https://doi.org/10.26599/JICV.2026.9210086

Research Article

Dangerous driving behavior detection method in vehicle intelligent cockpit based on transwindow multimodal fusion

Minghui Zhao[1,2], Bingtao Ren[1,✉], Shichun Yang[1,✉]

[1] School of Transportation Science and Engineering, Beihang University, Beijing 102206, China

[2] China Nuclear Control System Engineering Co., Ltd., Beijing 102401, China

✉ Corresponding author.

E-mail address: B. Ren, renbt1706@buaa.edu.cn; S. Yang, yangshichun@buaa.edu.cn

Received: October 25, 2025; Revised: January 27, 2026; Accepted: May 4, 2026

Abstract： Driver behavior is a critical determinant of road safety; thus, real-time monitoring of dangerous behaviors, such as fatigue and distraction, is essential for intelligent cockpit systems. However, existing in-vehicle sensing technologies are highly susceptible to environmental variations, particularly illumination changes, which result in unstable driver feature extraction. Moreover, Transformer-based models typically incur high computational overhead due to their global self-attention mechanisms, making it challenging to achieve an optimal balance between efficiency and accuracy in Driver Monitoring Systems (DMS). To address these challenges, a visual-tactile fusion framework for dangerous driving behavior detection is proposed. The framework integrates visual sensors with seat-embedded pressure sensors, enabling complementary perception across modalities. Specifically, visual cues facilitate accurate recognition of facial expressions and postures, while tactile signals remain robust under adverse conditions, such as low illumination and occlusion. This complementary fusion significantly enhances system robustness and reliability. Experimental results on a hybrid dataset collected from both simulated and real-world driving scenarios demonstrate the effectiveness of the proposed approach. Compared with state-of-the-art Transformer-based methods and their variants, the proposed TransWindow model achieves an accuracy of 98.77% while maintaining substantially lower computational complexity. These results indicate that the proposed method effectively balances accuracy and efficiency, satisfies the real-time performance of DMS, and provides a promising direction for the continuous optimization of intelligent cockpit systems.

Keywords : Intelligent Cockpit, Driver Monitoring System, Dangerous Driving Behavior Detection, Visual and Tactile Joint Perception, Transformer Multimodal Fusion

1 Introduction

Driver states and behaviors play a crucial role during the driving process and directly affect overall traffic safety. To reduce the risk of traffic accidents, Advanced Driver Assistance Systems (ADAS) have been widely deployed to monitor and analyze driver states and behaviors (Xu et al., 2019). According to the National Highway Traffic Safety Administration (NHTSA, 2024), a total of 40,990 people died in traffic accidents in the United States in 2023, with nearly 23% of the fatalities attributed to abnormal driver states. These statistics highlight the urgent need to develop reliable driver monitoring technologies to improve road safety.

With the emergence of Level 3 conditional autonomous driving technology, intelligent cockpits have become the core platform for in-vehicle human–machine interaction. To satisfy user requirements, intelligent cockpits generally follow a sequential decision-making and execution process, including interaction mode selection (passive, authorized, or proactive), interaction method selection (single-modal or multimodal), and task execution (Chen et al., 2024). Within this framework, Driver Monitoring Systems (DMS) are widely employed to continuously capture drivers' facial expressions, gaze behaviors, body postures, and physiological states. By dynamically evaluating drivers' attention levels and takeover readiness, DMS can provide more intelligent warnings and assistive decision-making. The development of Large Language Models (LLMs) and knowledge graph technologies has opened new directions for intelligent driving monitoring and decision-making. (Liu et al., 2025) enhanced state analysis and anomaly recognition in complex scenarios by integrating LLMs with domain-specific knowledge graphs. (Kuang et al., 2024) utilized multimodal LLMs to construct traffic knowledge graphs and perform intelligent decision-making, demonstrating the significant potential of LLMs in driving behavior understanding and intelligent cockpit cognition.

Despite these advances, existing DMS technologies still face numerous challenges. Dangerous driving behaviors are often subtle, diverse, and highly context-dependent, such as fatigue, distraction, emotional fluctuations, and risky operations, making them difficult to comprehensively capture using single-modal approaches. In addition, various real-world factors, including complex driving environments, lighting variations, and individual differences, can adversely affect detection accuracy and real-time performance. Therefore, enhancing dangerous driving behavior detection through multimodal fusion and intelligent recognition models is essential for improving the robustness and reliability of the system, ultimately ensuring safe human–machine collaboration in Level 3 and higher autonomous driving systems.

Current driver state detection methods can be broadly categorized into invasive and non-invasive approaches. Invasive detection typically relies on physiological signals,

such as electroencephalogram (EEG) and heart rate variability (HRV). Among these, EEG is widely regarded as a benchmark modality for fatigue detection (Sikander and Anwar, 2019), as it analyzes fluctuations across specific frequency bands (θ, α, β, γ waves) to infer drivers' emotional states, fatigue levels, and distraction tendencies (Chu et al., 2019; Xu et al., 2019). However, invasive approaches often suffer from poor user acceptance and limited practicality in real-world driving scenarios. In contrast, non-invasive detection methods monitor driver states and behaviors either directly or indirectly through multiple sensing modalities and have been extensively studied within Driver Monitoring Systems (DMS). Vision-based methods utilize in-cabin cameras to capture facial cues, such as eye closure and mouth movements, for fatigue and emotion assessment, while body posture, head orientation, and gestures provide additional behavioral indicators. Radar-based approaches, particularly those based on millimeter-wave (mmWave) frequency-modulated continuous wave (FMCW) radar, employ signal processing techniques such as short-time Fourier transform (STFT) and wavelet analysis to extract micro-Doppler signatures, enabling the detection of subtle driver motions associated with distraction (Liu et al., 2018). Furthermore, vehicle dynamics analysis incorporates parameters including vehicle speed, steering angle, acceleration, and lane deviation to indirectly infer driving behaviors and identify risky driving patterns (Li et al., 2017). Although these sensing modalities provide complementary perspectives for driver monitoring, each suffers from limitations in robustness and generalization when deployed under complex real-world driving conditions, underscoring the necessity of multimodal fusion strategies.

Traditional computer vision–based methods for driver fatigue and behavior detection primarily rely on the Percentage of Eyelid Closure over the Pupil Time (PERCLOS) algorithm. These approaches typically employ facial landmark extraction tools, such as Dlib integrated with OpenCV (Rajpathak et al., 2009) or more advanced libraries like MediaPipe (Zulkarnanie et al., 2022), to detect key facial regions including the eyes and mouth, and estimate fatigue levels by calculating eyelid closure duration. Subsequent improvements incorporate eye aspect ratio (EAR)–based algorithms (Chen et al., 2024), adaptive eyelid closure thresholds, and Euler angle–based feature correction methods (Fu et al., 2016) to enhance robustness. Additionally, object detection frameworks such as YOLO have been introduced to localize facial regions under complex driving conditions (Li et al., 2022). Although these methods are computationally efficient, they heavily depend on high-quality facial imagery and are therefore sensitive to illumination changes, occlusion, and head pose variations, resulting in limited representational capacity. With the development of deep learning, convolutional neural networks (CNNs) (Xing et al., 2019), Vision Transformers (ViT) (Han et al., 2023), and their variants have been widely adopted to extract richer representations from facial appearance, body posture, and even physiological signals for driver behavior classification. Representative studies include dual-stream CNN architectures that fuse spatial and

temporal information for night-time driver monitoring (Ma et al., 2017), as well as lightweight models such as MobileNet designed to achieve real-time and efficient behavior recognition (Kim et al., 2019; Reddy et al., 2017; Wang, 2020). To better capture both local details and global spatiotemporal characteristics, multiscale network designs have been proposed, including bidirectional feature pyramid networks (BiFPN) (Li et al., 2022), multi-receptive-field feature extraction strategies (Hu et al., 2019), and Fast R-CNN–based frameworks for detecting face–hand interactions (Le et al., 2016). Furthermore, attention mechanisms have been integrated into these models to enhance discriminative feature learning, such as combinations of Swin Transformer and SENet (Hu et al., 2020; Liu et al., 2021; Song et al., 2024), as well as multiscale efficient channel attention modules (Mei, 2025). Beyond unimodal approaches, multimodal fusion strategies have been explored to integrate complementary information from heterogeneous sensors. Examples include the fusion of facial expressions with physiological signals (Oh et al., 2021), as well as non-invasive multimodal data combining ocular features, vehicle dynamics, and environmental information (Mou et al., 2023). These deep learning-based multimodal methods significantly improve detection accuracy and robustness. However, models based on three-dimensional convolutional neural networks (3D CNNs) (Wang, 2023), while capable of jointly modeling spatiotemporal features, incur extremely high computational and memory costs due to 3D convolution operations, far exceeding those of 2D CNNs. Similarly, the Video Vision Transformer (ViViT) (Arnab et al., 2021) employs self-attention mechanisms to capture long-range spatiotemporal dependencies, but its computational complexity increases quadratically with both the number of input frames and video resolution. Consequently, both 3D CNN-based and ViViT-based approaches-whether unimodal or multimodal-suffer from excessive computational overhead and memory consumption, which severely limits their scalability and practical deployment in real-world in-vehicle environments for driver behavior detection.

Although substantial progress has been made in driver behavior detection methods and modeling approaches, existing technologies still face significant challenges. First, mainstream non-invasive detection methods exhibit inherent limitations. Vision-based approaches are highly sensitive to illumination variations and occlusion; millimeter-wave (mmWave) radar suffers from limited spatial resolution, making it difficult to accurately capture subtle driver movements and reliably distinguish multiple targets; and vehicle operation data, such as steering or speed signals, cannot directly reflect specific driver behaviors and are strongly influenced by individual driving styles. Second, video data acquired by visual sensors contain rich spatiotemporal information, which poses considerable challenges for efficient modeling. Processing long video sequences or high-resolution inputs remains difficult, resulting in poor scalability for practical applications. In particular, the high computational cost and memory overhead associated with 3D convolutional neural networks (3D CNNs) and Video Vision Transformers (ViViT)

severely constrain their deployment in real-world in-vehicle environments for driver video behavior detection.

To overcome the above limitations, this study proposes a novel multimodal fusion framework for dangerous driving behavior detection. By jointly integrating visual information with pressure-sensing signals, the framework enables complementary representations of driver behavior. Unlike conventional vision-based approaches that are highly sensitive to illumination variations and occlusions, the pressure-sensing modality captures the distribution and temporal dynamics of the driver's body pressure, providing stable and interpretable biomechanical cues that effectively compensate for the shortcomings of visual sensors. On this basis, a dynamic weighting mechanism is introduced to adaptively regulate the contribution of each modality: when one modality is degraded or unreliable, the model automatically increases the weight of the complementary modality, thereby significantly enhancing recognition accuracy and environmental robustness. In addition, a lightweight TransWindow network architecture is developed. The proposed model adopts window-based attention in conjunction with feature compression operations, which substantially reduce computational overhead compared with state-of-the-art methods, including hybrid Transformer–CNN architectures and multiscale Transformer models. Compared with the windowing strategy of the Swin Transformer, TransWindow further incorporates feature compression and efficient cross-window interaction mechanisms, enabling effective global context modeling while achieving lower computational complexity. Overall, the proposed approach achieves a favorable balance among recognition accuracy, environmental robustness, and computational efficiency, making it particularly suitable for real-time dangerous driving behavior detection in resource-constrained in-vehicle environments.

2 Design of the Dangerous Driving Behavior Detection Method

2.1 Visual and tactile joint detection method

The fusion of pressure-sensing and video modalities effectively overcomes the inherent limitations of single-modality approaches. Pressure sensors capture the distribution and dynamic variations of the driver's body pressure and pedal forces, enabling effective monitoring of sitting posture patterns and body twisting behaviors. In contrast, video data complement the pressure modality by providing posture-based visual analysis (e.g., torso inclination), which compensates for the limited capability of pressure signals in recognizing fine-grained hand movements. Through this complementary integration, a deep fusion of biomechanical cues and visual behavioral

information is achieved, resulting in a more comprehensive representation of driver behaviors.

2.1.1 Piezoresistive Pressure Array Sensor

The pressure signals acquired by piezoresistive pressure array sensors are closely associated with the biomechanical interactions between the driver's body and vehicle interfaces. Variations in contact pressure primarily arise from changes in driving posture, body weight distribution, and force transmission through the lower limbs during vehicle operation. Driving posture determines the spatial distribution of contact forces between the driver and the seat or pedals. In particular, pressure distributions over the pelvis and thighs reflect the driver's center-of-pressure (CoP) pattern, which is indicative of body weight allocation, while asymmetric pressure distributions suggest lateral leaning or habitual postural bias. Consequently, spatial pressure maps provide an intuitive and physically interpretable representation of the driver's sitting posture. Beyond static posture characteristics, posture adjustments and body movements induce dynamic variations in pressure signals. Minor torso and lower-limb movements generate transient pressure changes and low-frequency oscillations, which correspond to behaviors such as body twisting, restlessness, and postural instability. These pressure dynamics are consistent with established biomechanical responses of the human body under conditions of discomfort, fatigue, or increased cognitive load. From a biomechanical perspective, the driver-seat-pedal system can be modeled as a coupled human-machine interaction system. Variations in pedal operation forces are transmitted along the lower-limb kinematic chain, leading to corresponding changes in contact pressure between the thighs and the seat. Stable driving behavior is typically characterized by smooth and coordinated pressure patterns, whereas fatigue or impaired motor control results in increased variability and irregular fluctuations in both pedal and seat pressure signals. This coupling mechanism provides a theoretical foundation for extracting driver-state-related features from pressure measurements. The effectiveness of pressure-based features lies in their strong interpretability and robustness. Statistical and temporal characteristics of pressure signals-such as mean pressure distribution, variance, center-of-pressure trajectories, and low-frequency spectral components-can be directly linked to specific posture adjustments and driving behaviors. Unlike purely data-driven representations, these features correspond to observable physical phenomena and human movement patterns, thereby enhancing model interpretability. Moreover, pressure sensing is non-intrusive and insensitive to illumination variations or visual occlusion, making it well suited for long-term and real-world driving scenarios. When integrated with other sensing modalities, pressure signals provide complementary information that enhances the reliability and completeness of driver state assessment.

2.1.2 Visual and Tactile joint detection architecture design

The proposed architecture employs a side-view camera to acquire visual signals of the driver. This viewpoint allows simultaneous observation of the driver's head, torso, and

hand movements within a single field of view, thereby enabling a more comprehensive characterization of driving behavior. Such a configuration is particularly well suited for recognizing distraction-related behaviors (e.g., mobile phone usage and eating) as well as fatigue-related behaviors (e.g., nodding and torso sway). Moreover, the use of a single camera simplifies system deployment and enhances the feasibility of real-world in-vehicle applications. By comparison, camera configurations commonly adopted in existing public datasets exhibit notable limitations. Cameras mounted near the rearview mirror are primarily designed to capture the forward road scene and offer limited visibility of the driver's torso and upper-limb movements. Cameras installed near the steering wheel focus mainly on facial and head cues, which are effective for fatigue analysis but are highly sensitive to illumination variations and occlusions, leading to degraded robustness in practical driving environments. Top-mounted cameras positioned above the driver's seat emphasize hand-steering wheel interactions but lack sufficient coverage of overall upper-body dynamics, thereby constraining their ability to capture posture-related behaviors. Although visual sensing can provide valuable posture-related cues, such as torso inclination and neck orientation, it remains susceptible to illumination interference and incomplete capture of fine-grained details. To mitigate these limitations, the proposed architecture integrates piezoresistive pressure sensing, which measures variations in the driver's body center of gravity and posture-induced biomechanical responses. These pressure-derived signals provide stable and interpretable information that effectively complements visual cues, particularly under challenging lighting conditions or partial visual occlusion. By jointly modeling visual posture features and pressure-derived biomechanical information, the proposed visual and tactile joint detection architecture achieves a more comprehensive, robust, and reliable representation of dangerous driving behaviors, making it well aligned with the requirements of real-world intelligent driving monitoring systems. Figure 1 illustrates the overall visual–tactile joint detection architecture.

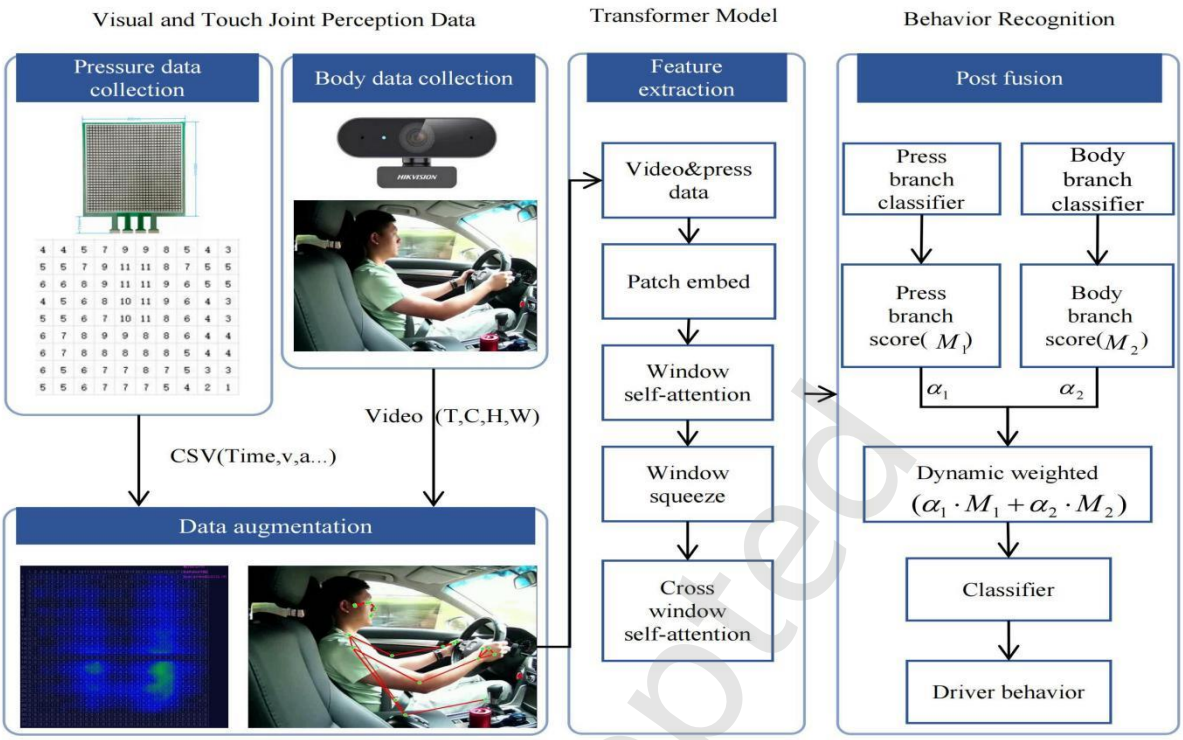


Fig.1 Visual and tactile joint detection model architecture

The first branch is dedicated to driver body pressure acquisition. A piezoresistive pressure array sensor with a resolution of 32×32 channels is embedded within the seat cushion to continuously capture the spatial distribution and temporal variation of contact pressure across different regions of the driver's body. A pressure-to-electrical signal conversion module transforms the raw physical pressure values into digital signals, enabling precise measurement of pressure changes at the driver-seat interface. The resulting numerical pressure matrix is subsequently mapped into a heatmap representation, thereby explicitly encoding spatial structure and facilitating downstream spatiotemporal feature learning. The second branch focuses on driver posture acquisition via visual sensing. A camera is used to capture key behavioral regions, including facial expressions, head posture, and upper-limb movements. Based on the MediaPipe framework, 3D skeletal keypoint detection is performed to extract body posture information. To enhance the expressiveness of the posture representation, additional keypoints are incorporated, and the detected skeletal landmarks are overlaid onto the original video frames, yielding enriched spatial-semantic descriptions of body motion and configuration. Subsequently, a Transformer-based feature extraction architecture is applied independently to the pressure and visual streams. This design enables effective modeling of spatiotemporal characteristics in pressure distribution patterns, subtle posture adjustments, and the spatial relationships and angular configurations among different body parts (e.g., head, torso, and limbs). Each modality is then processed by an independent classifier to produce modality-specific confidence scores.

To effectively integrate information from visual and tactile modalities, we introduce a learnable weighting

mechanism. Specifically, a set of trainable parameters is employed to model the relative importance of each modality, which are subsequently normalized via a softmax function to obtain modality-wise weights that sum to one. Given the modality feature set ($\{M_i\}_{i=1}^{N}$) Eq. (1), the fused representation is computed as a weighted summation of all modalities. This formulation enables the model to emphasize more reliable modality features while suppressing less informative ones. By adaptively adjusting the contribution of each modality, the proposed fusion mechanism enhances robustness against environmental variations and improves the accuracy of dangerous driving behavior recognition.

$$F = \sum_{i=1}^{N} \alpha_i \cdot M_i, \quad where \quad \sum_{i=1}^{N} \alpha_i = 1$$
$$\alpha_i = \frac{\exp(w_i)}{\sum_{j=1}^{N} \exp(w_j)} \tag{1}$$

M_i stands for The i-th modal feature, w_i is learnable parameters, α_i is normalized weights and F is Features after fusion.

2.1.3 Synthetic data collection based on driving simulator and real vehicle testing

Data acquisition using driving simulators offers several advantages, including the elimination of safety risks and precise control over experimental conditions such as weather, traffic density, and road scenarios, thereby enabling systematic and repeatable testing. In contrast, real-world driving data collected from actual vehicle operations capture rare stochastic events and complex driving maneuvers that are difficult to reproduce in simulated environments, providing higher ecological validity and representativeness. Motivated by these complementary strengths, this study employs a hybrid data collection strategy that integrates both driving simulator experiments and real-vehicle driving tests. Data are collected across seven categories of driving behavior: angry driving, depressed driving, fatigued driving, mobile phone usage while driving, conversational driving, eating while driving, and normal driving. For multimodal data acquisition, the system is equipped with a Hikvision USB camera for visual sensing and a Roxi Technology RX-M3232L piezoresistive pressure array sensor for tactile sensing, as illustrated in Figure 2.

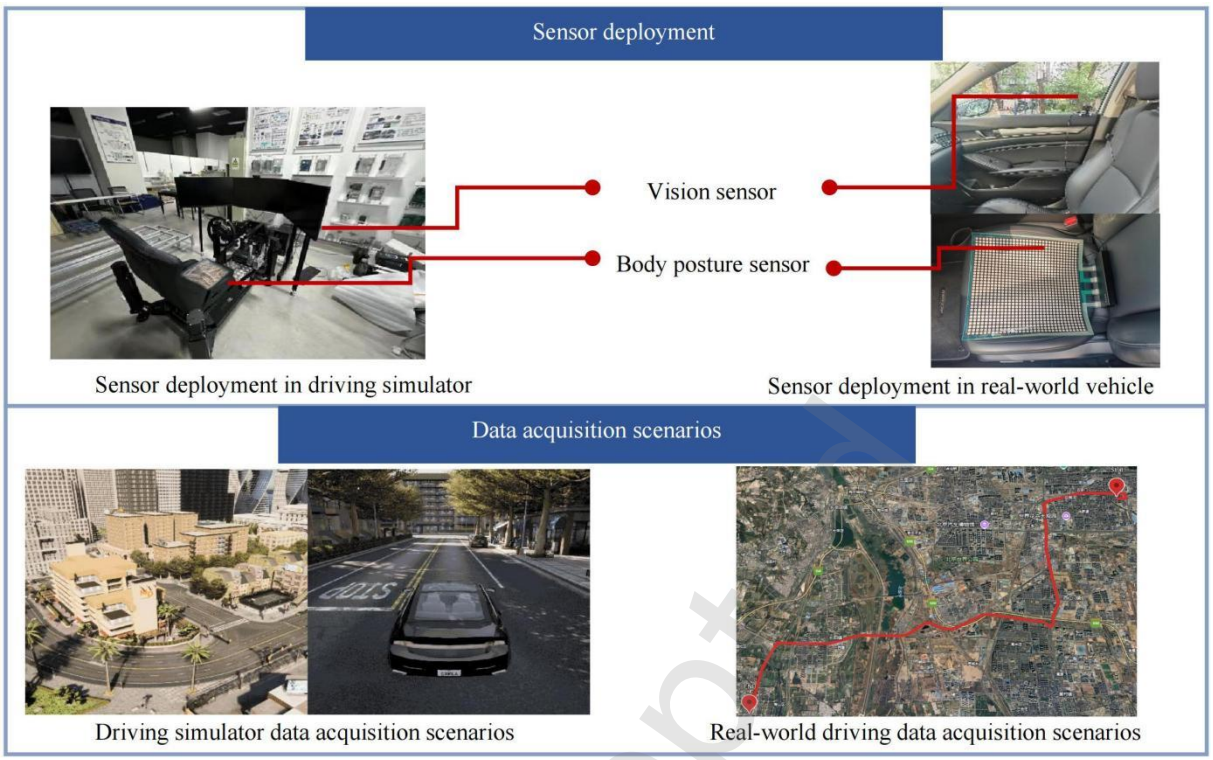


Fig.2 Sensor deployment

The visual camera was mounted on the right side of the driver, while the piezoresistive pressure array sensor was installed beneath the driver's seat cushion. To ensure strict temporal alignment between the visual and tactile modalities, precise sensor calibration and software-based synchronization were performed during system deployment. For simulated driving data acquisition, dangerous and normal driving scenarios were constructed within the CARLA simulation environment, using a four-degrees-of-freedom driving simulator. The simulation route was configured as a daytime urban road network with one to two lanes in each direction and a speed limit of 40 km/h. For real-world data collection, daytime driving experiments were conducted on Jingkai Expressway, Jingliang Road, and multiple urban streets, covering a total driving distance of approximately 30 km. These routes featured a posted speed limit of 80 km/h and encompassed diverse traffic conditions and typical driving behaviors.

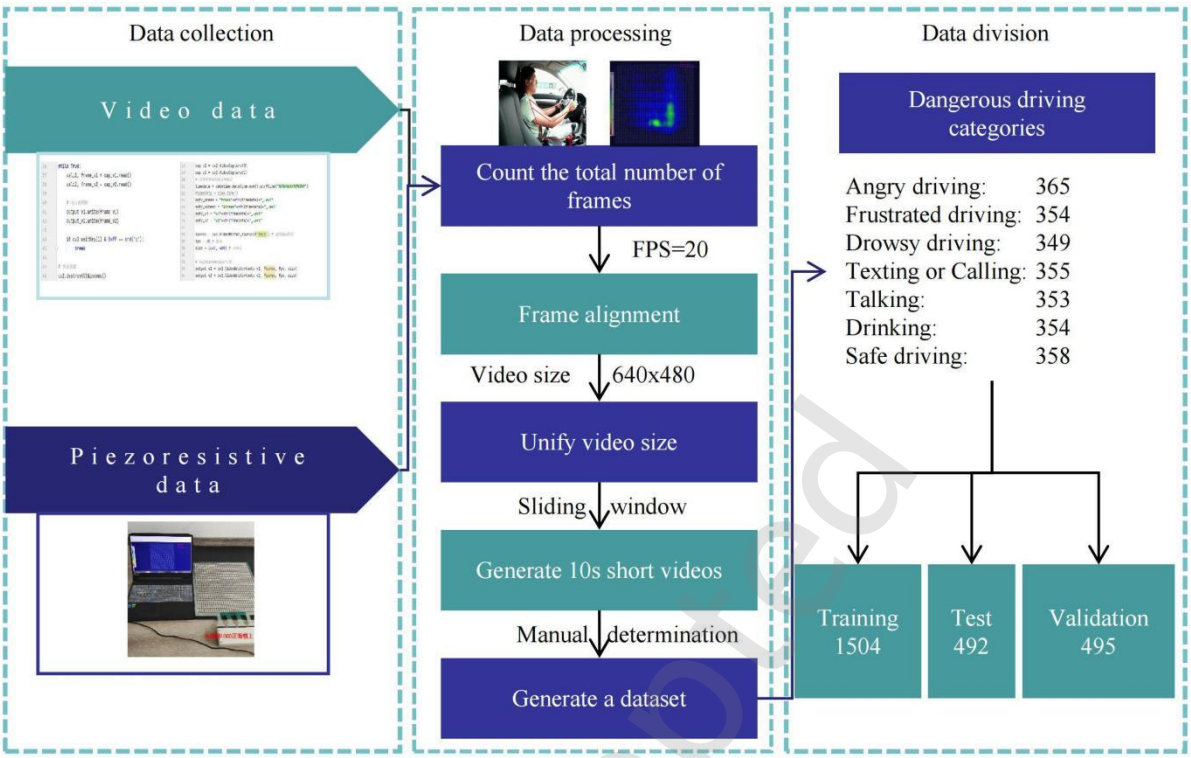


Fig.3 Data processing flow

The dataset comprises seven categories of driving behavior, defined as follows:

(1) Angry Driving: yelling, aggressive maneuvers, tailgating, and prolonged honking;

(2) Frustrated Driving: low emotional state, inattention, delayed reactions, and distracted behavior;

(3) Drowsy Driving: frequent yawning, slow response times, and visible signs of fatigue;

(4) Texting or Calling: operating a mobile phone with one hand, making or receiving calls, or sending voice messages while driving;

(5) Talking: turning to converse with passengers in the front or rear seats;

(6) Drinking While Driving: opening a water bottle and drinking while steering;

(7) Safe Driving: normal, attentive, and focused driving behavior.

Video data were recorded at 30 frames per second (fps), while pressure signals were sampled at 1000 Hz. To achieve multimodal synchronization, all sensor streams were aligned to a unified timeline through software calibration. Video sequences were temporally trimmed by removing redundant frames using a 10-frame sliding window, and the pressure data were downsampled to 20 Hz. Additionally, video clips exhibiting excessive occlusion or severe shadowing (exceeding 80% of the clip duration), as well as samples with abnormal sensor readings, were filtered out to ensure data quality. Each sample was annotated with four coarse-level labels (Emotional, Fatigued, Distracted, and Normal) and six fine-grained labels (Angry, Frustrated, Drowsy, Texting or Calling, Drinking, and Talking). The multimodal data and corresponding annotations were associated via file paths, then randomly shuffled and partitioned into training, validation, and test

sets with a ratio of 6:2:2. A corresponding JSON metadata file was generated to record label definitions, sample counts per class, and other key dataset statistics, ensuring the reliability and reproducibility of model training and evaluation. The final distribution of the dataset is illustrated in Figure 4.

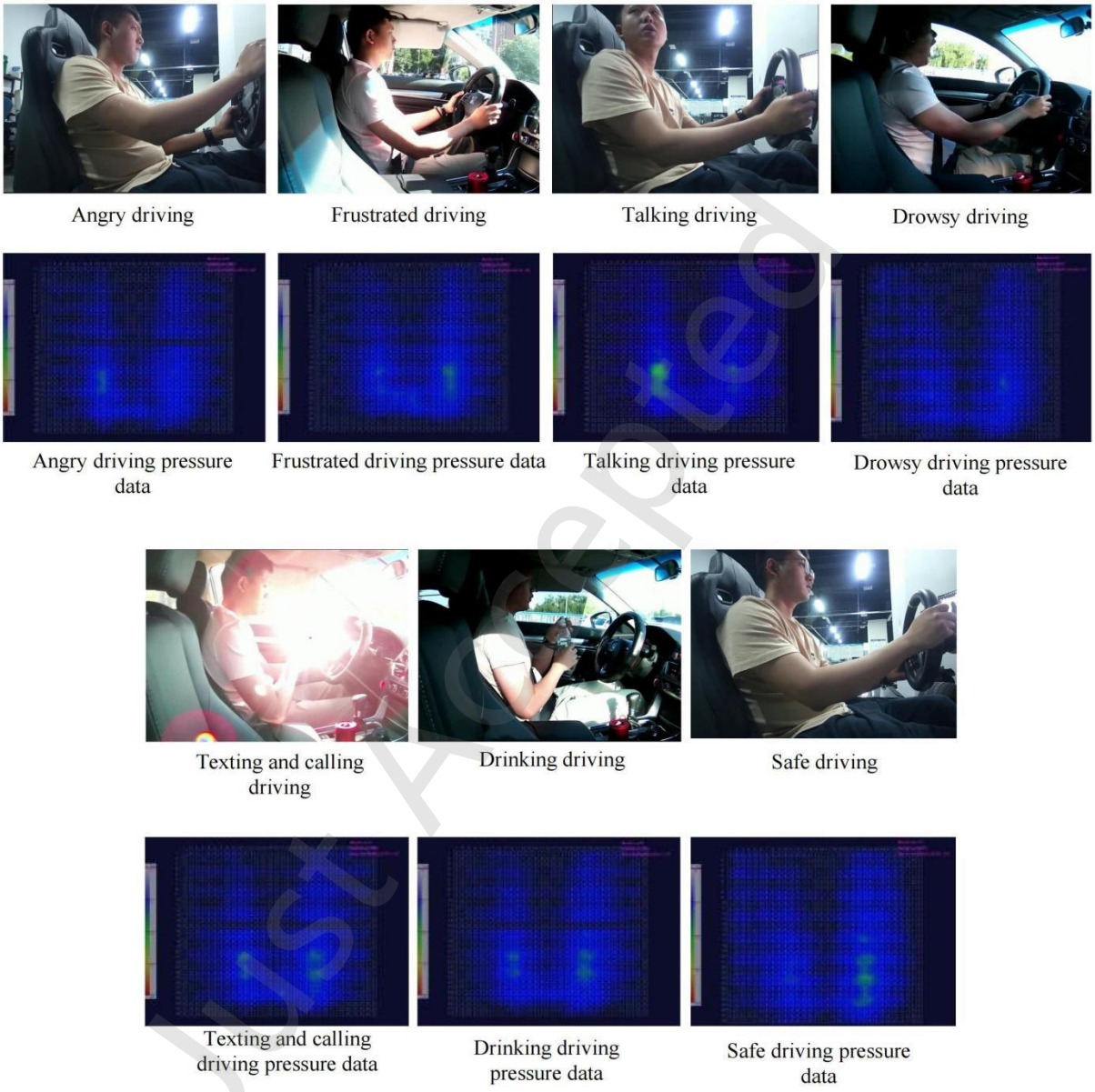


Fig.4 Dangerous driving datasets

2.1.4 Data preprocessing based on MediaPipe

For real-time driving behavior detection, MediaPipe(Lugaresi et al., 2019) is adopted for visual feature extraction due to its high processing speed and low computational overhead, while still enabling accurate estimation of three-dimensional skeletal keypoints of the driver (e.g., hands, elbows, shoulders, and knees). In contrast, although pose estimation frameworks such as OpenPose (Cao et al., 2021) and AlphaPose (Fang et al., 2023) achieve superior accuracy in complex multi-person

scenarios, their reliance on heavyweight network architectures results in substantial computational costs and inference latency, which significantly limits their applicability in real-time driver monitoring systems. The visual processing pipeline begins with video frame preprocessing, including BGR-to-RGB color space conversion, spatial resizing, and optional grayscale transformation to accelerate subsequent feature extraction. The core detection modules then operate in parallel. Specifically, the BlazePalm detector rapidly localizes hand regions, followed by a lightweight convolutional neural network that regresses the coordinates of 21 hand keypoints. Simultaneously, the BlazePose detector estimates the driver's full-body pose, and the Pose Landmark model further refines this output by predicting the positions of 17 essential skeletal landmarks, including the head, neck, shoulders, elbows, hips, and knees. Each landmark is represented by three-dimensional coordinates (X, Y, Z), where the depth component (Z) enables effective recognition of behaviors involving spatial displacement, such as forward leaning, which may indicate fatigue or distraction. To ensure temporal consistency and robustness, a pose constraint module dynamically calibrates landmark positions across adjacent frames, reducing abrupt positional fluctuations and maintaining coherent motion trajectories. In addition, Kalman filtering is applied to predict landmark positions based on a kinematic motion model, effectively suppressing noise caused by occlusion, motion blur, or transient detection failures. The resulting 3D skeletal landmark trajectories are overlaid in real time onto the original video frames for visual inspection and are simultaneously recorded as multivariate time-series data. These trajectories serve as the primary visual representation for downstream behavior analysis. A Transformer-based recognition model is subsequently employed to model the full temporal evolution of driver actions—from initiation to execution and completion—thereby enabling fine-grained discrimination between behaviors such as normal steering, mobile phone operation, and fatigue-related torso leaning. This temporal modeling capability significantly enhances the accuracy and timeliness of dangerous driving behavior detection and early warning generation.

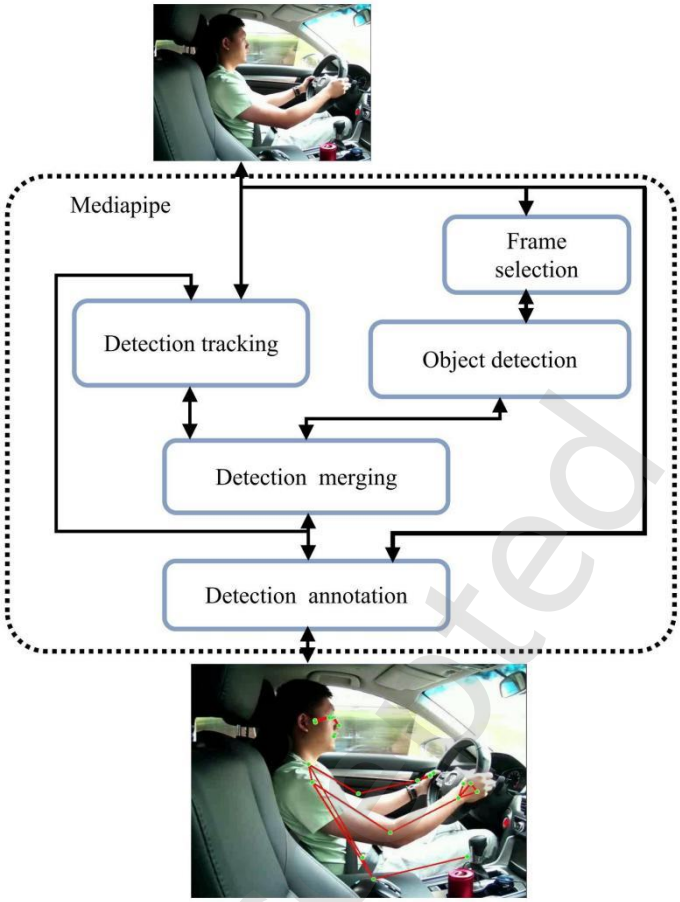


Fig.5 Driver pose adds skeleton key points through Mediapipe

For tactile sensing, the driver posture pressure analysis module acquires real-time seat pressure distribution data using a 32×32 pressure array sensor embedded beneath the driver's seat. To improve the interpretability and analyzability of the raw matrix data, the system maps the pressure values into a two-dimensional pressure heatmap, thereby introducing explicit spatial structure,

$$M'_{[i][j]} = \frac{M_{[i][j]} - M_{\min}}{M_{\max} - M_{\min}} \tag{2}$$

Through color linear interpolation, the normalized matrix M' is mapped to a predefined color map.

$$C(M) = (1 - M') \cdot C_{\min} + M' \cdot C_{\max} \tag{3}$$

$C_{\min} = (R_{\min}, G_{\min}, B_{\min})$ and $C_{\max} = (R_{\max}, G_{\max}, B_{\max})$ are the given two endpoint colors. The pressure intensity of each sensor cell matrix is visually displayed by the color depth (high value dark color, low value light color). Based on the color correlation of adjacent blocks in the thermal map, the system can quantitatively analyze the spatial correlation characteristics of pressure distribution, and then identify the driver's posture change patterns such as sitting posture deviation and center of gravity transfer.

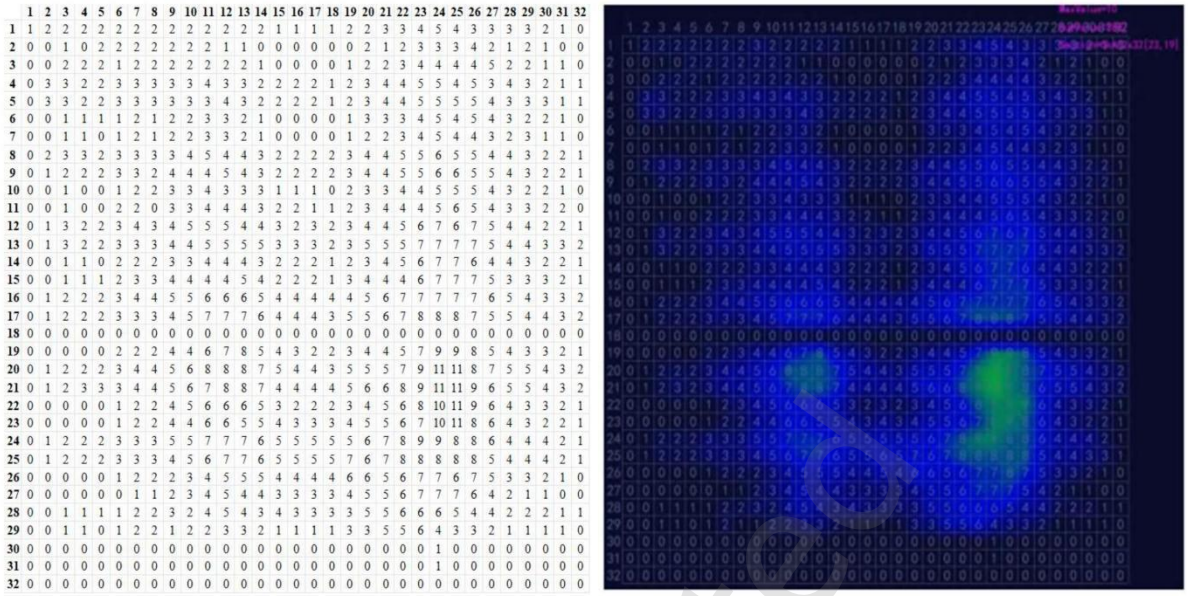


Fig.6 Converting a 2D matrix to a heat map

2.1.5 Late-fusion of visual and tactile features

Feature extraction is performed independently in the visual and tactile branches and subsequently fused at a high semantic level, preserving modality-specific characteristics while compensating for missing or unreliable information in individual modalities. This late-fusion strategy offers superior flexibility and scalability compared with early- or mid-level fusion approaches. Since each modality is processed by a dedicated branch-specific architecture, there is no need to temporally or spatially align raw sensor data at the input stage. When an additional sensing modality is introduced, it can be seamlessly incorporated by adding a new processing branch and integrating its output at the fusion layer, thereby avoiding modifications to existing feature extractors and significantly reducing overall system complexity. Moreover, late fusion is particularly effective in capturing high-level semantic correlations across modalities. Unsafe driving behaviors-such as mobile phone usage or aggressive horn pressing-often involve subtle motion patterns and facial or postural cues that only become semantically meaningful after deep feature extraction. By enabling interactions between modalities at the feature level and leveraging attention-based or adaptive weighting mechanisms, the proposed fusion scheme selectively emphasizes informative cues from each branch. This design enhances the modeling of cross-modal dependencies and substantially improves recognition accuracy and robustness. In addition, late fusion mitigates the impact of data quality imbalance, such as visual occlusion or low-light noise, by reducing the influence of unreliable modality-specific features on the final decision.

2.2 Design of TransWindow model for extracting driver behavior features

The core idea of the TransWindow model is to decompose conventional global self-attention into two complementary components: intra-window self-attention and cross-window global interaction, thereby avoiding the $O(N^2)$ computational explosion of a traditional

Transformer on large inputs while still retaining long-range dependency modeling: first, efficient self-attention is performed within each fixed-size window, reducing overall complexity to approximately $O(N)$; then, a cross-window attention step captures interactions between any two distant regions in one shot, completely overcoming the need for CNN to stack many convolution layers just to propagate global information. Unlike the Swin Transformer, which relies on a four-stage hierarchical structure with repeatedly stacked BasicLayers, TransWindow achieves local-to-global feature aggregation using a single layer composed of window-based attention and cross-window interaction modules. As a result, TransWindow attains substantially lower model complexity and computational cost than conventional multi-layer Transformers or deep CNN architectures. This lightweight design makes it particularly well suited for scenarios with limited training data or constrained computational resources, such as real-time in-vehicle driver monitoring. With appropriately selected window sizes and token abstractions, TransWindow maintains strong representational capacity while delivering performance comparable to significantly larger models in driver behavior recognition and multimodal fusion tasks.

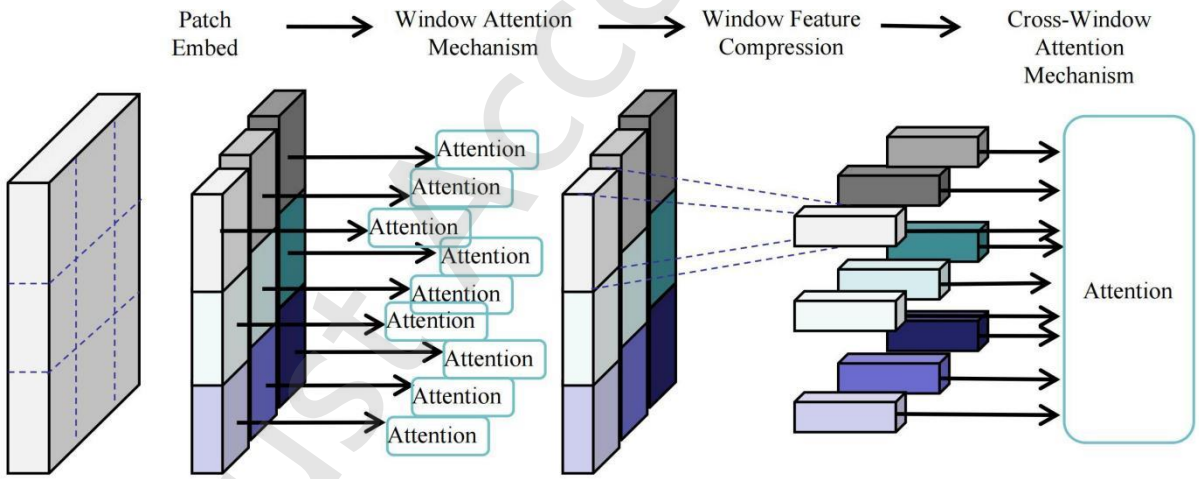


Figure 7 TransWindow model architecture diagram

2.2.1 Patch embed

The patch embedding stage employs a 3D convolutional operation to simultaneously perform feature extraction and downsampling along the temporal and spatial dimensions. By applying convolutional kernels with a stride of (2, 4, 4), the input video sequence is reduced to half its original temporal resolution and one-quarter of its original spatial resolution in both height and width. This aggressive but structured downsampling substantially reduces downstream computational and memory costs while preserving essential local spatiotemporal structures. Compared with a purely linear projection, convolution inherently provides local receptive fields and parameter sharing, enabling the extraction of low-level semantic cues such as texture patterns, motion boundaries, and short-

term temporal variations. The temporal dimension of the 3D convolution further equips the model with an initial capability for dynamic motion modeling, allowing it to capture inter-frame changes such as gesture transitions or posture shifts at an early stage. By ensuring that the resulting spatiotemporal patches are non-overlapping and semantically independent, the patch embedding module establishes a uniform and controllable token grid. This structured representation simplifies subsequent window-based attention computation, facilitates efficient cross-window interactions, and supports stable token-level processing in the TransWindow architecture.

2.2.2 Window attention mechanism

The window based attention mechanism constrains the scope of self-attention by partitioning the input into fixed-size unit windows and performing attention computations independently within each window.

$$Attn(W_{p,q,r}) = Softmax(\frac{QK^T}{\sqrt{d}})V \quad (5)$$
$$Q, K, V \in \square^{(M_T, M_H, M_W)\times d}$$

Where (M_T, M_H, M_W) given window size .

By enforcing this fixed-window design, the computational complexity is reduced from $O(N^2)$ to $O(N)$ approximately, greatly improving efficiency and alleviating the resource bottleneck caused by large input volumes, which allows the model to process large-scale video data more quickly and effectively. Window partitioning not only saves computation but also encourages the model to focus on the key information in local regions, enhancing its ability to capture fine-grained

The input patch corresponding to each output position (i, j, k) is:

$$\begin{aligned} &Patch(i, j, k) = X[2i:2i+1, 4j:4j+3, 4k:4k+3] \\ &i \in [0, T/2-1] \\ &j \in [0, H/4-1] \\ &k \in [0, W/4-1] \end{aligned} \quad (4)$$

Where $X \in \square^{T\times H\times W\times C}$ i represents the input data as a three-dimensional tensor, T, H, and W represent the time dimension, height, and width respectively. height, and width respectively,C is the number of channels.

For each window $W_{p,q,r}$, self-attention is calculated only on the patch token within the window:

temporal and spatial features. The localized attention mechanism enables the model to maintain high computational efficiency while exhibiting strong local pattern recognition, which in turn improves generalization and robustness to detailed behaviors in complex backgrounds. Prior to entering the attention module, windowed features undergo Layer Normalization, stabilizing the feature distribution across batches and layers and facilitating faster convergence. Multiple independent attention heads divide the input features into parallel

subspaces to capture semantic relationships in different parts of the sequence. Finally, residual connections and MLP feed-forward layers are added after each attention module to mitigate vanishing gradients in deep networks and to enable cross layer information fusion.

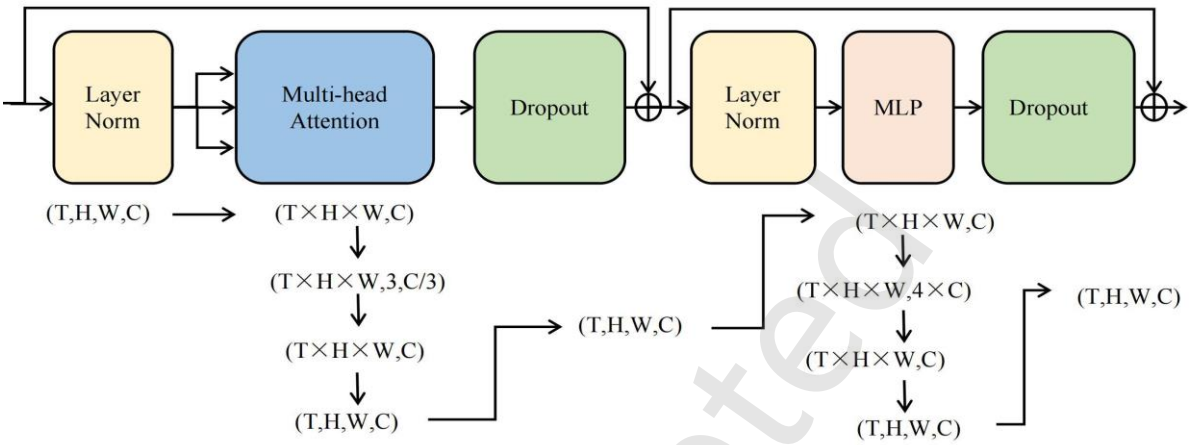


Fig 8 Window multi-head attention mechanism

2.2.3 Window feature compression

Before performing inter-window attention, which involves multi-head attention across multiple windows and incurs a relatively high computational cost, the TransWindow model first applies a window-level feature compression module. This module is designed to preserve globally informative cues while suppressing redundant information and reducing feature dimensionality. Specifically, spatial aggregation is employed to compress each window's feature map, while the channel dimension is correspondingly increased. This design encourages the model to gradually shift its representational focus from low-level visual cues (e.g., edges and textures) to higher-level semantic features, such as object configurations and behavior categories. To further enhance computational efficiency and representational capacity, a depthwise separable convolution structure is incorporated. In this structure, depthwise convolution is first applied independently within each channel to extract spatial features, followed by pointwise convolution to fuse information across channels. This decomposition significantly reduces computational overhead compared with standard convolution, while enabling more fine-grained and higher-order feature representations. To explicitly model inter-channel dependencies, the compressed feature vectors are passed through a channel attention module. The channel descriptors are first projected to a lower-dimensional space via a fully connected layer, followed by a ReLU activation to amplify salient channels and suppress less informative responses. A second fully connected layer then restores the original channel dimensionality. Finally, a Sigmoid gating function is applied to model nonlinear channel-wise interactions, adaptively enhancing informative features and attenuating

irrelevant ones. This mechanism improves both feature selection and computational efficiency prior to inter-window attention fusion.

2.2.4 Cross-Window attention mechanism

Relying solely on intra-window self-attention has inherent limitations. Because attention computation is confined to local windows, the model forms isolated "information islands," preventing effective modeling of long-range dependencies and global contextual relationships. Although window-based attention excels at capturing fine-grained local details, it inevitably overlooks interactions across distant windows, which restricts the model's overall semantic understanding of complex driver behaviors. To overcome this limitation, a cross-window global attention mechanism is introduced to explicitly bridge information gaps between windows. This mechanism enables efficient information exchange across spatially or temporally distant regions, allowing the model to capture global context and long-range dependencies in a single attention operation. As a result, the proposed cross-window attention effectively resolves the information isolation problem inherent in local attention, significantly enhancing holistic scene perception and the model's ability to reason about complex, multi-stage driving behaviors.

3. Experimental verification and analysis

3.1 Model training

To evaluate TransWindow's overall performance during training and inference, we built a collaborative software-hardware experimental platform. The system is based on PyTorch 2.0 and Python 3.8, running on an AutoDL L20 GPU (119.5 TFLOPS mixed-precision tensor performance, 48 GB VRAM) with CUDA 11.8 acceleration and 100 GB of system memory to support large batch training, ensuring efficient training and validation of both the TransWindow model and the integrated visual and tactile joint detection pipeline.

During data augmentation, input video frames are first adjusted to 224×224 to match the network's input requirements. We then apply random region cropping to break the model's reliance on the image center and encourage it to attend to informative regions across the frame, mitigating spatial bias. Horizontal flipping is also introduced so the model learns to recognize symmetric driving actions and scene variations, improving directional robustness. Finally, random brightness adjustments generate virtual training samples under diverse lighting conditions, enhancing the model's illumination invariance for night-time or tunnel scenarios. For optimizer scheduling, we adopt a ReduceLROnPlateau policy to dynamically adjust the learning rate: if validation metrics (e.g., loss or accuracy) do not significantly improve over several epochs, the learning rate is automatically reduced by a factor of 0.1 (new_lr = old_lr×0.1), which helps fine-tune parameters in later training stages and avoid local minima. Additionally, we fix the channel dimension to 192 in the window attention layers to strike a balance between representational power and computational efficiency, enabling strong feature encoding even under resource constraints.

3.2 Evaluation indicators and test condition

To comprehensively evaluate the performance, efficiency, and overall advantages of the proposed TransWindow feature extraction model in driver behavior detection tasks, a series of multi-dimensional comparative experiments were conducted. The comparison benchmarks include the Video Vision Transformer (ViViT) (Arnab et al., 2021) and several representative Transformer-based variants, specifically: hybrid Transformer–CNN architectures (ConViT (d'Ascoli et al., 2021)), sliding-window-based efficient attention models (Swin Transformer (Song et al., 2024)), multiscale attention mechanisms (MViTv2 (Li et al., 2022)), factorized spatiotemporal attention models optimized for efficiency (TimeSformer (Bertasius et al., 2021)), and pyramid attention models emphasizing local feature continuity (PVTv2 (Wang et al., 2022)). These models are selected as baseline methods for comparison. In the classification evaluation, we employed the confusion matrix-a standard tool for performance assessment-using the following metrics: accuracy, precision, recall, and F1-score. The formulas for these metrics are defined as follows:

$$Accracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{8}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{9}$$

TP stands for the number of true positives, that is, the number of correctly classified positive samples; TN stands for the number of true negatives, that is, the number of correctly classified negative samples; FP stands for the number of false positives, that is, the number of negative samples misclassified as positive; and FN stands for the number of false negatives, which means the number of positive samples misclassified as negative.

3.3 Model comparison test and result analysis

We conducted training and comparative experiments separately on driver body posture data, pressure sensor array data, and their multimodal fusion. The training process is shown in Figure 9, and the results are shown in Table 1. The results demonstrate that, in terms of single-modality performance, the pressure array sensor exhibits slightly slower convergence and lower final accuracy than the body posture modality. However, when the two modalities are fused, the resulting model achieves higher accuracy than either unimodal configuration, demonstrating the effectiveness of multimodal fusion. in driver behavior detection.

In addition, the proposed dual-stream architecture incorporates a dynamic weighting mechanism. When one modality becomes unreliable or partially unavailable, the model automatically reduces its corresponding weight while increasing the contribution of the more reliable modality. This adaptive weighting strategy ensures robust detection performance under conditions of partial signal degradation or modality failure. Regarding model selection, the window-based Transformer architecture (TransWindow) significantly

reduces computational complexity through localized attention computation. Compared with representative models of the same class, TransWindow requires fewer computations than ConViT, ViViT, and MViTv2, while its recognition accuracy is only 0.55% lower than that of the best-performing MViTv2. Although TimeSformer and PVTv2 achieve lower computational costs, their detection accuracy is substantially inferior to that of TransWindow. Overall, TransWindow achieves a favorable trade-off between recognition accuracy and real-time performance, making it well suited for practical in-vehicle driver monitoring applications.

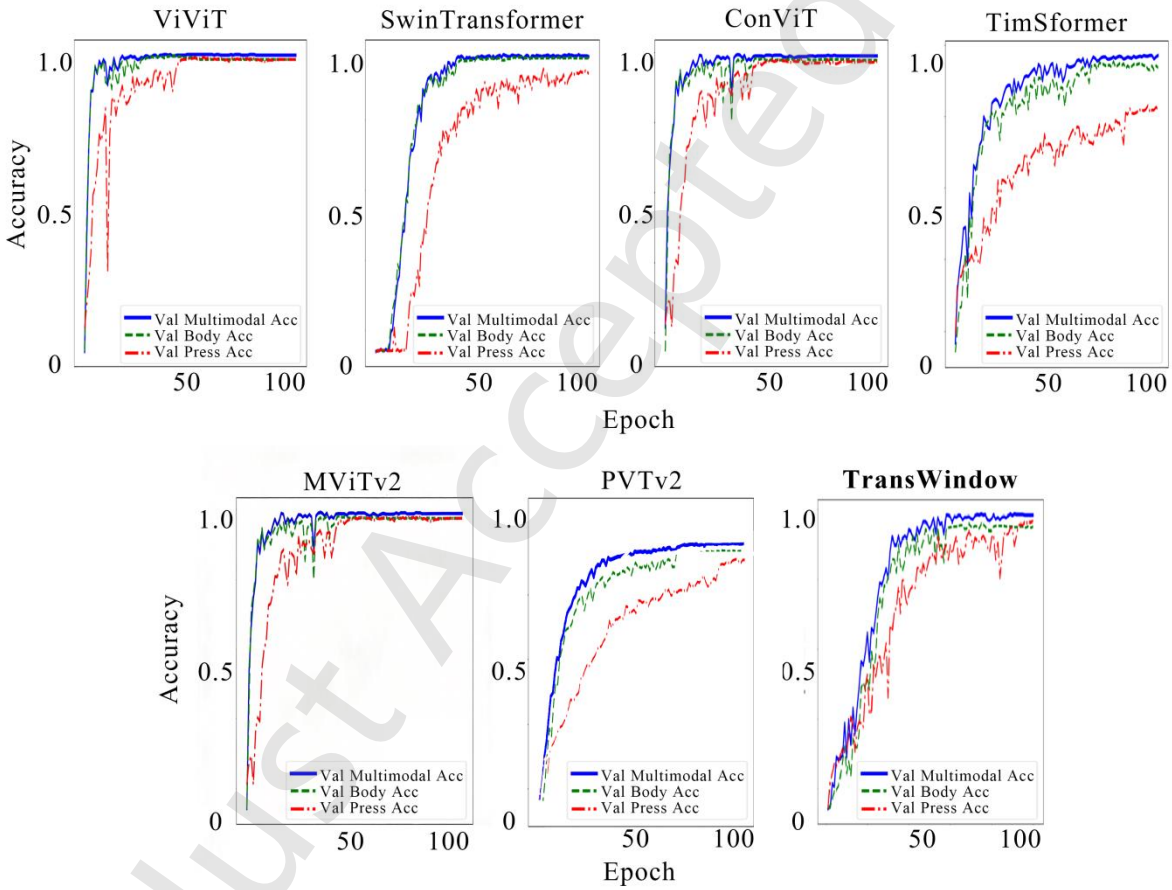


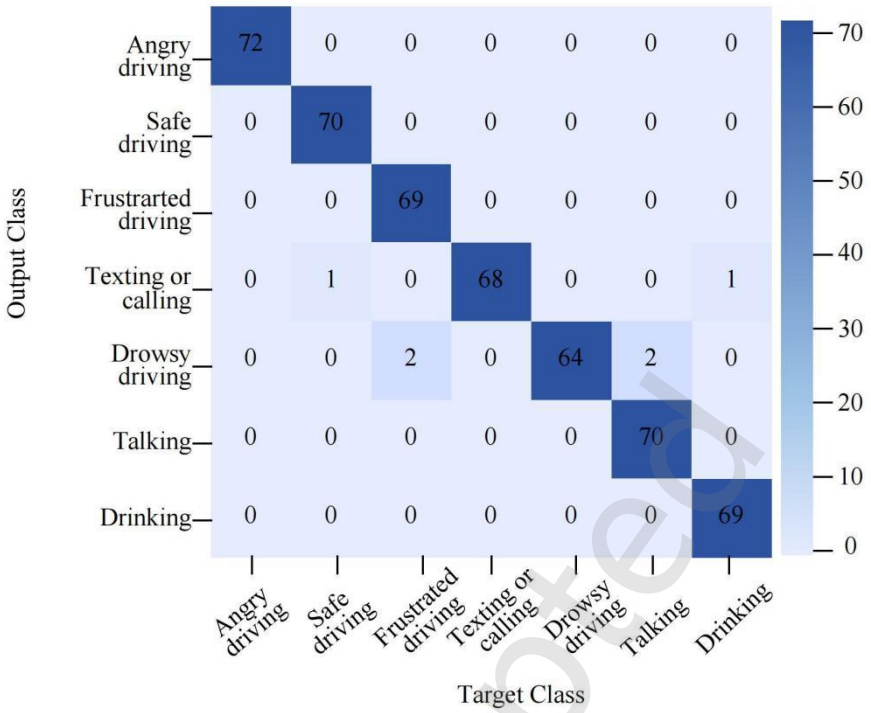
Fig 9 Model training process on the validation set

Table 1 Comparative evaluation of models on the test set

Method	Modal	Acc.(%)	Percision	Recall	F1-score	FLOPs	Params(M)	MACs(M)
ViViT	Body & Press Fusion	98.57	0.9860	0.9857	0.9856	1.465T	172.493	226,959,8.202
	Body	97.34	0.9735	0.9734	0.9733			
	Press	96.72	0.9688	0.9672	0.9673			
Swin Transformer	Body & Press Fusion	97.75	0.9779	0.9775	0.9774	282.87G	55.03	299,545.952

ConViT	body	95.90	0.9594	0.9590	0.9589	1.821T	200.841	227,008,1.139
	Press	91.60	0.9177	0.9160	0.9160			
	Body & Press Fusion	98.16	0.9818	0.9816	0.9816			
TimSformer	body	97.95	0.9798	0.9795	0.9795	8.362G	3.826	627,100,8.768
	Press	96.93	0.9698	0.9693	0.9692			
	Body & Press Fusion	94.31	0.9438	0.9431	0.9431			
MViTv2	body	93.44	0.9346	0.9344	0.9341	1.462T	172.484	152,309,6.789
	Press	90.53	0.9221	0.9053	0.9076			
	Body & Press Fusion	99.32%	0.9933	0.9932	0.9931			
PViTv2	body	98.32%	0.9833	0.9832	0.9831	20.037G	5.442	206,26.763
	Press	95.45%	0.9574	0.9545	0.9537			
	Body & Press Fusion	71.97%	0.7335	0.7197	0.7193			
TransWindow	body	71.21%	0.7251	0.7121	0.7089	175.502G	5.984	187,220.132
	Press	69.70%	0.7072	0.6970	0.6933			
	Body & Press Fusion	98.77%	0.988	0.9877	0.9876			
	body	96.29%	0.9644	0.9629	0.9627			
	Press	95.08%	0.9515	0.9508	0.9506			

The visual–tactile joint detection model demonstrates excellent performance on the driver behavior recognition task, achieving high recognition rates across all seven behaviors (angry, normal, frustrated, mobile-phone use, fatigued, conversational, and eating), as illustrated in Figure 10. The confusion matrix shows that the angry, texting and calling, and drowsy classes each attain 100% recognition accuracy. The frustrated and talking classes suffer two misclassifications, each being erroneously labeled as drowsy. This confusion likely arises because the sighing, head-dropping, and slowed reactions characteristic of frustrated driving-as well as the mouth movements associated with talking driving-closely resemble certain manifestations of drowsy driving. Additionally, one safe driving instance and one drinking driving instance were each misclassified as texting or calling while driving.



.

Fig 10 Test set confusion matrix

To further assess model interpretability, Grad-CAM visualizations (Selvaraju et al., 2017) were employed to analyze feature activations within the self-attention layers of the TransWindow model, as shown in Figure 11. The visualized results, corresponding to two consecutive video frames, indicate that the model's attention focuses on semantically meaningful regions consistent with human cognitive understanding of driving behaviors. Specifically, TransWindow effectively captures temporal posture changes and dynamically tracks critical behavioral cues, such as torso movement, hand detachment from the steering wheel during distracted driving, and head scanning behavior associated with angry driving. These findings support the reliability and interpretability of the proposed model for safety-critical driver monitoring applications.

(a) Driver behavior at time T. The left side is drinking driving, the middle is angry driving, and the right side is texting or calling driving

(b)Driver behavior at time T+1: left is driving while drinking, middle is driving while angry, and right is driving while texting or calling.

Fig 11 Visualize model feature heatmaps using GradCAM

3.4 Model robustness verification analysis

To further evaluate the robustness of the proposed TransWindow model, we constructed a challenging dataset based on raw data collected from real-world driving scenarios, in which noise-corrupted samples were explicitly isolated. The dataset contains 1,035 samples and focuses on complex and adverse driving conditions. Environmental disturbances were deliberately intensified, including low-light scenarios (e.g., underground parking lots and underpasses), strong direct illumination, and adverse weather conditions such as rain. In addition, highly dynamic traffic situations, including rush-hour congestion and dense vehicle interactions, were incorporated to comprehensively reflect drastic variations in illumination and driving environments. furthermore, by artificially introducing row and column-level data anomalies (such as partial row and column signal failure due to poor contact) during the pressure sensor array data acquisition process, the adaptability of the model under sensor performance degradation and complex interference was tested. Further verification was conducted to determine whether the model possesses good generalization ability and robustness in harsh environments with significant noise interference. As shown Figer12 and Table 2.

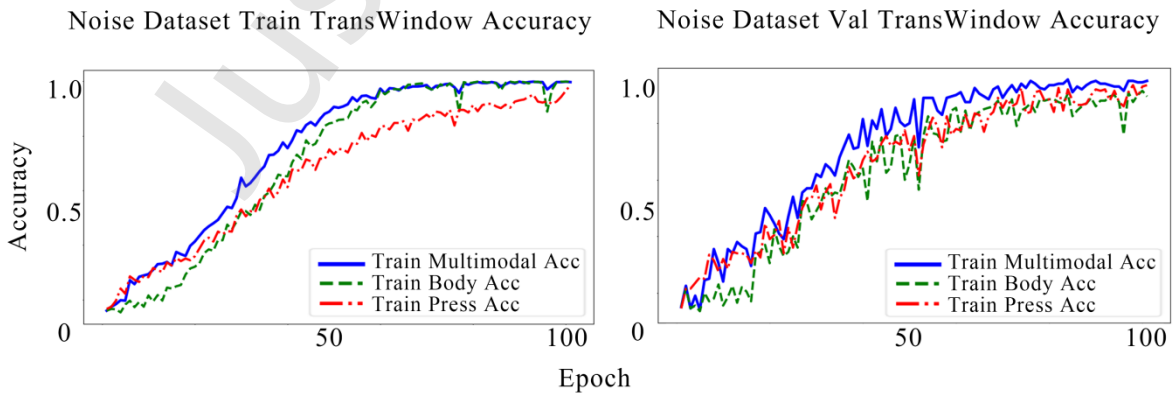


Fig 12 Training process of the TransWindow model on a noisy dataset

Table 2 Noise datasets model evaluation

Method	Acc.(%)	Precision	Recall	F1-score

Multimodal	97.00	0.9708	0.9700	0.9700
Body	89.50	0.9071	0.8950	0.8965
Press	96.00	0.9623	0.9600	0.9602

The model's adaptability and fault tolerance under normal operating conditions further demonstrate its reliability in complex and interference-prone scenarios. When a single modality is severely affected by external disturbances or suffers from information degradation, the model leverages a dynamic weighting mechanism to adaptively down-weight the impaired modality while simultaneously increasing the contribution of the other, more reliable modality, thereby effectively ensuring the stability and accuracy of in-cabin perception. Whether under normal operating conditions or in complex traffic environments with strong noise interference, the proposed model consistently exhibits excellent robustness and reliability, providing strong support for driver behavior monitoring and safety assessment.

3.5 Model ablation experiment verification analysis

To validate the effectiveness of each module in TransWindow, we conducted a comprehensive ablation study and carried out experiments using unimodal body posture data. The proposed model consists of three key modules: Window attention mechanism, Window feature compression, and Cross-window attention mechanism. Each module aims to address different aspects of spatiotemporal representation learning, including local context modeling, channel-level feature recalibration, and global dependency capture.

The first configuration uses only the Window attention mechanism as a baseline model to capture fine-grained spatiotemporal patterns within local windows. This configuration demonstrates that the local attention mechanism effectively models short-range dependencies but has inherent limitations in capturing long-range interactions across windows. The second configuration combines the Window attention mechanism with Window feature compression. By compressing features within each window, this mechanism highlights informative features while suppressing redundant features. The model's performance continues to improve, indicating that window-level feature compression enhances the discriminative power of local representations. The third configuration combines the Window attention mechanism and the Cross-window attention mechanism to achieve information exchange between different windows. This design explicitly models long-range spatiotemporal dependencies and global contextual relationships that cannot be captured by local attention mechanisms alone. This highlights the importance of cross-window interaction in complex video understanding tasks. Specifically, we evaluated different experiments, as shown in Figure 13. The model configurations are shown in Table 3.

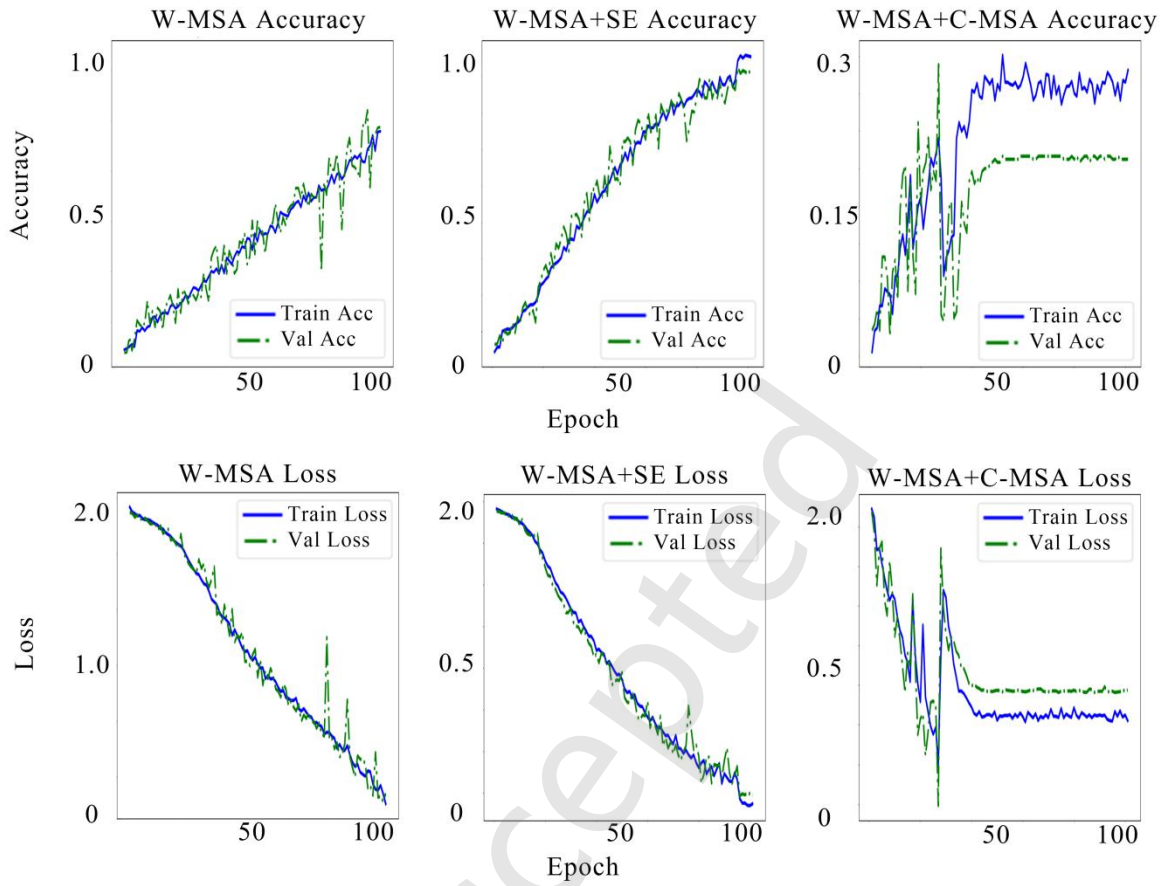


Fig 13 Ablation Experiment Process Based on Unimodal TransWindow

Table 3 TransWindow ablation experiment

Ablation	Acc.(%)	FLOPs(G)	Params(K)	MACs(M)
Window attention mechanism	67.21	182.695	457.543	212,213.173
Window attention mechanism+Window feature compression	89.96	182.849	475.975	212,367.389
Window attention mechanism+Cross-window attention mechanism	34.43	95.701	774.055	109,896.014

The ablation study demonstrates the contribution of each component in the proposed framework. The window-based attention mechanism effectively captures local dependencies but exhibits limited global modeling capability. The introduction of feature compression significantly improves performance with negligible computational overhead, indicating its effectiveness in enhancing feature representation. In contrast, the current cross-window attention design leads to performance degradation, suggesting that it does not effectively facilitate inter-window information exchange. In summary, adding feature compression after window-based attention can effectively improve accuracy, while cross-window attention requires further optimization to achieve efficient utilization of inter-window information.

4. CONCLUSION AND FUTURE WORK

This study proposes a multimodal visual and tactile framework for dangerous driving behavior detection by integrating pressure sensing and visual perception. The proposed TransWindow model achieves high recognition accuracy while maintaining low computational complexity, making it suitable for real-time deployment in intelligent cockpit systems. Experimental results demonstrate that multimodal fusion significantly improves detection performance compared with unimodal approaches. However, a limitation of this study is the reliance on data collected from a limited number of drivers, which may affect generalization. Future work will focus on expanding the dataset and improving cross-subject generalization to enhance the robustness and scalability of the proposed method.

Replication and data sharing

The source code and source data used in this study can be obtained by emailing the corresponding author.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (52302486, U22A202101).

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C., 2021. ViViT: A Video Vision Transformer. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 6816–6826.

Bresnahan, D.G., Li, Y., 2022. Measurement of Multiple Simultaneous Human Vital Signs using a Millimeter-Wave FMCW Radar. 2022 IEEE Radar Conference (RadarConf22), 1–5.

Bertasius, G., Wang, H., Torresani, L., 2021. Is Space-Time Attention All You Need for Video Understanding?. In: Proceedings of the 38th International Conference on Machine Learning (ICML).

Cao, Z., Hidalgo, G., Simon, T., Wei, S., Sheikh, Y., 2021. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. IEEE Trans. Pattern Anal. Mach. Intell., 43, 172–186.

Chen, H., Gao, R., Fan, L., Liu, E., Li, W., Tan, R., Li, Y., He, L., Cao, D., 2024. Scenario-Function System for Automotive Intelligent Cockpits: Framework, Research Progress and Perspectives. IEEE Trans. Intell. Veh., 9, 4890–4904.

Chen, J., Dey, S., Wang, L., Bi, N., Liu, P., 2021. Multi-Modal Fusion Enhanced Model For Driver's Facial Expression Recognition. 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 1–4.

Chen, W., Zhang, X., Chen, S., 2024. Fatigue Detection System for Extracting Driver's Eye Features. 2024 7th International Conference on Advanced Algorithms and Control Engineering (ICAACE), 891–894.

Chu, Y., Chowdhury, H.R., Mitul, A., Uddin, N., Hossain, M.J., Aslam, D., 2019. Nature of distracted driving in various

physiological conditions. Proc. SPIE 11139, Applications of Machine Learning, 111391B.

d'Ascoli, S., Touvron, H., Leavitt, M. L., Morcos, A. S., Biroli, G., Sagun, L., 2021. ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada, October 11-17, 12259–12268.

Fang, K., Xiang, S., Sminchisescu, C., 2023. AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR).

Fu, R.R., Wang, H., Zhao, W.B., 2016. Dynamic driver fatigue detection using hidden Markov model in real driving condition. Expert Systems with Applications, 63, 397-411.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., Tao, D., 2023. A Survey on Vision Transformer. IEEE Trans. Pattern Anal. Mach. Intell., 45, 87–110.

Hara, K., Kataoka, H., Satoh, Y., 2018. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6546–6555.

Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E., 2020. Squeeze-and-Excitation Networks. IEEE Trans Pattern Anal Mach Intell, 42, 2011–2023.

Hu, Y., Lu, M., Lu, X., 2019. Driving behaviour recognition from still images by using multi-stream fusion CNN. Machine Vision and Applications, 30, 851–865.

Huang, J., Xin, W., Shaojie, Q., 2023. An adaptive dangerous driving behavior analysis and prediction framework for Internet of Vehicles. Radio Engineering, 53, 1311–1320.

Kim, W., Jung, W.S., Choi, H.K., 2019. Lightweight driver monitoring system based on multi-task mobilenets. Sensors, 19, 3200.

Kuang, S., Liu, Y., Wang, X., Wu, X., Wei, Y., 2024. Harnessing multimodal large language models for traffic knowledge graph generation and decision-making. Communications in Transportation Research, 4, 100146.

Le, T.H.N., Zheng, Y., Zhu, C., Luu, K., Savvides, M., 2016. Multiple Scale Faster-RCNN Approach to Driver's Cell-Phone Usage and Hands on Steering Wheel Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 46–53.

Li, K., Gong, Y., Ren, Z., 2020. A Fatigue Driving Detection Algorithm Based on Facial Multi-Feature Fusion. IEEE Access, 8, 101244–101259.

Li, T., Zhang, Y., Li, Q., Zhang, T., 2022. AB-DLM: An Improved Deep Learning Model Based on Attention Mechanism and BiFPN for Driver Distraction Behavior Detection. IEEE Access, 10, 83138–83151.

Li, Y., Wu, C.-Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., Feichtenhofer, C., 2022. MViTv2: Improved Multiscale Vision Transformers for Classification and Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA, June 19-24, 4804–4814.

Liu, Y., Zhou, Y., Liu, Y., Xu, Z., He, Y., 2025. Intelligent fault diagnosis for CNC through the integration of large language models and domain knowledge graphs. Engineering, 53, 311-322.

Li, Z., Li, S., Li, R., Cheng, B., Shi, J., 2017. Online detection of driver fatigue using steering wheel angles for real driving conditions. Sensors, 17, 495.

Liu, Z., Chen, H., Lu, W., et al., 2018. Radar vital signal detection based on improved complete ensemble empirical mode decomposition with adaptive noise. Chinese Journal of Scientific Instrument, 39, 171–178.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 9992–10002.

Liu, Z., Ning, J., Cao, Y., Wei, Y.X., Zhang, Z., Lin, S., Hu, H., 2021. Video Swin Transformer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 3202–3211.

Lugaresi, C., Tang, M., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C., Yong, M.G., Lee, J., et al., 2019. MediaPipe: A Framework for Building Perception Pipelines. arXiv preprint arXiv:1906.08172.

Ma, X., Chau, L.P., Yap, K.H., 2017. Depth video-based two-stream convolutional neural networks for driver fatigue detection. 2017 International Conference on Orange Technologies (ICOT), 155–158.

Mei, B., 2025. Research on a DSM-YOLO-Based Detection System for Dangerous Driver Behavior in Auto-Driving. 2025 8th International Conference on Advanced Algorithms and Control Engineering (ICAACE), 713–716.

Mou, L., Zhao, Y., Zhou, C., Nakisa, B., Rastgoo, M.N., Ma, L., Huang, T., Yin, B., Jain, R., Gao, W., 2023. Driver Emotion Recognition With a Hybrid Attentional Multimodal Fusion Framework. IEEE Trans Affect Comput, 14, 2970–2981.

National Highway Traffic Safety Administration, 2024. Early Estimate of Motor Vehicle Traffic Fatalities in 2023. Washington, DC: National Highway Traffic Safety Administration.

Oh, G., Ryu, J., Jeong, E., Yang, J.H., Hwang, S., Lee, S., Lim, S., 2021. DRER: Deep Learning–Based Driver's Real Emotion Recognizer. Sensors, 21, 2166.

Rajpathak, T., Kumar, R., Schwartz, E., 2009. Eye Detection Using Morphological and Color Image Processing. Florida Conference on Recent Advances in Robotics, FCRAR 2009, 1–6.

Reddy, B., Kim, Y.H., Yun, S., Seo, C., Jang, J., 2017. Real-Time Driver Drowsiness Detection for Embedded System Using Model Compression of Deep Neural Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 438–445.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, D., Parikh, D., Batra, D., 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based

Localization. 2017 IEEE International Conference on Computer Vision (ICCV), 618–626.

Sikander, G., Anwar, S., 2019. Driver fatigue detection systems: A review. IEEE Trans. Intell. Transp. Syst., 20, 2339–2352.

Song, C., Song, Q., 2024. A Novel Driver Fatigue and Distraction Detection Method Based on Improved Swin Transformer. 2024 7th International Symposium on Autonomous Systems (ISAS), 1–6.

Wang, C., 2023. A Review on 3D Convolutional Neural Network. 2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA), 1204–1208.

Wang, K., 2020. Research on monitoring method of driver distraction and fatigue driving behavior based on deep learning. M.S. Thesis. Wuhan, China: Wuhan University of Technology.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L., 2022. PVT v2: Improved baselines with Pyramid Vision Transformer. Computational Visual Media 8, 415–424.

Xing, Y., Lv, C., Wang, H., Cao, D., Velenis, E., Wang, F.Y., 2019. Driver Activity Recognition for Intelligent Vehicles: A Deep Learning Approach. IEEE Trans. Veh. Technol., 68, 5379–5390.

Xu, D., Zhao, H., Guillemard, F., Geronimi, S., Aioun, F., 2019. Aware of scene vehicles—probabilistic modeling of car-following behaviors in real-world traffic. IEEE Trans. Intell. Transp. Syst., 20, 2136–2148.

Xu, X., Gu, H., Yan, S., Pang, G., Gui, G., 2019. Fatigue EEG Feature Extraction Based on Tasks With Different Physiological States for Ubiquitous Edge Computing. IEEE Access, 7, 73057–73064.

Yang, X., 2020. Vehicle trajectory feature recognition and analysis based on deep learning. M.S. Thesis. Nanjing, China: Nanjing University of Posts and Telecommunications.

Zulkarnanie, M.A., Shanmugam, K.S., Badruddin, N., Saad, M.N.M., 2022. Enhancements to PERCLOS Algorithm for Determining Eye Closures. 2022 International Conference on Future Trends in Smart Communities (ICFTSC), 76–81.

Minghui Zhao pursued his Master's degree at the School of Transportation Science and Engineering, Beihang University, under the supervision of Associate Professor Ren Bingtao, from September 2022 to June 2025. His research focused on intelligent cockpit systems, where he investigated driver behavior recognition through multimodal sensor fusion technology. Since 2016, he has been employed as a System Engineer at China Nuclear Control System Engineering Co., Ltd., where he currently leads the Intelligent Assembly Project, overseeing its development and implementation.

Bingtao Ren, Ph.D. in Engineering, is currently an Associate Professor and Master's Supervisor in the Department of Automotive Engineering, School of Transportation Science and Engineering at Beihang University. His primary research focuses on optimization control methods for intelligent vehicle safety and simulation modeling techniques.He has led and conducted a number of competitive research projects, including the National Natural Science Foundation of China (NSFC) Youth Program, sub-projects under the National Key R&D Program on New Energy Vehicles, specialized initiatives of the Ministry of Industry and Information Technology, and the Beijing Natural Science Foundation Youth Program.

Shichun Yang , Ph.D. in Engineering, serves as the Party Committee Secretary and Dean of the School of Transportation Science and Engineering at Beihang University, where he also holds the positions of Professor and Doctoral Supervisor. He concurrently directs the Zhejiang New Energy Vehicle Research Institute of Beihang University. Recognized as a leading scholar in the field of intelligent connected and new energy vehicles, he has been honored with several national distinctions, including National Leading Talent and Mid-aged & Young Leading Scientist in Technological Innovation by the Ministry of Science and Technology. He is also a Fellow of the China Society of Automotive Engineers.