACM, New York, NY, USA, Article 41, 1–10. doi:10.1145/2971485.2971516

[12] Anders Bruun, Effie Lai-Chong Law, Thomas Dyhre Nielsen, and Matthias Heintz. 2021. Do You Feel the Same? On the Robustness of Cued-Recall Debriefing for User Experience Evaluation. *ACM Transactions on Computer-Human Interaction (TOCHI)* 28, 4, Article 25 (2021), 45 pages. doi:10.1145/3453479

[13] Shengli Cao, Peihua Feng, Wei Kang, Zeyi Chen, and Bo Wang. 2025. Optimized driver fatigue detection method using multimodal neural networks. *Scientific Reports* 15, 1 (2025), 12240. doi:10.1038/s41598-025-86709-1

[14] Yetong Cao, Fan Li, Xiaochen Liu, Song Yang, and Yu Wang. 2023. Towards Reliable Driver Drowsiness Detection Leveraging Wearables. *ACM Transactions on Sensor Networks* 19, 2, Article 39 (2023), 23 pages. doi:10.1145/3560821

[15] Haolin Chen, Xiaohua Zhao, Haijian Li, Jianguo Gong, and Qiang Fu. 2024. Predicting driver's takeover time based on individual characteristics, external environment, and situation awareness. *Accident Analysis & Prevention* 203 (2024), 107601. doi:10.1016/j.aap.2024.107601

[16] Sheldon Cohen, Tom Kamarck, and Robin Mermelstein. 1983. A Global Measure of Perceived Stress. *Journal of Health and Social Behavior* 24, 4 (1983), 385–396. doi:10.2307/2136404

[17] S. Costantini, M. Chiappini, G. Malerba, C. Dei, A. Falivene, S. Arlati, V. Colombo, E. Biffi, and F. A. Storm. 2023. Wrist-Worn Sensor Validation for Heart Rate Variability and Electrodermal Activity Detection in a stressful Driving Environment. *Sensors* 23, 20 (2023), 8423. doi:10.3390/s23208423

[18] Roddy Cowie. 2015. Ethical Issues in Affective Computing. In *The Oxford Handbook of Affective Computing*, Rafael A. Calvo, Sidney D'Mello, Jonathan Gratch, and Arvid Kappas (Eds.). Oxford University Press, Oxford, UK, 334–348. doi:10.1093/oxfordhb/9780199942237.013.006

[19] Rory Coyne, Michelle Hanlon, Alan F. Smeaton, Peter Corcoran, and Jane C. Walsh. 2024. Understanding drivers' perspectives on the use of driver monitoring systems during automated driving: Findings from a qualitative focus group study. *Transportation Research Part F: Traffic Psychology and Behaviour* 105 (2024), 321–335. doi:10.1016/j.trf.2024.07.015

[20] Omid Dehzangi, Vaishali Sahu, Vikas Rajendra, and Mojtaba Taherisadr. 2019. GSR-based distracted driving identification using discrete & continuous decomposition and wavelet packet transform. *Smart Health* 14 (2019), 100085. doi:10.1016/j.smhl.2019.100085

[21] Nachiket Deo and Mohan M. Trivedi. 2020. Looking at the Driver/Rider in Autonomous Vehicles to Predict Take-Over Readiness. *IEEE Transactions on Intelligent Vehicles* 5, 1 (2020), 41–52. doi:10.1109/TIV.2019.2955364

[22] Andrea Di Credico, David Perpetuini, Pascal Izzicupo, Giulia Gaggi, Daniela Cardone, Chiara Filippini, Arcangelo Merla, Barbara Ghinassi, and Angela Di Baldassarre. 2022. Estimation of Heart Rate Variability Parameters by Machine Learning Approaches Applied to Facial Infrared Thermal Imaging. *Frontiers in Cardiovascular Medicine* 9 (2022), 893374. doi:10.3389/fcvm.2022.893374

[23] James Diebel. 2006. *Representing Attitude: Euler Angles, Unit Quaternions, and Rotation Vectors.* Technical Report TR-CST-05-02. Stanford University, Stanford, CA, USA. https://www.astro.rug.nl/software/kapteyn-beta/_downloads/attitude.pdf

[24] Daghan Dogan, Seta Bogosyan, and Tankut Acarman. 2022. Evaluation of Takeover Time Performance of Drivers in Partially Autonomous Vehicles Using a Wearable Sensor. *Journal of Sensors* 2022, 1 (2022), 7924444. doi:10.1155/2022/7924444

[25] Na Du, Feng Zhou, Elizabeth M. Pulver, Dawn M. Tilbury, Lionel P. Robert, Anuj K. Pradhan, and X. Jessie Yang. 2020. Examining the effects of emotional valence and arousal on takeover performance in conditionally automated driving. *Transportation Research Part C: Emerging Technologies* 112 (2020), 78–87. doi:10.1016/j.trc.2020.01.006

[26] Hesham M. Eraqi, Yehya Abouelnaga, Mohamed H. Saad, and Mohamed N. Moustafa. 2019. Driver distraction identification with an ensemble of convolutional neural networks. *Journal of Advanced Transportation* 2019, 1 (2019), 4125865. doi:10.1155/2019/4125865

[27] Carlos H. Espino-Salinas, Huizilopoztli Luna-García, José M. Celaya-Padilla, Cristian Barría-Huidobro, Nadia Karina Gamboa Rosales, David Rondon, and Klinge Orlando Villalba-Condori. 2024. Multimodal driver emotion recognition using motor activity and facial expressions. *Frontiers in Artificial Intelligence* 7 (2024), 1467051. doi:10.3389/frai.2024.1467051

[28] Xinyu Feng, Yu Gu, Yuming Lin, and Yaojun Cai. 2025. Exploring the Impact of Drivers' Emotion and Multi-task Learning on Takeover Behavior Prediction in Multimodal Environment. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 490, 1–7. doi:10.1145/3706599.3719854

[29] Venere Ferraro and Secil Ugur. 2011. Designing wearable technologies through a user centered approach. In *Proceedings of the 2011 Conference on Designing Pleasurable Products and Interfaces (DPPI '11)*. Association for Computing Machinery, New York, NY, USA, Article 5, 8 pages. doi:10.1145/2347504.2347510

[30] Frank Flemisch, Matthias Heesen, Tobias Hesse, Johann Kelsch, Anna Schieben, and Johannes Beller. 2012. Towards a dynamic balance between humans and automation: authority, ability, responsibility and control in shared and cooperative control situations. *Cognition, Technology & Work* 14, 1 (2012), 3–18. doi:10.1007/s10111-011-0191-6

[31] Alícia Freitas, Rute Almeida, Hernâni Gonçalves, Glória Conceição, and Alberto Freitas. 2024. Monitoring fatigue and drowsiness in motor vehicle occupants using electrocardiogram and heart rate — A systematic review. *Transportation Research Part F: Traffic Psychology and Behaviour* 103 (2024), 586–607. doi:10.1016/j.trf.2024.05.008

[32] Rui Fu, Tao Huang, Mingyue Li, Qinyu Sun, and Yunxing Chen. 2023. A multimodal deep neural network for prediction of the driver's focus of attention based on anthropomorphic attention mechanism and prior knowledge. *Expert Systems with Applications* 214 (2023), 119157. doi:10.1016/j.eswa.2022.119157

[33] Adnan Ghaderi, Javad Frounchi, and Alireza Farnam. 2015. Machine learning-based signal processing using physiological signals for stress detection. In *Proceedings of the 2015 22nd Iranian Conference on Biomedical Engineering (ICBME '15)*. IEEE, Tehran, Iran, 93–98. doi:10.1109/ICBME.2015.7404123

[34] Letizia Gionfrida, Daekyum Kim, Davide Scaramuzza, Dario Farina, and Robert D. Howe. 2024. Wearable robots for the real world need vision. *Science Robotics* 9, 90 (2024), eadj8812. doi:10.1126/scirobotics.adj8812

[35] Christian Gold, Daniel Damböck, Lutz Lorenz, and Klaus Bengler. 2013. "Take over!" How long does it take to get the driver back into the loop?. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting (HFES '13)*. SAGE Publications, Los Angeles, CA, USA, 1938–1942. doi:10.1177/1541931213571433

[36] Yu Gu, Yibing Weng, Yantong Wang, Meng Wang, Guohang Zhuang, Jinyang Huang, Xiaolan Peng, Liang Luo, and Fuji Ren. 2024. EmoTake: Exploring Drivers' Emotion for Takeover Behavior Prediction. *IEEE Transactions on Affective Computing* 15, 4 (2024), 2112–2127. doi:10.1109/TAFFC.2024.3399328

[37] Amina Guettas, Soheyb Ayad, and Okba Kazar. 2020. Driver State Monitoring System: A Review. In *Proceedings of the 4th International Conference on Big Data and Internet of Things (BDIoT '19)*. Association for Computing Machinery, New York, NY, USA, Article 25, 1–6. doi:10.1145/3372938.3372966

[38] Gelareh Hajian, Ali Abedi, Bing Ye, Jennifer Campos, and Alex Mihailidis. 2025. Dynamics of Affective States During Takeover Requests in Conditionally Automated Driving Among Older Adults with and Without Cognitive Impairment. In *ArtifAI for Aging Rehabilitation and Intelligent Assisted Living*, Shehroz S. Khan, Luca Romeo, and Ali Abedi (Eds.). Springer Nature Singapore, Singapore, 29–43. doi:10.1007/978-981-95-0568-5_3

[39] Daniël D. Heikoop, Joost C. F. de Winter, Bart van Arem, and Neville A. Stanton. 2018. Effects of mental demands on situation awareness during platooning: A driving simulator study. *Transportation Research Part F: Traffic Psychology and Behaviour* 58 (2018), 193–209. doi:10.1016/j.trf.2018.04.015

[40] Mohammad Mehdi Hosseini, Ali Pourramezan Fard, and Mohammad H. Mahoor. 2025. Faces of fairness: Examining bias in facial expression recognition datasets and models. *arXiv preprint arXiv:2502.11049* abs/2502.11049 (2025), 1–15. https://arxiv.org/abs/2502.11049

[41] Wenbo Hu, Tingru Zhang, Yan Zhang, and Alan Hoi Shou Chan. 2024. Non-driving-related tasks and drivers' takeover time: A meta-analysis. *Transportation Research Part F: Traffic Psychology and Behaviour* 103 (2024), 623–637. doi:10.1016/j.trf.2024.05.012

[42] International Organization for Standardization. 2020. Road vehicles — Human performance and state in the context of automated driving. ISO Standard 21959:2020.

[43] Ishani Janveja, Akshay Nambi, Shruthi Bannur, Sanchit Gupta, and Venkat Padmanabhan. 2020. InSight: Monitoring the State of the Driver in Low-Light Using Smartphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 4, 3, Article 83 (2020), 29 pages. doi:10.1145/3411819

[44] Myounghoon Jeon. 2017. Emotions and Affect in Human Factors and Human–Computer Interaction: Taxonomy, Theories, Approaches, and Methods. In *Emotions and Affect in Human Factors and Human-Computer Interaction*. Academic Press, London, UK, 3–26. doi:10.1016/B978-0-12-801851-4.00001-X

[45] Lisheng Jin, Xingchen Liu, Baicang Guo, Zhuotong Han, Yinlin Wang, Yuan Cao, Xiao Yang, and Jian Shi. 2025. Impact of non-driving related task types, request modalities, and automation on driver takeover: A meta-analysis. *Safety Science* 181 (2025), 106704. doi:10.1016/j.ssci.2024.106704

[46] Sungchul Kang, Woojae Choi, Cheon Y. Park, et al. 2023. K-EmoPhone: A Mobile and Wearable Dataset with In-Situ Emotion, Stress, and Attention Labels. *Scientific Data* 10, 351 (2023), 1–15. doi:10.1038/s41597-023-02248-2

[47] Robert Kaufman, Jean Costa, and Everlyne Kimani. 2024. Effects of multimodal explanations for autonomous driving on driving performance, cognitive load, expertise, confidence, and trust. *Scientific Reports* 14, 1 (2024), 13061. doi:10.1038/s41598-024-62052-9

[48] Daniel Laidig and Thomas Seel. 2023. VQF: Highly accurate IMU orientation estimation with bias estimation and magnetic disturbance rejection. *Information Fusion* 91 (2023), 187–204. doi:10.1016/j.inffus.2022.10.014

[49] Xinyue Li, Zhenpeng Chen, Jie M. Zhang, Federica Sarro, Ying Zhang, and Xuanzhe Liu. 2025. Bias behind the Wheel: Fairness Testing of Autonomous Driving Systems. *ACM Transactions on Software Engineering and Methodology (TOSEM)*

34, 3, Article 82 (2025), 24 pages. doi:10.1145/3702989
[50] Weimin Liu, Qingkun Li, Wenjun Wang, Zhenyuan Wang, Chao Zeng, and Bo Cheng. 2024. Deep Learning Based Take-Over Performance Prediction and Its Application on Intelligent Vehicles. *IEEE Transactions on Intelligent Vehicles* 9 (2024), 1–15. doi:10.1109/TIV.2024.3374528
[51] Zhenji Lu, Riender Happee, Christopher D. D. Cabrall, Miltos Kyriakidis, and Joost C. F. de Winter. 2016. Human factors of transitions in automated driving: A general framework and literature survey. *Transportation Research Part F: Traffic Psychology and Behaviour* 43 (2016), 183–198. doi:10.1016/j.trf.2016.10.007
[52] R. Duncan Luce. 1986. *Response Times: Their Role in Inferring Elementary Mental Organization*. Oxford University Press, Oxford, UK.
[53] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Chi. 2018. Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*. Association for Computing Machinery, New York, NY, USA, 1930–1939. doi:10.1145/3219819.3220007
[54] Yunsheng Ma and Ziran Wang. 2024. ViT-DD: Multi-Task Vision Transformer for Semi-Supervised Driver Distraction Detection. In *Proceedings of the 2024 IEEE Intelligent Vehicles Symposium (IV '24)*. IEEE, Washington, DC, USA, 417–423. doi:10.1109/IV55156.2024.10588802
[55] Kimberly D. Martinez and Gaojian Huang. 2024. From young to old: The effects of information presentation type, multimodal display, and age on situation awareness and processing time in automated vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour* 103 (2024), 128–140. doi:10.1016/j.trf.2024.04.002
[56] Jolieke Mesken, Marjan P. Hagenzieker, Talib Rothengatter, and Dick de Waard. 2007. Frequency, determinants, and consequences of different drivers' emotions: An on-the-road study using self-reports, (observed) behaviour, and physiology. *Transportation Research Part F: Traffic Psychology and Behaviour* 10, 6 (2007), 458–475. doi:10.1016/j.trf.2007.05.001
[57] Guangrui Mu, Yang Zhang, Zhonghong Yan, Qinming Yu, and Qifan Wang. 2025. Recent advancements in wearable sensors: integration with machine learning for human-machine interaction. *RSC Advances* 15, 10 (2025), 7844–7854. doi:10.1039/d5ra00167f
[58] National Highway Traffic Safety Administration. 2022. *Automated Driving Systems (ADS) Crash Reporting*. Technical Report. U.S. Department of Transportation, Washington, DC, USA. https://www.nhtsa.gov/sites/nhtsa.gov/files/2022-06/Standing-General-Order-Crash-Report-Summary-2022.pdf
[59] Eilidh Noyes, Josh P. Davis, Nikolay Petrov, Katie L. H. Gray, and Kay L. Ritchie. 2021. The effect of face masks and sunglasses on identity and expression recognition with super-recognizers and typical observers. *Royal Society Open Science* 8, 3 (2021), 201169. doi:10.1098/rsos.201169
[60] Juan Diego Ortega, Neslihan Kose, Paola Cañas, Min-An Chao, Alexander Unnervik, Marcos Nieto, Oihana Otaegui, and Luis Salgado. 2020. DMD: A Large-Scale Multi-modal Driver Monitoring Dataset for Attention and Alertness Analysis. In *Computer Vision – ECCV 2020 Workshops*, Adrien Bartoli and Andrea Fusiello (Eds.). Springer International Publishing, Cham, Switzerland, 387–405. doi:10.1007/978-3-030-66823-5_27
[61] 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 (CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 103, 14 pages. doi:10.1145/3411764.3445563
[62] Gabriele Pätsch, Thomas Mandl, and Christa Womser-Hacker. 2014. Using sensor graphs to stimulate recall in retrospective think-aloud protocols. In *Proceedings of the 5th Information Interaction in Context Symposium (IIiX '14)*. Association for Computing Machinery, New York, NY, USA, 303–307. doi:10.1145/2637002.2637048
[63] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2018. FiLM: visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI '18)*. AAAI Press, Palo Alto, California, USA, Article 483, 10 pages. doi:10.1609/aaai.v32i1.11671
[64] Martin Pielot, Karen Church, and Rodrigo de Oliveira. 2014. An in-situ study of mobile phone notifications. In *Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices & Services (MobileHCI '14)*. Association for Computing Machinery, New York, NY, USA, 233–242. doi:10.1145/2628363.2628364
[65] Lukasz Piwek, David A. Ellis, Sally Andrews, and Adam Joinson. 2016. The Rise of Consumer Health Wearables: Promises and Barriers. *PLoS Medicine* 13, 2 (2016), e1001953. doi:10.1371/journal.pmed.1001953
[66] Christian H. Poth, Anja Petersen, and Wilfried Kunde. 2021. Urgency forces stimulus-driven action by overcoming cognitive control. *eLife* 10 (2021), e73682. doi:10.7554/eLife.73682
[67] R. Gnana Praveen and Jahangir Alam. 2024. Recursive Joint Cross-Modal Attention for Multimodal Fusion in Dimensional Emotion Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. Computer Vision Foundation, Seattle, WA, USA, 4803–4813. https://openaccess.thecvf.com/content/CVPR2024W/CVPRW/html/Praveen_Recursive_Joint_Cross-Modal_Attention_for_Multimodal_Fusion_in_Dimensional_Emotion_CVPRW_2024_paper.html
[68] J. C. Rann and A. Almor. 2022. Effects of verbal tasks on driving simulator performance. *Cognitive Research: Principles and Implications* 7, 12 (2022), 1–15. doi:10.1186/s41235-022-00357-x
[69] Mahdi Rezaei and Mohsen Azarmi. 2025. Driver-Net: Multi-Camera Fusion for Assessing Driver Take-Over Readiness in Automated Vehicles. In *Proceedings of the 2025 IEEE Intelligent Vehicles Symposium (IV '25)*. IEEE, Cluj-Napoca, Romania, 1841–1848. doi:10.1109/IV64158.2025.11097677
[70] Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. 2019. Latent multi-task architecture learning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI '19)*. AAAI Press, Palo Alto, California, USA, Article 592, 8 pages. doi:10.1609/aaai.v33i01.33014822
[71] Adil Ali Saleem, Hafeez Ur Rehman Siddiqui, Muhammad Amjad Raza, Furqan Rustam, Sandra Dudley, and Imran Ashraf. 2023. A systematic review of physiological signals based driver drowsiness detection systems. *Cognitive Neurodynamics* 17, 5 (2023), 1229–1259. doi:10.1007/s11571-022-09898-9
[72] Kevin Joel Salubre and Dan Nathan-Roberts. 2021. Takeover Request Design in Automated Driving: A Systematic Review. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 65, 1 (2021), 868–872. doi:10.1177/1071181321651296
[73] Andres Sanchez-Comas, Kåre Synnes, Diego Molina-Estren, Alexander Troncoso-Palacio, and Zhoe Comas-González. 2021. Correlation Analysis of Different Measurement Places of Galvanic Skin Response in Test Groups Facing Pleasant and Unpleasant Stimuli. *Sensors* 21, 12 (2021), 4210. doi:10.3390/s21124210
[74] Klaus R. Scherer. 2005. What Are Emotions? And How Can They Be Measured? *Social Science Information* 44, 4 (2005), 695–729. doi:10.1177/0539018405058216
[75] Elisabeth Shi, Tom Michael Gasser, Andre Seeck, and Rico Auerswald. 2020. The Principles of Operation Framework: A Comprehensive Classification Concept for Automated Driving Functions. *SAE International Journal of Connected and Automated Vehicles* 3, 1 (2020), 12–03–01–0003. doi:10.4271/12-03-01-0003
[76] Sónia Soares, António Lobo, Sara Ferreira, Liliana Cunha, and António Couto. 2021. Takeover performance evaluation using driving simulation: a systematic review and meta-analysis. *European Transport Research Review* 13, 1 (2021), 47. doi:10.1186/s12544-021-00505-2
[77] Roberto Valenti, Nicu Sebe, and Theo Gevers. 2012. Combining Head Pose and Eye Location Information for Gaze Estimation. *IEEE Transactions on Image Processing* 21, 2 (2012), 802–815. doi:10.1109/TIP.2011.2162740
[78] Suvin P. Venthuruthiyil and Mallikarjuna Chunchu. 2022. Anticipated Collision Time (ACT): A two-dimensional surrogate safety indicator for trajectory-based proactive safety assessment. *Transportation Research Part C: Emerging Technologies* 139 (2022), 103655. doi:10.1016/j.trc.2022.103655
[79] Tao Wang and Hanling Zhang. 2022. Using Wearable Devices for Emotion Recognition in Mobile Human-Computer Interaction: A Review. In *HCI International 2022 - Late Breaking Papers: Multimodality in Advanced Interaction Environments*. Springer, Berlin, Heidelberg, 205–227. doi:10.1007/978-3-031-17618-0_16
[80] Xiaoming Wang, Hongliu Yu, Søren Kold, Ole Rahbek, and Shaoping Bai. 2023. Wearable sensors for activity monitoring and motion control: A review. *Biomimetic Intelligence and Robotics* 3, 1 (2023), 100089. doi:10.1016/j.birob.2023.100089
[81] Yantong Wang, Yu Gu, Tong Quan, Jiaoyun Yang, Mianxiong Dong, Ning An, and Fuji Ren. 2025. ViE-Take: A Vision-Driven Multi-Modal Dataset for Exploring the Emotional Landscape in Takeover Safety of Autonomous Driving. *Research* 8 (2025), 0603. doi:10.34133/research.0603
[82] Bradley W. Weaver and Patricia R. DeLucia. 2022. A Systematic Review and Meta-Analysis of Takeover Performance During Conditionally Automated Driving. *Human Factors* 64, 7 (2022), 1227–1260. doi:10.1177/0018720820976476
[83] Sazali Yaacob, Nur Afrina Izzati Affandi, Pranesh Krishnan, Amir Rasyadan, Muhyi Yaakop, and Firdaus Mohamed. 2020. Drowsiness detection using EEG and ECG signals. In *Proceedings of the 2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET '20)*. IEEE, Kota Kinabalu, Malaysia, 1–5. doi:10.1109/IICAIET49801.2020.9257867
[84] Sol Hee Yoon and Yong Gu Ji. 2019. Non-driving-related tasks, workload, and takeover performance in highly automated driving contexts. *Transportation Research Part F: Traffic Psychology and Behaviour* 60 (2019), 620–631. doi:10.1016/j.trf.2018.11.015
[85] Kathrin Zeeb, Axel Buchner, and Axel Schrauf. 2016. Is take-over time all that matters? The impact of visual-cognitive load on driver take-over quality after conditionally automated driving. *Accident Analysis & Prevention* 92 (2016), 230–239. doi:10.1016/j.aap.2016.04.002
[86] Lu Zheng, Ye Zhu, Jun Zhang, Zhiqiang Li, Rui Wang, and Lei Zhang. 2020. Sensor-Based and Vision-Based Human Activity Recognition: A Comprehensive Survey. *Pattern Recognition* 108 (2020), 107561. doi:10.1016/j.patcog.2020.107561
[87] Daoming Zong, Chaoyue Ding, Baoxiang Li, Jiakui Li, and Ken Zheng. 2024. Balancing Multimodal Learning via Online Logit Modulation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI '24)*. International Joint Conferences on Artificial Intelligence Organization, California, USA, 5753–5761. doi:10.24963/ijcai.2024/636

A PPG Robustness and Cross-Device Validation

A.1 Experimental Protocol

We conducted a validation study with 15 participants ($N = 15$; 10 males, 5 females) to evaluate the noise robustness of our custom wrist-worn PPG device against an ECG system (Movesense HR2) and a commercial smartwatch (Galaxy Watch 7). To ensure the direct relevance of this validation to our main results, we re-recruited participants from the original study pool based on their availability for follow-up testing. The average age of this sub-group was 25.1 years ($SD = 1.8$), closely mirroring the demographic characteristics of the main study's population. The participants in this validation study were all of East Asian descent, consistent with the primary study cohort. While this ensures internal consistency for the current dataset, we acknowledge the importance of sensor reliability across broader skin tone variations (e.g., Fitzpatrick scale V-VI) as a critical factor for future generalizability.

The protocol consisted of four 1-minute tasks designed to introduce varying levels of motion artifacts: (i) seated baseline (minimal motion), (ii) steering-wheel turning (rotational motion), (iii) video watching (low motion), and (iv) phone use (irregular, high motion). The entire sequence was repeated twice, yielding 8 minutes of data per participant.

For analysis, HR and HRV metrics were computed using 120-second sliding windows. While the custom device and ECG provided raw inter-beat intervals (IBIs), the Galaxy Watch did not grant access to raw IBI data due to proprietary signal processing. Consequently, the commercial device was excluded from HRV analysis and used solely for HR benchmarking.

A.2 Motion Robustness and Signal Quality

The custom PPG sensor demonstrated high signal integrity across all conditions. Even during high-motion tasks (e.g., phone use), the signal loss rate remained below 3%. No significant signal degradation or burst artifacts were observed, and HRV metrics showed minimal deviation (< 7% relative change) compared to the baseline.

These findings indicate that the optical and mechanical design of the device effectively suppresses motion-induced PPG noise.

A.3 Validation Results

Table 8 summarizes the agreement between the custom PPG device and the ECG reference. The custom device achieved high correlation across all metrics ($r > 0.91$) with low MAE.

Table 8: HR and HRV agreement between ECG and custom PPG ($N = 15$).

Metric	ECG (M ± SD)	PPG (M ± SD)	MAE	Corr (r)
BPM	71.76 ± 10.44	73.20 ± 11.27	1.49	0.937
RMSSD (ms)	84.73 ± 47.52	82.79 ± 53.87	12.79	0.926
SDNN (ms)	89.39 ± 38.21	91.93 ± 40.13	11.35	0.914

Statistical analysis showed no significant differences between the custom PPG and ECG-derived metrics. In contrast, the Galaxy Watch 7 exhibited an HR deviation of approximately 3–5 bpm relative to ECG—likely due to commercial smoothing—while HRV could not be validated because raw IBI data were not accessible, preventing continuous peak/IBI acquisition.

Therefore, while current commercial devices have practical constraints that limit direct integration into our autonomous-driving simulation pipeline, the accuracy and robustness of our custom wrist-worn device under motion suggest strong potential for using wearable sensing to support takeover-readiness assessment in automated driving.

B Sensitivity Analysis of TOT Discretization

B.1 Experimental Overview

To assess the robustness of our labeling strategy, we compared the proposed quartile-based discretization (Fast: bottom 25%, Normal: middle 50%, Slow: top 25%) against three alternative strategies:

- **Median Split** (Binary: Fast/Slow),
- **Absolute Thresholds** (1.0 s and 2.0 s),
- **Asymmetric Quantiles** (20–40–40 split).

All models were trained using identical hyperparameters and the same training protocol, ensuring that performance differences arise solely from the class-definition strategies.

B.2 Performance Comparison

Table 9 summarizes the classification performance across the different discretization schemes. As expected, the binary Median Split yielded the highest accuracy due to the reduced complexity of the two-class problem. However, across all multi-class variants, the macro-averaged F1 scores remained highly stable (approximately 0.89–0.90).

Table 9: Sensitivity analysis of TOT discretization strategies.

Method	Classes	TOT Acc	TOT F1
Global Quartiles (Ours)	**3**	**0.91**	**0.90**
Median Split	2	0.93	0.93
Absolute (1.0 s, 2.0 s)	3	0.90	0.89
Quantile 20–40–40	3	0.89	0.89

Notably, the performance consistency across both data-driven (quantiles) and heuristic (absolute thresholds) labeling schemes indicates that the model's predictive power stems from robust feature representation rather than artifacts of specific class boundaries. The proposed framework scales effectively from binary to granular multi-class tasks without significant performance degradation.

C Emotion Labeling Method Validation

C.1 Experimental Protocol

To empirically validate the retrospective CRD labeling approach, we conducted an auxiliary experiment comparing CRD-based emotion ratings against ground-truth in-situ reports. The protocol used the same simulator settings, synchronization procedures, and takeover scenarios as the main study.

Participants ($N = 5$) completed the driving tasks with a modified procedure: immediately after each event, the simulation was paused, and participants reported their valence and arousal for three temporal phases (before, during, and after). After completing all events, participants viewed the recorded playback and retrospectively labeled their emotions using the CRD interface. This design enabled a direct, event-aligned comparison between immediate and retrospective emotion reports, consistent with prior CRD validation practices.

C.2 Results Analysis

Table 10 summarizes the agreement between the two labeling methods. We evaluated consistency using: (1) three-level categorical agreement (Low: 1–3, Mid: 4–6, High: 7–9), (2) MAE on the continuous 1–9 scale, (3) discrepancy magnitude (1-step, 2-step, and non-adjacent errors), and (4) participant-level Pearson correlation coefficients (r).

Table 10: Agreement statistics between in-situ and CRD emotion ratings.

Metric	Outcome
3-Class Agreement Accuracy	84.2%
Mean Absolute Error (MAE)	0.8 (Scale: 1–9)
Mean Pearson Correlation (r)	0.84
Median Pearson Correlation (r)	0.83
Correlation Range ($r_{\min}$–$r_{\max}$)	0.79 – 0.93
p-value (all participants)	< 0.001
Discrepancy Characteristics	
Valence : Arousal Mismatch Ratio	1 : 4
1-step Deviation	84% of errors
2-step Deviation	15% of errors
3-step Deviation	$< 1\%$
Non-adjacent Mismatches	0%

The correlation values reflect data from four representative participants for whom full paired ratings were available. Individual correlations were $r = 0.79$, 0.82, 0.84, 0.83, and 0.90 (all $p < 0.001$), indicating strong trial-level correspondence between in-situ and CRD emotional trajectories.

C.3 Interpretation

The results indicate strong reliability of the retrospective labeling process. The 3-class agreement of 84.2% shows that CRD labels preserved the overall ordinal affective structure. The MAE of 0.8 further suggests that CRD ratings deviated by less than one scale point on average.

Participant-level correlations were high ($r = 0.79$–0.93), exceeding the typical ranges reported in prior CRD validation studies (often $r \approx 0.5$–0.7) [11]. Crucially, 84% of all discrepancies were only one-step deviations (e.g., Low → Mid or Mid → High), and no non-adjacent misclassifications were observed. This indicates that CRD did not introduce qualitative distortions but only minor boundary shifts. The higher mismatch rate in arousal (4:1 relative to valence) follows known patterns of increased perceptual variability in arousal reporting.

C.4 Summary and Limitations

Overall, the validation results indicate that CRD is a robust and valid proxy for in-situ emotion reporting in our high-arousal driving context. The method preserves both the temporal ordering and ordinal dynamics of emotional responses during takeover events, while avoiding the immersion-breaking interruptions inherent to real-time reporting.

While this paired in-situ vs. CRD comparison should be interpreted as a pilot-level validation—given the modest sample size and limited population diversity—the strong correlations, high categorical agreement, and absence of non-adjacent errors collectively support the reliability of the CRD labeling procedure used in the main study. We therefore view these results as initial but rigorous evidence that retrospective CRD can provide sufficiently accurate labels for model training and analysis, with broader validation across more diverse participants and settings as an important next step.