vehicles

MDPI

Article

Mitigating Laboratory-to-Vehicle (Sim-to-Real) Spatial Shift in Edge-ADAS Driver Vigilance via Quantized Two-Stream Bottleneck Transformers

Mostafa Badr [1,2], Batoul R. Haidar [2] and Gamal A. Ebrahim [1,*]

[1] Computer and Systems Engineering Department, Faculty of Engineering, Ain Shams University, Cairo 11517, Egypt; 2300325@eng.asu.edu.eg or mostafa.badr@tkh.edu.eg
[2] School of Computing, Coventry University—Egypt Branch, Hosted at The Knowledge Hub Universities, Cairo 4841303, Egypt; batoul.haidar@tkh.edu.eg
* Correspondence: gamal.ebrahim@eng.asu.edu.eg

Abstract

Deploying complex driver vigilance models on resource-constrained edge hardware remains a challenging task, further obscured by the ubiquitous illusion of data leakage in the current literature. Unfrozen networks overfit static backgrounds during Laboratory-to-Vehicle (Sim-to-Real) transfer: 2×3 ablations on UTA-RLDD ($N = 25{,}178$) and DriverMVT ($N = 66$, $N = 270$) show partially unfrozen backbones (MobileNetV3-Small: 385,282; MobileViT-v2: 9,401,269 trainable parameters) collapse to 0.00% Specificity (TN = 0), while fully unfrozen models collapse in Sensitivity (33.33% and 18.18% TPR). Under density-matched augmentation (+1584 sequences), Generative AI matting (GenAIAug) introduces +63.0% Laplacian flicker ($p = 1.00 \times 10^{-7}$) and +107.8% zero-motion pixels ($p = 3.27 \times 10^{-15}$). This study proposes a Two-Stream Bottleneck Transformer with a 100% frozen MobileViT-v2 backbone (41,906 trainable [0.24%] of 17,466,235 parameters; 66.87 MB FP32) versus MobileNetV3-Small (34,738 trainable [3.61%] of 961,746). Complete freezing with motion isolation recovers 56.97% mean accuracy on MobileNetV3-Small and 71.21% FP32 ensemble accuracy (0.8838 ROC-AUC; 87.88% accuracy, 12.12% FPR at Youden's J^*) on MobileViT-v2. Selective FX INT8 Post-Training Quantization (53.10 MB, −20.6%) acts as a negative logit recenterer ($\Delta z_0 = -10.878$) on traditional augmentations (TradAug), boosting Fold 1 accuracy (53.03% → 69.70%, FP → TN = 20) and 5-fold INT8 Ensemble accuracy to 74.24% (15.15% FPR on $N = 66$) and 76.30% (80.48% Specificity on $N = 270$). Desktop x86 CPU simulation reaches 4.99 FPS (200.42 ms), establishing algorithmic feasibility, while multi-window temporal hysteresis for suppressing residual false alarms (12.12–19.52% FPR) and hardware-in-the-loop profiling on physical automotive SoCs remain essential for commercial deployment.

Keywords: Edge-ADAS; laboratory-to-vehicle (sim-to-real) domain shift; post-training quantization; architectural inductive bias; driver vigilance monitoring; spatial shortcut learning

Check for updates

Received: 14 October 2026
Revised: 16 November 2026
Accepted: 19 November 2026
Published: 24 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Advanced Driver Assistance Systems (ADAS) and Driver Monitoring Systems (DMS) are increasingly important parts of today's vehicle technology [1]. These systems are key to both road safety and user satisfaction. Their importance is especially relevant as autonomous driving advances from Level 2 to Level 3 conditional automation. Traffic accidents represent a major global public health problem, causing approximately 1.35 million deaths annually.

More than 90% of such crashes are attributed to human error, such as inattention, distraction, and fatigue [2–6]. For example, drowsy driving is a major cause of injuries and deaths, conservatively estimated to be involved in about 91,000 crashes in the U.S. annually [7–10]. Therefore, there is a strong need for immediate and reliable intervention systems to detect and mitigate these risks [11].

Initial driver vigilance detection methods were based on the monitoring of vehicle dynamics (e.g., steering wheel movement, deviation of lane position) or physiological signals (e.g., Electroencephalography (EEG), Electrocardiography (ECG)) [12,13]. While physiological methods can achieve high accuracy, they are often intrusive, and require the use of cumbersome wearable sensors, which limits their acceptance and applicability in real-world driving conditions [14]. In response to these limitations, vision-based driver monitoring systems (V-DMS) that make use of in-cabin cameras and deep learning (DL) techniques have emerged as the preferred non-intrusive solution [15,16]. They track key behavioral metrics, such as eye state, mouth activity, and head pose, to evaluate the driver's attention and fatigue level. Convolutional Neural Networks (CNNs) have played a major role in driving this field to great lengths [6,10,12].

Despite the technological momentum, there are two critical challenges that limit the wide application of DL-based DMS: the requirement for strict real-time performance and the need for robust generalization in the face of a broad range of driving scenarios. Firstly, V-DMS must provide real-time inference with low latency on resource-constrained embedded platforms with limited computational power and memory [17–22]. Secondly, models must be robust and generalize well to complex real-world scenarios, which involve a variety of challenging conditions such as varying illuminations, occlusions, and driver movements [6,12]. Such fragility is caused by the difficulty and cost of collecting and annotating a sufficiently large real-world dataset that captures the diverse edge cases required for robust driving safety [23]. This domain gap represents a fundamental Laboratory-to-Vehicle (Lab-to-Vehicle, often termed Sim-to-Real) spatial transfer challenge from stationary indoor laboratories to dynamic real-road vehicle cabins.

Compounding the generalization deficit, the recent literature suffers from a pervasive methodological flaw: the common practice of the "data leakage illusion." In an attempt to benchmark performance, many works test complex models on small-scale laboratory databases (e.g., UTA Real-Life Drowsiness Detection (UTA-RLDD) [24]), and report unrealistically high, nearly perfect classification accuracies often over 95% under randomized frame-level splits [13,25]. However, as exposed by Ghoddoosian et al. [24] and Magán et al. [18], this anomalous high performance is an artifact of frame-level data leakage rather than true algorithmic intelligence, with subject-independent 5-fold cross-validation dropping baseline accuracy to 63.0–65.2%.

This flawed validation procedure produces an inherent "memory effect" where the convolutional network bypasses extraction of generalizable fatigue semantics. The convolutional network simply overfits to static laboratory backgrounds, personal accessories, and identities of subjects present in the leaked frames. Consequently, models evaluated under such procedures yield high apparent accuracy in controlled settings, but perform poorly in real-world driving scenarios [3,18,26].

Historically, the literature tackles the aforementioned bottlenecks of computational efficiency and diversity in an isolated manner, thus resulting in a dichotomy in the proposed solutions. On the one hand, to satisfy the constraints of edge-hardware capabilities, prior works focus on Model Compression methodologies [27,28] such as Knowledge Distillation (KD) [9], pruning, and Post-Training Quantization (PTQ) [21,22,29] to optimize the accuracy–speed trade-off. In these directions, PTQ is usually regarded as purely de-

structive to accuracy, imposing strict constraints on memory footprint in exchange for degraded performance.

However, this view does not account for a special case of discrete 8-bit integer (INT8) mapping that inherently carries a regularizing inductive bias. Specifically, the discretization of weight topologies formed under traditional data perturbations allows for effective damping of spatially varying artifacts and overfitted continuous activations. As a result, targeted numerical compression exposes an embedded generalizable feature subspace, enabling high performance under non-ideal operational conditions.

However, applying compression to models trained on non-diverse data results in severe failures in the compressed model. It becomes generalization-lossy and can be "almost inoperative" in cross-dataset evaluations [3,30]. Moreover, the compression process can amplify predictive degradation: quantization or distillation often induces large accuracy drops (between 3.6% and 17%) if the model fails to learn diverse features during training [20,22,26,29].

On the other hand, to tackle this lack of robustness, the field has turned towards Generative Data Augmentation (GDA) to synthetically generate training data to diversify the environment [23,31]. However, this does not address the underlying mechanisms of the Lab-to-Vehicle domain shift. In practice, it is observed that when a conventional, unfrozen network is deployed from constrained laboratory environments into a novel vehicle cabin, it collapses spatially into zero True Negatives. This occurs because a conventional network is prone to overfitting to environmental invariants: it memorizes a static background as a spatial anchor instead of generalizing fatigue semantics. Moreover, simply attempting to bridge this gap using generative synthesis is not a complete solution: these GDA pipelines introduce temporal artifacts that disrupt the continuous optical flow required by sequential models [32].

Specifically, generative VFX alpha-matting (GenAIAug) introduces +63.0% active-foreground Laplacian boundary flicker ($p = 1.00 \times 10^{-7}$) and +107.8% static-plate zero-motion pixels ($p = 3.27 \times 10^{-15}$), severely corrupting frame differencing.

To truly address this vulnerability, a structural intervention is needed. For example, Figure 1 shows that a conventional unfrozen architecture learns to overfit on static lab environments, achieving artificially high frame-level accuracy, but failing severely in unseen domains. By contrast, the proposed framework freezes a domain-aligned hybrid CNN-Transformer visual backbone (MobileViT-v2 initialized from `mosesb/drowsiness-detection-mobileViT-v2`, pretrained on ImageNet-1K and fine-tuned on >40,000 facial drowsiness images with 0.0% subject or frame overlap with UTA-RLDD or DriverMVT). This structural intervention leaves only 41,906 trainable temporal parameters (0.24%) out of 17,466,235 total inference parameters (66.87 MB FP32)—compared to the 961,746-parameter MobileNetV3-Small baseline (34,738 trainable [3.61%] when frozen)—enforcing a strict architectural inductive bias to extract invariant fatigue dynamics for robust Laboratory-to-Vehicle transfer.

A critical synthesis of the literature reveals model compression and data augmentation to be individually mature fields of study. With this said, there remains a fundamental research gap in the domain concerning which augmentation strategy (Generative AI vs. algorithmic perturbations) engenders the most quantization-resilient feature topologies within resource-constrained ADAS environments. This work investigates the aforementioned optimization space in depth by applying strict 8-bit PTQ to distinct training corpora, forcing a revision of the understanding of how features survive the truncation process.

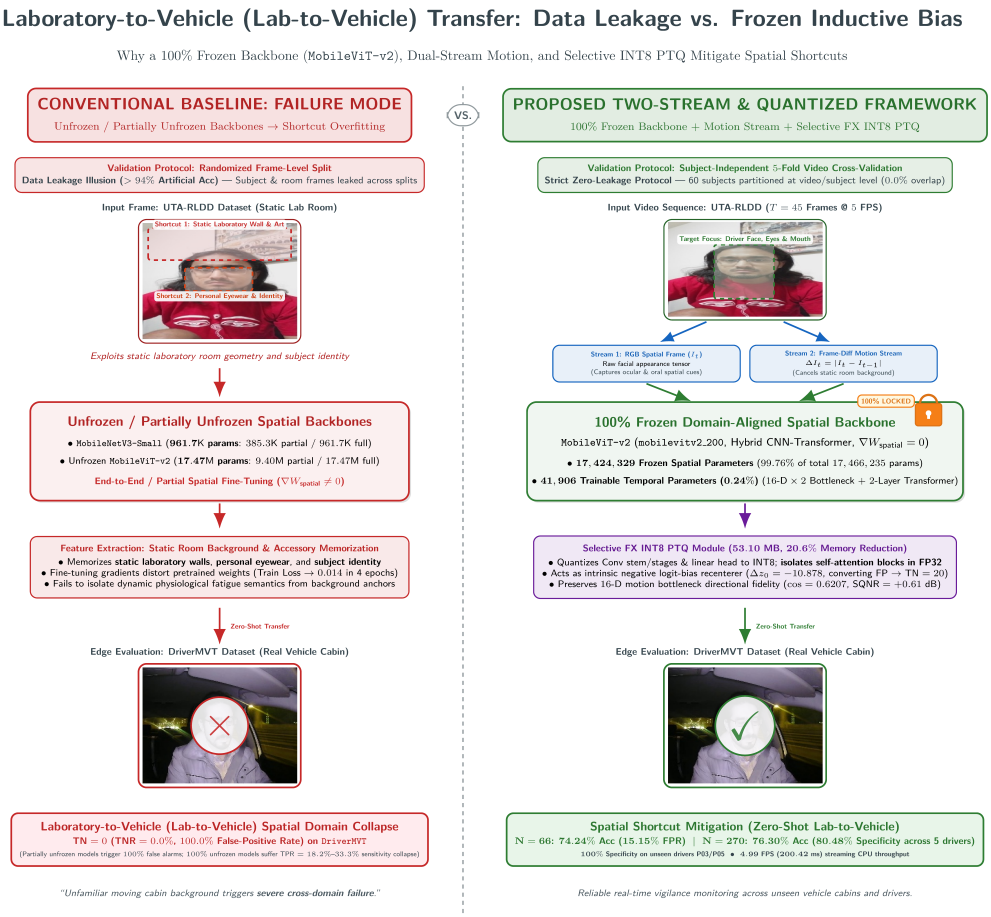

Figure 1. Conceptual illustration of the Lab-to-Vehicle spatial domain shift and the Data Leakage Illusion between conventional unfrozen architectures (**left**) and the proposed frozen-backbone zero-leakage framework (**right**). Unmodified human facial frames originate from publicly available academic benchmarks: Participant 09 from UTA-RLDD [24] and Participant P05 from DriverMVT [33].

To bridge this critical gap, this paper proposes an end-to-end driver vigilance monitoring framework, which is engineered to be efficient, robust, and numerically resilient even under stringent hardware constraints, thus rendering itself amenable to real-world ITS deployment. By rigorously comparing synthetic and traditional data pipelines under edge compression, the primary contributions are as follows:

- **Quantitative Scarcity & Density-Matched Pipeline:** This study establishes an unsupervised scarcity metric across $N = 565{,}273$ UTA-RLDD frames using DINOv2 ($K = 200$, DBI = 1.3260), CLIP ($K = 100$), and ResNet-50 ($K = 100$). This quantifies a 13.75:1 static-to-scarce frame imbalance ($\rho_{\max/\min}$ =103.94:1) and targets high-scarcity subjects P26 (Fold 3) and P55 (Fold 5) with +1584 participant- and density-matched augmented sequences each.
- **Controlled** 2×3 **Freezing & Stream-Isolation Ablations:** Controlled ablations on UTA-RLDD and DriverMVT reveal that partially unfrozen backbones (MobileNetV3 and MobileViT-v2) collapse to 0.00% Specificity (TN = 0) and fully unfrozen models collapse to 33.33% and 18.18% Sensitivity. Enforcing 100% spatial backbone freezing recovers 56.97% mean accuracy on MobileNetV3-Small and 71.21% FP32 ensemble accuracy (0.8838 ROC-AUC; 87.88% accuracy/12.12% FPR at Youden's J^*) on MobileViT-v2, while Stream 2 motion isolation suppresses false alarms (45.45% vs. 15.15% TNR when masked).

- **Physical Optical-Flow Proof of Generative Distortions:** Physical motion analysis proves that Generative AI VFX alpha-matting (GenAIAug) introduces +63.0% active-foreground Laplacian boundary flicker ($p = 1.00 \times 10^{-7}$) and +107.8% static-plate zero-motion pixels ($p = 3.27 \times 10^{-15}$). This corrupts frame-differencing features and degrades cross-domain specificity (27.27% TNR).
- **Selective FX INT8 PTQ & Edge Benchmarks:** Mechanistic logit analysis demonstrates that Selective FX INT8 PTQ (53.10 MB, keeping self-attention in FP32) acts as an activation-norm damper and negative logit-bias recenterer ($\Delta z_0 = -10.878$, converting 20 False Positives to True Negatives). This elevates 5-Fold INT8 Ensemble accuracy to 74.24% (15.15% FPR on $N = 66$) and 76.30% (80.48% Specificity across 5 drivers on $N = 270$; 100% TNR on unseen P03/P05), achieving 4.99 FPS (200.42 ms) 4-thread CPU streaming throughput.

The remainder of this paper is organized as follows. Section 2 reviews SOTA driver vigilance monitoring, data leakage, edge compression, and data augmentation. Section 3 details the zero-leakage protocol, frozen-backbone architecture, and Selective FX INT8 pipeline, followed by experimental evaluations in Section 4. Section 5 analyzes the Lab-to-Vehicle spatial collapse, generative augmentation artifacts, and INT8 regularization mechanics, and Section 6 concludes the paper.

2. Related Work

2.1. Deep Learning for Driver Vigilance Monitoring

Rapid progress in Computer Vision (CV) and Image Processing (CVIP) has established vision-based approaches as the primary focus in non-intrusive driver monitoring. Such systems continuously monitor behavioral cues such as facial expressions, eye movements, and head positions using in-vehicle cameras [7,12,15]. Driver vigilance monitoring systems are based mainly on the analysis of visual cues that indicate fatigue or distraction. These systems leverage advanced CNNs and recurrent architectures to extract spatio-temporal cues.

The eye and mouth states are the most important indicators of driver fatigue. Deep learning models are commonly used to determine if a driver's eyes are open or closed [12]. This classification forms the basis for important metrics like the Percentage of Eye Closure (PERCLOS) [8,34]. Drowsiness behaviors such as prolonged eye closure and microsleep have different temporal dependencies, thus hybrid models combining CNN with Long Short-Term Memory (LSTM) units have been proposed [12,35]. Yawning is another important indicator which is usually detected by analyzing the Mouth Aspect Ratio (MAR) or the Frequency of Mouth Opening (FOM) [17,34]. Finally, multi-task CNNs have appeared to combine the features of both eyes and mouth in a single framework [12,36]. This allows the simultaneous extraction of the facial states and the direct determination of the overall vigilance status (awake or drowsy) of the driver.

Driver distraction is often posed as an action recognition problem, whereby DL systems classify Non-Driving-Related Activities (NDRAs) [37]. Head pose, gaze direction and hand gestures [7,34,35,37] are crucial behavioral features. Deep networks, e.g., Multi-task Cascaded Convolutional Networks (MTCNN), are necessary to localize the driver's face and extract the facial landmarks [9]. Head pose estimation is important for predicting the gaze direction of the driver and detecting inattentive behavior [4,7,35] when the attention is not on the road. These traditional approaches heavily rely on geometrical algorithms such as solvePnP. To identify more complex and time-varying distracted behaviors, researchers are increasingly adopting advanced deep learning techniques, such as spatio-temporal CNNs (3D-CNNs) and transformer-based attention mechanisms [35].

Beyond visual sensors, alternative approaches optimize sequence modeling on vehicle dynamic signals. For instance, Li et al. [38] combined a Gated Recurrent Unit (GRU) with

an attention mechanism and the Whale Optimization Algorithm (WOA) to weight time-series features from steering wheel angles, throttle positions, and vehicle speed, achieving 89.84% accuracy on real-road data. However, while dynamic telemetry avoids facial camera occlusions, it cannot observe micro-facial fatigue dynamics (such as PERCLOS or MAR) and remains dependent on vehicle-specific CAN-bus hardware. Consequently, camera-based visual monitoring remains essential, necessitating solutions to Laboratory-to-Vehicle (Lab-to-Vehicle) spatial domain transfer across diverse vehicle interiors.

Although deep learning models often report higher classification accuracy than traditional CV approaches, early convolutional architectures established the foundational baseline for non-invasive driver monitoring. Initial deep feature extraction studies demonstrated that neural networks can be trained to learn non-trivial relationships between multiple physiological parameters, including eye state and mouth aspect ratios, simultaneously. Thus, such models capture the patterns required for vigilance estimation in a controlled environment [2,12]. However, these accuracy scores must be interpreted with caution due to potential flaws in model evaluation. Indeed, heavily parameterized architectures tend to overfit the training distribution, particularly when the background is static, as in laboratory environments. As a result, such models fail to learn generalizable representations of fatigue, rendering their reported in-distribution accuracy unrepresentative of real-world driving performance.

Furthermore, the adoption of such models in commercial ADAS is severely limited by their extensive computational requirements [12,20,34,39]. The complexity of such DL architectures requires substantial processing resources [12,17,34] beyond the constrained memory footprint and processing power of low-cost embedded hardware [20,34]. Due to the necessity of real-time processing (typically 10–30 FPS and sub-second latencies) for accident prevention [20], this efficiency-versus-generality trade-off has motivated extensive research into model optimization techniques. However, the perceived superiority of many existing state-of-the-art approaches is undermined by a pervasive methodological issue typically termed the "data leakage illusion."

To evaluate performance, many studies benchmark heavily parameterized models across constrained, homogeneous, laboratory-limited datasets (e.g., UTA-RLDD) using randomized, frame-level train-test splits [13,25]. Such approaches routinely report over 95% accuracy because this evaluation protocol introduces an inherent "memory effect" [26], wherein convolutional networks memorize subject identity, personal accessories (e.g., glasses, necklaces), and the static background of the training set rather than invariant fatigue-related facial transformations. By contrast, pioneering zero-leakage evaluations by Ghoddoosian et al. [24] (65.2% accuracy) and Magán et al. [18] (63.0% 5-fold CV accuracy) exposed this memory effect. To properly evaluate algorithmic generalization and ensure robustness in vehicular environments, the field must adopt rigorous subject- and video-independent cross-validation protocols without data leakage (Table 1).

Table 1. Comparison of SOTA driver vigilance monitoring frameworks regarding validation protocols, cross-domain testing, and edge deployment capabilities.

Reference	Validation Protocol	Cross-Domain Testing	Edge Optimization	Parameter Efficiency (Trainable/Total)
Prior Works [13,25]	Frame-Level (Leakage)	No	None	Low (Heavy CNNs)
Ghoddoosian et al. [24]	Zero-Leakage (Video)	No	None	Medium (HM-LSTM)
Magán et al. [18]	Zero-Leakage (5-Fold)	No	None	Low (1.20M/5.21M)
Proposed Framework	**Zero-Leakage (5-Fold)**	**Yes (DriverMVT $N = 66$ & $N = 270$)**	**Selective PTQ INT8**	**High (41.9K/17.47M [0.24%])**

Note: Bold formatting highlights the proposed framework.

2.2. Model Compression in Embedded Vision Systems

Given the need to deploy high-capacity deep learning models for real-time driver monitoring, optimizing their computational efficiency is essential. As discussed in the previous section, DL models capable of achieving high predictive accuracy are often associated with considerable computational complexity. This characteristic limits their direct application in commercial ADAS, which are typically equipped with resource-constrained hardware, such as NVIDIA Jetson or automotive system-on-chip (SoC) platforms [6,20,29]. To meet the strict real-time requirements of ADAS (e.g., 5–30 FPS), model compression and acceleration (MCA) techniques [29] are employed to balance model size, computational cost, and predictive accuracy across four major optimization strategies.

Parameter Quantization reduces the numerical precision of model parameters and activations, transforming them from high-precision floating-point representations (e.g., 32-bit FP32) to lower-precision encodings (e.g., INT8) [40,41]. This significantly lowers storage demands and accelerates inference on embedded systems [22,29]. However, the existing literature has predominantly examined PTQ from the perspective of hardware optimization, whereas the algorithm-level regularization effects of numerical truncation remain underexplored. Specifically, prior work lacks a mechanistic analysis of how mapping continuous activations into discrete 8-bit bins interacts with differently augmented feature topologies, and whether numerical truncation can selectively attenuate high-frequency spatial noise without degrading core temporal semantics.

Alternative model compression strategies, such as model pruning and knowledge distillation, are valid and explored techniques for reducing network redundancy and transferring learned representations [28,42]. However, while pruning aims at minimizing theoretical floating-point operations (FLOPs), unstructured sparse topologies often require specialized sparse co-processors, in order to obtain actual hardware speedups in general-purpose embedded environments [28,43,44]. On a similar note, KD incurs an important computational overhead, during the development phase, since it requires the simultaneous training and optimization of large teacher networks, in conjunction with the targeted compact student architectures [9,40].

Recently, Wang et al. [45] demonstrated selective knowledge distillation to adapt compact student networks for resource-constrained edge transport, while Zhuang et al. [46] deployed lightweight MobileNetV3 backbones for in-vehicle infrared object detection on embedded SoCs. However, full compression across all layers can degrade critical spatio-temporal features. By contrast, the proposed framework employs Selective FX INT8 quantization (53.10 MB), retaining self-attention blocks in FP32 to avoid the 50.00% specificity collapse observed under full INT8 quantization.

In addition to algorithmic approaches to model compression, several architectural-level design paradigms have pursued intrinsically efficient networks. Architectures such as MobileNet, SqueezeNet, and EfficientNet employ depthwise separable convolutions or fire modules, respectively, to substantially reduce computational operations from the ground up [40]. Such networks are well suited for execution on low-cost microcontrollers or embedded platforms such as the Raspberry Pi [16,22]. Crucially, parameter efficiency under backbone freezing must distinguish between total inference parameters and the small subset of trainable parameters restricted to the temporal fusion head (0.24% of total parameters in MobileViT-v2 versus 3.61% in MobileNetV3-Small, detailed in Section 3.1.4).

By reducing theoretical FLOPs and parameter counts, algorithmic compression and architectural optimization satisfy tight single-frame latency budgets (often under 100 ms) on off-the-shelf embedded vision hardware [20,28].

However, a common assumption across these techniques is an inevitable trade-off between compression ratio and classification accuracy, where aggressive quantization is viewed solely as a risk to predictive reliability [28,40].

This assumption is, however, primarily evaluated in-distribution and does not account for the interplay between quantization operators and the topological stability of learned features under domain shift. Rather than universally degrading performance, it is hypothesized—and empirically demonstrated in this study—that when applied to structurally stable, traditionally augmented representations, selective numerical truncation can act as an implicit regularizer: by damping and recentering shifted latent activations, quantization can improve cross-domain specificity for edge deployment.

Ultimately, however, conventional compression methods are primarily validated in homogeneous, sanitized settings rather than across varying vehicle cabin configurations. Consequently, the literature lacks a unified framework that first enforces a structural inductive bias on an overfitted spatial backbone before applying targeted INT8 quantization for generalizable edge efficiency.

2.3. Generative Data Augmentation for Driver Analysis

To address the limitation of scarce diverse annotated naturalistic driving data, recent ADAS research works have begun to adopt GDA to generate synthetic samples and simulate unpredictable or occluded driver behavior using advanced generative models like Generative Adversarial Networks, Variational Autoencoders, and Diffusion Models [23,32,47,48]. Such methods augment the training corpora, enabling discriminative models to learn from a more diverse distribution of environmental conditions while avoiding the high cost and danger of collecting real-world data [31,49,50].

Static generative models have made significant strides in synthesizing high-fidelity spatial priors and generating rare but safety-critical edge cases such as extreme environmental occlusions or illumination conditions. Such cases are both unsafe and extremely challenging to capture in real-world naturalistic driving studies [47,48]. For example, Li et al. [51] developed VehicleGAN for pose-guided synthetic image generation and joint metric learning in vehicle re-identification, demonstrating that synthetic data can effectively augment static visual representations across surveillance cameras. However, while pose-guided image generation and feature blending succeed in static image tasks, sequential video-based driver fatigue monitoring relies on strict inter-frame optical-flow continuity ($\Delta I_t = |I_t - I_{t-1}|$), where frame-wise synthesis fails.

However, the practical applicability of these generative techniques for sequential video modeling is severely challenged by accompanying temporal artifacts. Generative synthesis, while providing strong spatial priors, is challenging to apply to sequential video data due to frame-to-frame inconsistencies. The process of generating or editing images frame-by-frame (via latent diffusion, image-to-image translation, or alpha-matte compositing) introduces boundary and pixel-level artifacts [32,52].

These microscopic synthetic artifacts are highly detrimental to sequential fatigue modeling. Fatigue and inattention are gradual physiological processes involving smooth behavioral transitions over time. When frames are synthesized or composited independently, subtle inter-frame inconsistencies accumulate, introducing pronounced discontinuities in the optical-flow domain. This poses a severe challenge for temporal architectures, such as temporal transformers, which model the evolution of features across time. Instead of capturing smooth physiological transitions of fatigue, the transformer's temporal attention mechanisms are confounded by spurious high-frequency boundary flicker introduced by frame-wise synthesis, causing temporal feature discontinuity and fold-wise cross-validation volatility.

In addition to the previously mentioned challenges, most GDA research evaluates synthesized models solely in restricted, in-distribution settings. The current literature rarely investigates whether synthetic generative features can withstand the Lab-to-Vehicle spatial domain shift when transferred to novel, dynamic vehicular cabins. Moreover, hardware-aware analysis of these augmented topological structures is absent in the literature. In particular, the literature does not investigate whether the fragile, artificially generated features can survive the numerical truncation imposed by the 8-bit PTQ. This step is mandatory for real-world automotive edge deployment [20,30,41].

Specifically, frame-wise VFX alpha-matting and static background plate compositing (GenAIAug) introduce two physical spatio-temporal distortions into the motion stream (ΔI_t). First, alpha-matte boundary instability causes +63.0% active-foreground Laplacian boundary flicker ($p = 1.00 \times 10^{-7}$ vs. No-Aug; +77.2% vs. TradAug, $p = 1.93 \times 10^{-7}$) and +22.7% Farnebäck optical-flow spatial shear ($p = 3.68 \times 10^{-4}$). Second, static background vehicle plates induce +107.8% zero-motion pixels (81.42% vs. 39.19%, $p = 3.27 \times 10^{-15}$), depriving the motion stream of natural background flow and triggering 8-bit PTQ collapse.

3. Methodology

Modern computer-vision-based driver monitoring systems face the dual challenge of achieving real-time performance on embedded hardware while maintaining robust perception under varying environmental conditions [12,20]. Previous work struggles to resolve this dilemma, typically requiring parameter-heavy deep learning architectures that either exceed the latency budgets of embedded systems or lack generalization ability under out-of-distribution conditions [7,8,12]. To address this trade-off, a two-stage experimental framework is proposed to decouple data-level topology optimization from hardware-oriented numerical compression—thereby enabling independent analysis of cross-domain generalization and quantization robustness [40].

Phase 1, namely Spatio-Temporal Optimization and Zero-Leakage Evaluation, focuses on developing a structurally stable visual perception engine capable of withstanding Lab-to-Vehicle domain shifts. Acknowledging that conventional DMS models frequently undergo spatial background memorization when transitioning from confined laboratory environments [18,24] to real-world vehicle cabins [53], this phase enforces a strict zero-leakage validation protocol. In addition, this phase conducts a controlled comparative analysis between traditional data perturbations and generative visual effects synthesis [23,50]. This enables the proposed framework to empirically isolate which augmentation strategy yields the most invariant spatio-temporal topologies prior to hardware compression.

Phase 2, Edge-Aware Model Compression, takes the unique feature topologies from Phase 1, and applies them to a targeted 8-bit post-training quantization pipeline [21,40,44]. This is motivated by memory and latency constraints of embedded automotive systems [20]. However, the work does not view PTQ as a mere memory reduction tool, but rather analyzes the interplay of severe numerical precision reduction and spatial feature regularity as a way to maximize throughput in INT8 deployment while maintaining the solution accuracy [20].

3.1. Phase 1: Spatio-Temporal Optimization and Zero-Leakage Evaluation

To achieve structural generalization prior to edge compression [12,34], Phase 1 intentionally utilizes a constrained single-modal laboratory benchmark (Section 3.1.1) to emulate the pervasive Lab-to-Vehicle spatial domain shift [24].

This engineering phase is conducted via a sequential two-stage pipeline. First, Scarcity Analysis provides a quantitative measure of feature space redundancy and class imbalance [54] to motivate the need for artificial data synthesis from a mathematical standpoint. Second, Spatio-Temporal Augmentation conducts a direct comparison study between

conventional algorithmic data perturbations and state-of-the-art generative image synthesis [23,48–50]. The framework is evaluated using a strict zero-leakage subject-independent cross-validation protocol. Models trained on the baseline, unaugmented data are compared with models that utilize the traditional and generative pipelines. This comparison aims to empirically isolate which augmentation topology is successful in extracting invariant motion dynamics and curbing spatial shortcut learning, thus providing the structural resilience necessary for subsequent numerical compression (Phase 2).

3.1.1. Dataset Selection and Domain Gap Analysis (UTA-RLDD and DriverMVT)

For the first phase, the development and optimization of the core visual perception engine, the UTA-RLDD dataset was selected as the main single-modal benchmark [14]. Widely adopted in Intelligent Transportation Systems research, this dataset contains around 30 h of RGB video from 60 distinct subjects [13,24]. The advantage of this dataset is the detailed annotation scheme that categorizes the driver vigilance into three significant states: Alert, Low Vigilance and Drowsy [13,24].

Furthermore, the UTA-RLDD corpus is valuable because it captures genuine fatigue events rather than simulated expressions by observing subjects as they naturally transition into drowsiness. This approach is effective for mitigating the inherent bias of datasets derived from posed drowsiness behavior [24]. To establish crisp physiological boundaries, the ambiguous transitional Low Vigilance state is excluded, mapping Alert to class 0 (12,468 sequences) and Drowsy to class 1 (12,710 sequences) so that subjective self-reported ambiguity at the intermediate vigilance boundary does not confound binary safety-intervention decision thresholds (Alert versus Drowsy warning) during cross-domain evaluation. This yields 25,178 base 45-frame sequences (9.0 s duration at 5 FPS, temporal stride of 15 frames) for subject-independent 5-fold cross-validation.

However, a closer technical review uncovers a critical environmental constraint that characterizes the methodological core. Although labeled as a realistic dataset, the videos were collected in controlled, static indoor environments—primarily academic laboratories or residential rooms where participants performed stationary tasks such as reading or watching a screen [18,24]. Consequently, the visual input space is restricted to static backgrounds, stable indoor lighting, and zero vehicle-induced camera motion. This limitation implies that the training corpus lacks critical real-world vehicular characteristics, such as rapid ambient illumination shifts, high-frequency cabin vibrations, and complex background clutter typical of vehicle interiors [24].

This representation gap induces a pronounced spatial domain shift between stationary laboratory training data and target in-vehicle ADAS deployment [3]. Models trained on UTA-RLDD with unfrozen spatial backbones suffer severe performance degradation when deployed in real vehicle cabins, as their spatial encoders overfit to static indoor backgrounds. To quantify this vulnerability, the Driver Monitoring including Video and Vehicle Telemetry Information (DriverMVT) dataset is utilized [33] as a realistic out-of-distribution vehicular benchmark. Captured directly in dynamic driving environments, DriverMVT provides the complex in-cabin visual stimuli absent from the laboratory-centric UTA-RLDD corpus.

Pairing UTA-RLDD with DriverMVT specifically isolates background clutter, natural sunlight and shadow transitions, and cabin vibration under comparable frontal to near-frontal camera framing, whereas alternative public benchmarks either remain confined to stationary laboratory/simulator cabins (e.g., NTHU-DDD) or introduce confounding multi-angle oblique camera and infrared sensor shifts (e.g., DMD, StateFarm).

For out-of-distribution evaluation on DriverMVT (5 distinct subjects P01–P05 recorded at 30 FPS and downsampled 6× to 5 FPS, with the bottom 8% telemetry banner cropped prior to 224×224 resizing), frame-level annotations in `info.csv` containing 619 '`drowsiness`'

frames and 179 '`yawning`' frames (798 fatigued frames forming 279 distinct anchor events) were mapped to binary Drowsy (1), while contiguous anomaly-free driving segments (≥300 frames) were mapped to Alert (0). Each evaluation sample represents a distinct, non-overlapping 45-frame (9.0 s) video sequence.

The historical $N = 66$ balanced benchmark comprises 33 positive Drowsy sequences (1.0 s anchor cooldown) paired with 33 subject-matched Alert sequences across P02 (29 Drowsy/29 Alert), P04 (2 Drowsy/2 Alert), and P05 (2 Drowsy/2 Alert), as P01's 5 drowsy anchors fell within 132 frames of the video boundary and P03's single drowsy recording had a corrupted MP4 `moov` header.

To address small-sample concerns and evaluate all five drivers, an expanded $N = 270$ benchmark was constructed across P01–P05, extracting 60 valid Drowsy sequences (0.5 s cooldown: 52 from P02, 4 from P04, and 4 from P05) and 210 non-overlapping Alert sequences (42 sequences per driver across all five subjects P01–P05). Together, these two test protocols serve complementary diagnostic roles: the 1:1 balanced $N = 66$ set provides an unconfounded equal-prior reference for sensitivity, specificity, and ROC-AUC comparisons without majority-class inflation, whereas the expanded $N = 270$ cohort (77.8% Alert prior across all five drivers) stresses false-alarm resilience under realistic alert-dominated highway driving priors.

Accordingly, Phase 1 trains baseline models on UTA-RLDD to evaluate whether traditional perturbations or synthetic cabin compositing better reduce spatial shortcut learning on the unseen DriverMVT domain.

3.1.2. Unsupervised Scarcity Analysis

In order to go beyond the qualitative claims about the lack of variety in either visual appearance or the surrounding environment in the UTA-RLDD dataset, an empirical evaluation of the number of spatial features present in the training set was carried out. Simply relying on the qualitative descriptions of the laboratory collection conditions [18,24] is not sufficient for selecting specific ranges for individual parameters during the augmentation process. Instead, an unsupervised clustering procedure was utilized to determine the distribution of visual diversity and evaluate the domain scarcity of the original input sample before any synthetic data inclusion.

The analysis was performed by projecting the raw image space $\mathcal{X}$ into a high dimensional latent manifold. Then, using pre-trained vision backbones $f_\theta : \mathcal{X} \to \mathbb{R}^d$—specifically self-supervised DINOv2 (ViT-S/14, $d = 384$), vision-language CLIP (ViT-B/32, $d = 512$), and ImageNet-pretrained ResNet-50 ($d = 2048$, final classification head removed)—high-level spatial feature vectors $v_i = f_\theta(x_i)$ were extracted for all frames in the dataset $\{x_i\}_{i=1}^N$ [8,37]. In order to identify the intrinsic sub-structures, the obtained vectors were then partitioned by a MiniBatch K-Means algorithm, which aims at optimizing the within cluster sum of squares:

$$\arg\min_{\mathbf{C}} \sum_{k=1}^{K} \sum_{v_i \in C_k} \|v_i - \mu_k\|^2 \tag{1}$$

where $\mathbf{C} = \{C_1, C_2, \ldots, C_K\}$ represents the distinct spatial clusters and μ_k denotes the mean centroid of cluster C_k.

This quantitative mapping revealed the representational deficit within the base corpora. A majority of the training data collapsed to high-density clusters (C_{static}) defined strictly by the static environmental invariants (same office walls, controlled lighting, fixed camera angles)—while the clusters capturing rare dynamic variations (C_{dynamic}), like extreme head

poses, lighting variations, or heavy occlusions, were severely under-represented – yielding an extreme visual Imbalance Ratio (ρ):

$$\rho = \frac{|C_{\text{static}}|}{|C_{\text{dynamic}}|} \gg 1 \tag{2}$$

To establish a reproducible mathematical assignment criterion between C_{static} and C_{dynamic}, each cluster C_k is assigned a log-inverse frequency scarcity score $S(C_k) = \ln(N/|C_k|)$, where clusters exceeding the scarcity threshold $\tau_{\text{scarce}} = 6.0$ (corresponding to $|C_k| < N/e^{6.0} \approx 1401$ frames) are classified as scarce dynamic clusters (C_{dynamic}), and clusters with $S(C_k) \leq 6.0$ ($|C_k| \geq 1401$ frames) are classified as high-density static clusters (C_{static}).

Across all $N = 565{,}273$ frames in UTA-RLDD, feature extractors were evaluated over $K \in [5, 600]$ minimizing the Davies-Bouldin Index (DBI). DINOv2 (ViT-S/14, 384-D) achieved optimal clustering at $K = 200$ (DBI = 1.3260, Calinski-Harabasz CHI = 15,764.77), identifying 154 static high-density clusters (526,959 frames, 93.22%) versus 46 dynamic high-scarcity clusters (38,314 frames, 6.78%, cluster count ratio $\rho_{\text{clusters}} = 154/46 = 3.35{:}1$). This quantifies a frame imbalance ratio of $\rho_{\text{frames}} = 13.75{:}1$ (526,959/38,314) and an extreme cluster frequency ratio of $\rho_{\text{max / min}} = 103.94{:}1$ (9978 frames in Cluster 178 vs. 96 in Cluster 80). Alternative backbones confirmed this deficit under the same $S(C_k) > 6.0$ criterion: CLIP (ViT-B/32, $K = 100$, DBI = 1.3968, 97 static vs. 3 dynamic clusters, $\rho_{\text{frames}} = 232.68{:}1$) and ResNet-50 ($K = 100$, DBI = 1.4361, 99 static vs. 1 dynamic cluster, $\rho_{\text{frames}} = 980.38{:}1$).

These mathematically quantified visual deficiencies serve as explicit targets for the comparative augmentation stage. By leveraging both traditional data perturbations and generative visual effects to synthetically recreate the missing dynamic variances, this stage aims to rebalance ρ and expand the decision boundary. Finally, evaluating the resulting models on the unseen vehicular DriverMVT dataset [33] identifies which augmentation strategy best bridges the vehicular domain shift without corrupting motion features.

3.1.3. The Comparative Augmentation Pipeline (GenAI vs. Traditional)

To thoroughly assess the impact of various augmentation topologies on the cross-domain spatial generalization, there were three training corpora employed for comparative experiments. The first one, namely $\text{Dataset}_{\text{Base}}$, was the original set of image frames extracted from the UTA-RLDD training partition with no modifications applied. Thus, it represents the minimum viable context that can be explored within the given environment. The second set, $\text{Dataset}_{\text{Trad}}$, was the standard augmentation benchmark used in the study. It was generated by applying the standard set of spatial and photometric transformations, including shifting, flipping horizontally, applying Gaussian blur, and reducing the brightness [10,13]. Finally, the third corpus, namely $\text{Dataset}_{\text{GenAI}}$, was generated to incorporate the absent dynamic contexts identified in the unsupervised scarcity analysis (Section 3.1.2).

This enables a direct evaluation of whether high-quality synthetic visual effects can shape a more generalizable spatial manifold than standard augmentation algorithms [53]. Crucially, to ensure a controlled head-to-head evaluation between TradAug and GenAIAug, both augmentation pipelines were applied to the exact same scarce target participants identified in the scarcity analysis: Participant 26 (Drowsy state, Fold 3) and Participant 55 (Alert state, Fold 5). Each pipeline generated 4 variants of 3000 frames (198 sequences of 45 frames) per participant, adding +1584 augmented sequences (+6.29% volume) and expanding the dataset from 25,178 to 26,762 sequences in both corpora.

Restricting augmentation to these two highest-scarcity participants ($S(C_k) > 6.0$) was formulated as a controlled intervention to test whether targeted enrichment of rare head-pose and illumination tail cases improves Lab-to-Vehicle generalization without the computational overhead and artifact exposure of re-synthesizing all 60 subjects, while acknowledging that localizing synthetic sequences to two participants introduces fold-level

distribution asymmetry between Folds 1, 2, and 4 (where P26 and P55 appear in training) and Folds 3 and 5 (where P26 or P55 is held out for validation).

Synthesizing $\text{Dataset}_{\text{GenAI}}$ required careful semantic preservation of subtle behavioral cues, such as eye openness and instantaneous head pose, across the domain shift. To avoid the risks of facial hallucination inherent to end-to-end image translation, a deterministic three-layer compositing pipeline was employed in which the driver foregrounds originate strictly from unmodified raw video frames of the publicly available UTA-RLDD benchmark dataset [24] (Participant 26 in Fold 3 and Participant 55 in Fold 5). Specifically, in accordance with editorial transparency standards for generative models, four static empty vehicle cabin background plates (`plate_1.png`–`plate_4.png`, Layer 1) were synthesized using Stable Diffusion v1.5 (`runwayml/stable-diffusion-v1-5`) executed via Hugging Face `diffusers` (v0.25.0), `transformers` (v4.36.0), and `PyTorch` (v2.0.1) with the `DPMSolverMultistepScheduler` (30 inference steps, classifier-free guidance scale 7.5, deterministic seeds $\{42, 137, 256, 512, 1024\}$).

These background plates were domain-adapted via custom Low-Rank Adaptation with asymmetric learning rates (LoRA+, `peft` v0.7.1, rank $r = 64$, $\alpha = 64$, learning rate ratio $\lambda = \text{lr}_B/\text{lr}_A = 16.0$) and Timestep-dependent LoRA (T-LoRA) fine-tuned for 3000 steps on in-cabin frames using the training prompt "a photo of sks driver inside a car, low vigilance, realistic, 4k". The empty vehicle cabin background plates were generated using the verbatim positive text prompt "interior of a car, empty driver seat, steering wheel, dashboard camera perspective, daytime, highly detailed, photorealistic, no driver" and the verbatim negative prompt "person, human, face, body, hands, driver, ghost, silhouette, cartoon, illustration, distorted, blurry, low quality".

From these four plates, the corresponding foreground steering wheel overlay masks (`plate_1_driving_wheel.png`–`plate_4_driving_wheel.png`, Layer 3) were extracted manually using Pen Tool masking in the GNU Image Manipulation Program (GIMP) to define the exact object boundaries and mask out the surrounding vehicle interior. Real driver foregrounds (Layer 2) were then segmented from the UTA-RLDD laboratory footage using the `rembg` framework backed by the pretrained U^2-Net human segmentation neural network (`u2net_human_seg`) [55] with alpha matting enabled (foreground threshold 240, background threshold 10, erosion size 10 px), scaled and shifted into the driver seat (scale = 0.85), refined with 5 px dual horizontal edge alpha-fading and CIE-LAB color-space histogram matching (strength = 0.35), and composited between the background cabin plate (Layer 1) and the GIMP-masked steering wheel overlay (Layer 3).

Figure 2 qualitatively compares a representative frame across the three training corpora. $\text{Dataset}_{\text{Base}}$ (Figure 2A) preserves the original laboratory background, whereas $\text{Dataset}_{\text{Trad}}$ (Figure 2B) applies traditional spatial and photometric perturbations. $\text{Dataset}_{\text{GenAI}}$ (Figure 2C) is a three-layer composite of a synthetic vehicle cabin plate, a LAB-color-harmonized real UTA-RLDD driver foreground, and a GIMP-masked steering wheel overlay introducing vehicular cabin context while preserving authentic facial semantics.

3.1.4. Spatio-Temporal Model Architecture (Two-Stream Transformer)

The main problem during the use of an accurate DMS on limited-edge devices is the choice of model architecture. The selected architecture must meet specific requirements, such as computational efficiency and the dynamic nature of driver fatigue. A spatio-temporal DL architecture is selected over a static 2D-CNN because driver sleepiness is a dynamic temporal process. It is associated with the prolonged duration and the development of specific physiological signs over time that are characteristic of this state, for example, PERCLOS or the number of yawns [39]. Because a 2D-CNN processes isolated frames

independently, it cannot capture these multi-frame dynamics. As a result, there may be ambiguities in the classification of instantaneous video frames, for example, a change in the position of the driver's eyes when looking away from the road in the frame. In this regard, processing the movement of objects in real-time allows their instantaneous characteristics to be separated from long-term trends, such as a marker of fatigue.

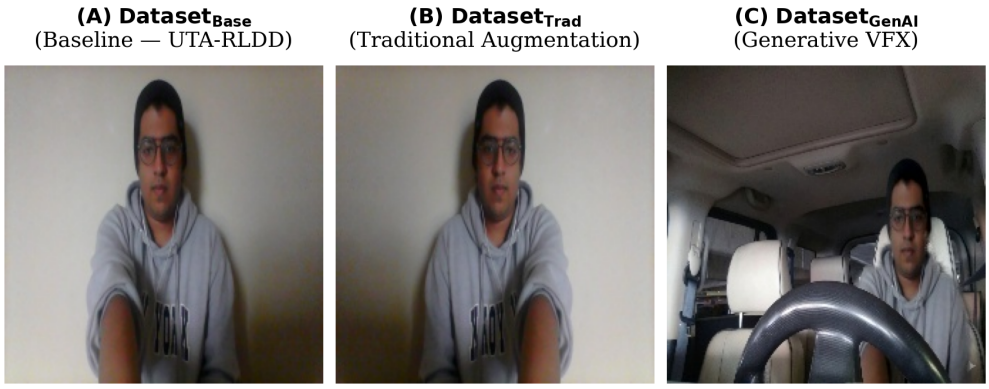

Figure 2. Qualitative comparison of a representative frame (Participant 55) across the three experimental training corpora: (**A**) Dataset$_{Base}$ (original laboratory frame), (**B**) Dataset$_{Trad}$ (spatial and photometric perturbations), and (**C**) Dataset$_{GenAI}$ (three-layer composite of a Stable Diffusion v1.5 + LoRA+ vehicle cabin plate, a U^2-Net matted driver foreground, and a GIMP Pen Tool steering-wheel overlay) [24]. The human driver face is an unmodified foreground frame from the public UTA-RLDD dataset [24].

To address these requirements, the resulting architecture, schematically presented in Figure 3, is designed as a specific Two-Stream Bottleneck-Transformer optimized in terms of trainable parameter efficiency and domain versatility. In particular, the model implements a clear separation between spatial feature extraction and temporal sequence modeling. To this end, the standard unfrozen CNNs used for the former task in most of the literature are substituted with a frozen, domain-aligned hybrid CNN-Transformer MobileViT-v2 spatial backbone coupled with an ultra-lightweight trainable temporal bottleneck, which avoids spatial backpropagation overhead during training and bounds the trainable parameter footprint. The spatial backbone (`mobilevitv2_200`, width multiplier 2.0) is initialized from `mosesb/drowsiness-detection-mobileViT-v2`, pretrained on ImageNet-1K and fine-tuned on >40,000 static facial drowsiness images, guaranteeing 0.0% subject or frame overlap with UTA-RLDD or DriverMVT.

On the other hand, the temporal stream exploits a transformer-based approach to effectively model the long-range temporal patterns of the previously extracted visual features. As a result, the network can track subtle fatigue micro-expressions accurately while keeping computational complexity low for real-time processing.

Aside from the two-stream design, the model enforces a strict inductive bias to suppress spatial shortcut learning, which is common in unfrozen, heavily parameterized networks. Namely, 100% of the spatial backbone's parameters (θ_{spatial}) are frozen during training. This complete parametric freeze prevents the network from updating its spatial weights to memorize static laboratory backgrounds or subject identities, which otherwise cause severe domain collapse during cross-dataset deployment. By freezing the spatial feature extractor, the model is forced to learn discriminative fatigue representations exclusively through the temporal bottleneck and attention layers, mitigating overfitting to environment-specific correlations.

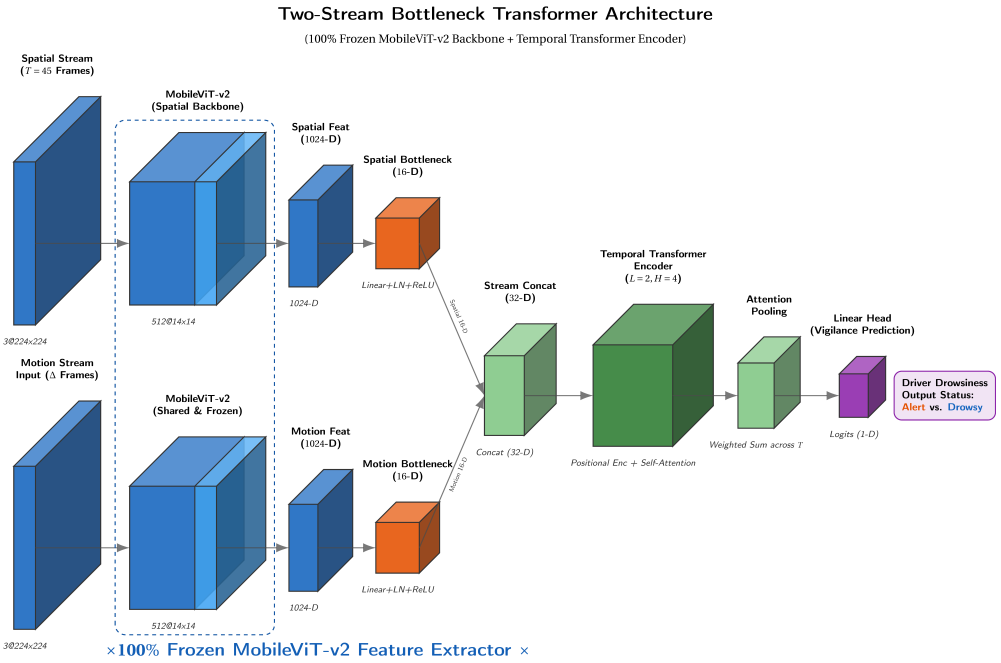

Figure 3. Detailed schematic of the proposed Two-Stream Bottleneck-Transformer architecture illustrating the 100% frozen spatial/motion backbones (99.76% of total parameters for MobileViT-v2) and the lightweight trainable temporal bottleneck and classification head (41,906 trainable parameters, 0.24%). Original schematic diagram created by the authors.

Enforcing 100% backbone freezing in the MobileViT-v2 configuration (17,466,235 total inference parameters, 66.87 MB FP32) leaves 17,424,329 parameters (99.76%) frozen within the weight-shared MobileViT-v2 spatial and motion feature extractor (`self.backbone`) and exposes only 41,906 trainable parameters (0.24%). Specifically, these 41,906 trainable parameters are distributed across the weight-shared 1024-D → 16-D linear bottleneck projection and LayerNorm (16,432 parameters; 9264 parameters for the 576-D → 16-D bottleneck in the 961,746-parameter MobileNetV3-Small baseline, which has 34,738 trainable parameters [3.61%]), the concatenated 32-D 2-layer 4-head temporal Transformer encoder (25,408 parameters; 12,704 per layer with $d_{\text{model}} = 32$ and $d_{\text{ff}} = 128$), the temporal attention pooling layer (33 parameters), and the single-logit linear classification head (33 parameters).

To preserve a compact trainable footprint suitable for edge platforms, the frozen spatial representations are routed through a trainable, low-dimensional Information Bottleneck ($d_{\text{model}} = 16$ per stream), which prunes spatial redundancies before projecting the compressed tokens to the unfrozen temporal transformer encoder. Such a design fulfills two important roles. First, it achieves an extremely compact footprint in terms of trainable parameters and intermediate token dimensions, protecting the pipeline from memory overflows on resource-constrained devices [20,21,56]. Secondly, it isolates the temporal sequence modeling from a computational perspective. The fused spatio-temporal output tensor is aggregated via temporal attention pooling (33 parameters) and processed by a linear classification layer (33 parameters) followed by a sigmoid activation function, yielding the final classification probability over the defined driver vigilance states.

3.2. Phase 2: Edge-Aware Model Compression

To bridge the gap between high-capacity models and the strict power, memory, and latency budgets of resource-constrained ADAS edge environments [20], Phase 2 applies hardware-aware numerical truncation to the trained topologies from Phase 1. This stage evaluates the conversion of the full-precision spatio-temporal network into a compressed,

real-time-capable counterpart and identifies which augmented feature topology best withstands quantization.

3.2.1. The Automotive Edge Environment and Rationale for Quantization

Deploying high-capacity DL architectures introduces computational bottlenecks at the automotive edge. Modern ADAS operate under strict multi-objective constraints, where intelligent subsystems must share limited power budgets, memory bandwidths, and thermal capacities [20,21]. Deploying conventional, uncompressed neural networks in these environments is computationally impractical, as millions of floating-point matrix multiplications per forward pass can exceed real-time driver monitoring latency thresholds [56].

To address these constraints without redesigning the spatio-temporal architecture from scratch, parameter quantization emerges as a practical solution. By reducing the numerical precision of network weights and activations from 32-bit floating-point (FP32) to 8-bit integers (INT8) [28,29], the memory footprint of quantized layers can be theoretically reduced by up to 4×, directly alleviating memory bandwidth bottlenecks [29]. Meanwhile, by mapping continuous activations to discrete integer registers, floating-point multiply-accumulate (MAC) operations in quantized blocks can be largely replaced by efficient integer-based MAC operations during inference [21].

The existing literature mainly focuses on the destructive nature of INT8 quantization as a means of achieving hardware-level optimizations. This phase, however, will primarily focus on the interaction between the aforementioned operations and the spatial topologies synthesized in Phase 1. It is hypothesized that aggressive numerical approximations may act as an implicit regularizer, screening spatially-localized fragile patterns. The present goal, therefore, is to evaluate whether traditional perturbations or VFX compositing yield representations resilient to integer mapping.

3.2.2. Isolation of Phase 1 Champion Models for Comparative Compression

From each of the three training procedures carried out in Phase 1 ($\text{Dataset}_{\text{Base}}$, $\text{Dataset}_{\text{Trad}}$, and $\text{Dataset}_{\text{GenAI}}$), the best-performing models were selected as the "champion models." Specifically, across all five subject-independent folds, per-fold checkpoints and single-fold champion models (`No-Aug Fold 4`, `TradAug Fold 1`, and `GenAIAug Fold 1`) were selected strictly by maximum internal validation Macro F_1-score on the held-out UTA-RLDD validation split of each respective fold, with zero exposure to the out-of-distribution DriverMVT test set (while `TradAug Fold 2` is additionally reported in Section 4.5 as a post-hoc reference representing the highest cross-domain FP32 fold).

The subsequent application of PTQ to these three topologies is motivated by an analysis of latent space stability. Model quantization inherently introduces rounding and clipping errors when mapping continuous, high-precision floating-point representations to constrained discrete integers [29]. By subjecting each model variant to identical hardware-constrained INT8 compression, this study objectively determines which learned feature topology is most resilient to numerical truncation.

In essence, this isolation strategy is aimed to verify the hypothesis that the choice of data augmentation strategies is decisive for the geometry of the learned feature manifold—in other words, whether the conventional algorithmic perturbations or image synthesis procedures would be sufficient to encode a latent spatial representation, which would survive the extreme precision loss required for the real-world automotive edge deployment. Furthermore, to verify that post-quantization behavior is not an artifact of a single validation split, PTQ is additionally evaluated across all five cross-validation folds (both per-fold mean ± SD and 5-fold INT8 ensembles) on the $N = 66$ and expanded $N = 270$ DriverMVT benchmarks.

3.2.3. Post-Training Quantization Protocol: FP32 to INT8 Mapping

The procedure used to map the network's high-precision parameters into a hardware-friendly representation is based on a static PTQ pipeline. This step is essential to achieve the inference throughput and memory footprint needed for real-time execution on downstream embedded ADAS platforms [28]. The decision to adopt PTQ rather than Quantization-Aware Training (QAT) is motivated by the fact that PTQ enables direct post-hoc deployment of trained checkpoints without requiring time- and memory-consuming retraining [30].

Procedurally, static PTQ converts the network's continuous FP32 weights and activations into a discrete INT8 space. Mathematically, this transformation is governed by the standard asymmetric quantization function:

$$x_{\text{INT8}} = \text{clip}\left(\text{round}\left(\frac{x_{\text{FP32}}}{S}\right) + Z, -128, 127\right) \tag{3}$$

where x_{FP32} is the real floating-point tensor, S is the computed step-size scaling factor, Z is the integer zero-point, and the boundaries $[-128, 127]$ represent the representable dynamic range of the signed 8-bit integer space [21,28]. To derive optimal values for S and Z and mitigate clipping errors, the protocol utilizes a targeted calibration dataset to precisely profile the minimum and maximum activation boundaries during a series of forward passes [29]. In the proposed PyTorch FX Graph Mode static PTQ pipeline (`prepare_fx` and `convert_fx` using the `fbgemm` engine, Figure 4), calibration is executed over 100 forward batches sampled strictly from the UTA-RLDD training split of each respective fold across three independent random seeds $(42, 100, 2026)$. Crucially, zero DriverMVT test data is exposed during calibration, guaranteeing zero data leakage.

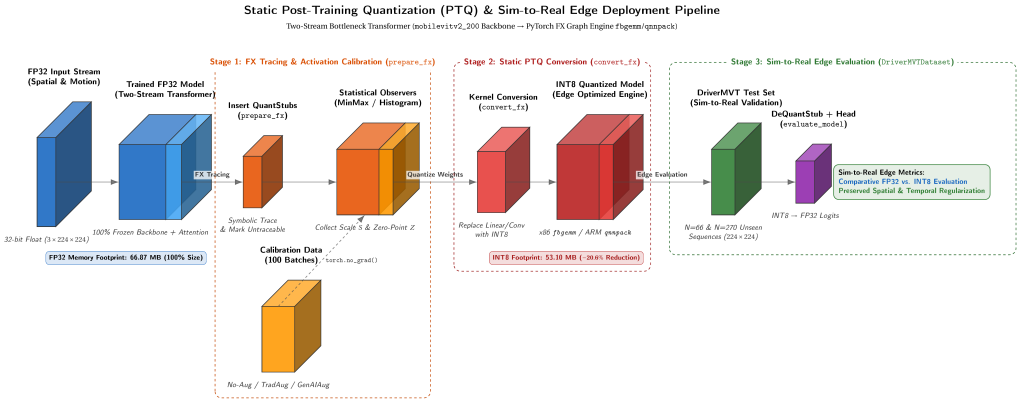

Figure 4. Selective PyTorch FX Graph PTQ pipeline demonstrating the hybrid precision methodology. Early MobileViT-v2 convolutional stages and the bottleneck/classification head are quantized to INT8 (53.10 MB, 20.6% compression from 66.87 MB FP32), while `MobileViTv2Block` linear self-attention modules and the temporal Transformer encoder are preserved in FP32 to prevent motion stream degradation. Original schematic diagram created by the authors.

Crucially, standard PTQ uniformly quantizes the entire network architecture, which in many cases results in severe information decay within highly sensitive sequential models. A strict selective quantization strategy in the PyTorch FX Graph Mode is proposed to circumvent these adverse effects. As illustrated in Figure 4, this partial quantization strategy maps the early frozen MobileViT-v2 convolutional stages and linear bottleneck/head projections to INT8 to obtain considerable memory reductions, whereas the `MobileViTv2Block` linear self-attention stages and the temporal Multi-Head Self-Attention (MHSA) layers of the Transformer encoder are explicitly isolated in full-precision FP32 to safeguard the sequential fatigue dynamics.

This hybrid-precision design maximizes memory compression while isolating sensitive temporal fatigue representations from signal decay. Specifically, quantizing early convo-

lutional/inverted-residual blocks (`stem`, `stages.0`, `stages.1`, `2.0`, `3.0`, `4.0`, `final_conv`) plus the bottleneck/head compresses MobileViT-v2 from 66.87 MB to 53.10 MB (20.6% reduction), whereas full-backbone convolution quantization (17.83 MB, 73.3% reduction) degrades linear self-attention motion representations and collapses all checkpoints across all 5 folds to 50.00% chance accuracy (0.3333 Macro F_1) on the balanced $N = 66$ Driver-MVT benchmark. Crucially, Selective FX INT8 PTQ is designed for memory compression and integer CPU execution rather than as a universal accuracy-enhancing operation; any accuracy gain observed under uncalibrated cross-domain deployment ($\tau = 0.50$) is a regime-dependent operating-point calibration effect requiring strong pre-existing FP32 rank separability paired with perturbation-hardened convolutional weights.

3.2.4. Trade-Off Analysis: Memory Compression Versus Accuracy Retention

The deployability of an ADAS perception module depends on its ability to execute within constrained hardware budgets without incurring severe predictive degradation. Automotive real-time processing requires intelligent subsystems to operate within strict memory bandwidth limits while sustaining high classification fidelity [20]. To quantify the hardware–accuracy trade-offs introduced by PTQ, the compressed INT8 topologies are systematically evaluated against their full-precision FP32 baselines using a dual-factor analytical framework.

First, memory footprint reduction is characterized through the absolute and relative decreases in physical checkpoint file size (in Megabytes) and layer bit-width precision, quantifying how Selective FX INT8 mapping alleviates storage bottlenecks.

Second, predictive retention is quantified by computing the performance delta (Δ) between the original FP32 accuracy and macro F_1-score and the quantized INT8 outputs. Because driver fatigue monitoring involves class-specific safety risks, the harmonic mean of precision and recall (F_1-score) alongside class-specific sensitivity (TPR) and specificity (TNR) must be tracked to ensure that the model does not suffer severe degradation on either vigilance state after quantization. Evaluating Δ across the three training regimes ($\text{Dataset}_{\text{Base}}$, $\text{Dataset}_{\text{Trad}}$, and $\text{Dataset}_{\text{GenAI}}$) directly addresses the central research question: whether traditional perturbations or generative VFX synthesis yields a feature manifold more resilient to quantization noise in edge ADAS.

While this study establishes PTQ's structural viability via physical memory footprint compression and cross-domain F_1-score evaluations, to evaluate operational runtime feasibility, streaming latency, throughput (FPS), and peak Resident Set Size (RSS) memory are profiled on an x86 CPU edge-simulation benchmark using the `fbgemm` engine across 1-thread and 4-thread execution. Meanwhile, direct on-target profiling of active power draw and thermal dissipation on physical automotive Electronic Control Units (ECUs) or NVIDIA Jetson modules will be addressed in future work.

4. Experiments and Results

4.1. Experimental Setup and Implementation Details

In order to thoroughly assess the proposed framework under resource constraints representative of edge ADAS development, all experiments were implemented using the PyTorch framework and executed on an NVIDIA GeForce RTX 3060 Laptop GPU with an allocated 6 GB of Video RAM (VRAM). The network was trained using the AdamW optimizer with a learning rate of 5×10^{-5} and a weight decay of 0.01. To maximize memory efficiency while maintaining training throughput, Automatic Mixed Precision (AMP) was employed with a gradient accumulation step of 4, resulting in an effective batch size of 16.

Crucially, to prevent spatial shortcut memorization, the high-dimensional spatial backbone (MobileViT-v2) is completely frozen during training. By freezing 100% of the

backbone parameters, the network is forced to optimize exclusively the low-dimensional information bottleneck and subsequent temporal transformer layers. This enables the model to learn generalizable motion representations and fatigue dynamics rather than static identity features.

Furthermore, to ensure strict numerical reproducibility and preserve the integrity of the temporal motion signals, deterministic execution protocols were enforced throughout all experiments. This included fixed CUBLAS workspace configurations and the explicit disabling of TensorFloat-32 (TF32) matrix multiplications, removing non-deterministic algorithmic behavior across training iterations. Specifically, PyTorch 2.x deterministic execution was enforced via `torch.use_deterministic_algorithms(True)` and `torch.backends.cudnn.deterministic = True`, while TF32 tensor core acceleration was explicitly disabled (`torch.backends.cuda.matmul.allow_tf32 = False` and `torch.backends.cudnn.allow_tf32 = False`) to ensure bitwise reproducibility across folds.

4.2. Zero-Leakage Cross-Validation Protocol

A pervasive issue in driver vigilance research is the "data leakage illusion." Several recent studies benchmarking on the UTA-RLDD dataset report inflated frame-level classification accuracies exceeding 95% [13,25]. These inflated figures stem from the use of randomized train/test splits at the frame level rather than the subject level, inducing severe data leakage and memory effects. Consequently, as demonstrated by Ghoddoosian et al. [24] and Magán et al. [18], the network memorizes subject facial identity, clothing, and static laboratory backgrounds instead of learning a generalizable representation of underlying fatigue dynamics [18,24,26].

To counteract this methodological bias, a strict, video-level, subject-independent 5-fold cross-validation protocol [18,24] is enforced by the evaluation framework. Specifically, the 60 participants in UTA-RLDD were partitioned into 5 subject-independent folds of exactly 12 participants per fold, guaranteeing 0.0% subject, video, or frame overlap between training and validation sets across all folds and ensuring methodologically sound SOTA comparisons for unseen drivers.

4.3. Internal Evaluation on UTA-RLDD (SOTA Comparison)

The internal 5-fold subject-independent cross-validation on the UTA-RLDD dataset assesses the performance of the frozen MobileViT-v2 spatial backbone across the three designated pipelines. According to Table 2, based on the aggregate Out-Of-Fold (OOF) predictions, the non-augmented baseline (No-Aug) attained the highest internal performance, reaching an OOF accuracy of 59.96% (per-fold mean 60.00 $\pm$ 5.59%) and an OOF macro F_1-score of 0.6374 (per-fold mean 0.6357). Traditional geometric and photometric augmentations (TradAug) achieved a per-fold mean validation accuracy of 58.01 $\pm$ 6.62% with a mean macro F_1-score of 0.6099. Meanwhile, the generative AI augmentation pipeline (GenAIAug) attained a per-fold mean accuracy of 56.13 $\pm$ 8.58% and a mean macro F_1-score of 0.5978.

These results must be put into perspective of the strict, zero-leakage baselines in the literature, specifically the 65.2% accuracy reported by Ghoddoosian et al. [24], and the 63.0% accuracy achieved by Magán et al. [18]. When benchmarked against these results, the internal OOF accuracy of 59.96% (0.6374 Macro F_1) should not be viewed as a performance shortcoming, but rather as a significant architectural advancement, since prior zero-leakage methodologies rely on multi-stage, non-differentiable preprocessing pipelines such as manual facial landmark extraction via `dlib` or hand-crafted fuzzy inference systems.

Table 2. Internal 5-Fold Subject-Independent Cross-Validation Benchmark Comparison on UTA-RLDD (N = 25,178 Base Sequences) against literature baselines and controlled freezing variants. Here and in the subsequent cross-domain evaluations, 5-fold cross-validation dispersion (±SD) denotes population standard deviation (ddof = 0) across the five folds.

Architecture & Pipeline	Backbone Freezing	5-Fold Mean Acc (%) [OOF]	Macro F_1-Score	Params (Trainable/Total)
Magán et al. [18] (EfficientNetB0_GRU)	Partially Unfrozen	51.09 ± 4.29	0.4207	1,197,697/5,205,245 (23.01%)
Single-Stream BottleneckTransformer (RGB-Only)	100% Frozen	63.19 ± 4.96	0.5659	15,841/942,849 (1.68%)
Single-Stream BottleneckTCN (Dilated TCN, RGB-Only)	100% Frozen	55.92 ± 3.75	0.5782	15,681/942,689 (1.66%)
Two-Stream MobileNetV3-Small (Frozen Baseline)	100% Frozen	60.03 ± 7.10	0.6490	34,738/961,746 (3.61%)
Two-Stream MobileNetV3-Small (TradAug)	Partially Unfrozen	56.87 ± 8.86	0.6315	385,282/961,746 (40.06%)
Two-Stream MobileNetV3-Small (TradAug)	100% Unfrozen	53.26 ± 5.94	0.4888	961,746/961,746 (100.0%)
Proposed Two-Stream MobileViT-v2 (No-Aug Baseline)	**100% Frozen**	**60.00 ± 5.59 [59.96]**	**0.6357 [0.6374]**	**41,906/17,466,235 (0.24%)**
Proposed Two-Stream MobileViT-v2 (TradAug)	**100% Frozen**	**58.01 ± 6.62**	**0.6099**	**41,906/17,466,235 (0.24%)**
Proposed Two-Stream MobileViT-v2 (GenAIAug)	100% Frozen	56.13 ± 8.58	0.5978	41,906/17,466,235 (0.24%)
Two-Stream MobileViT-v2 (TradAug)	Partially Unfrozen	49.16 ± 4.29	0.3900	9,401,269/17,466,235 (53.83%)
Two-Stream MobileViT-v2 (TradAug)	100% Unfrozen	58.32 ± 5.30	0.5630	17,466,235/17,466,235 (100.0%)

Note: Bold rows highlight the primary proposed 100% frozen Two-Stream MobileViT-v2 configurations.

The proposed framework, by contrast, achieves competitive zero-leakage predictive accuracy using an end-to-end differentiable architecture updating only its lightweight 41,906-parameter temporal bottleneck and classification head (0.24% of total parameters); hence, these results demonstrate that high-level spatio-temporal fatigue dynamics can be effectively extracted and classified with a compact temporal head tailored for edge deployment. Furthermore, when evaluated directly on identical 45-frame (5 FPS) subject-independent splits in Table 2, the proposed Two-Stream Frozen MobileViT-v2 outperforms the reproduced Magán et al. [18] EfficientNetB0_GRU baseline by +21.67% in macro F_1 (0.6374 vs. 0.4207) and outperforms both single-stream Transformer (0.5659) and TCN (0.5782) baselines, while unfreezing 100% of either MobileNetV3-Small or MobileViT-v2 rapidly overfits training subjects within 2 to 4 epochs (Train Loss → 0.014) and degrades internal macro F_1 to 0.4888 and 0.5630, respectively.

4.4. Cross-Domain Generalization: Mitigating the Spatial Domain Shift

The critical evaluation of cross-domain generalization was performed on the unseen real-world vehicular dataset DriverMVT to explicitly expose the Lab-to-Vehicle domain shift. To mitigate small-sample bias and evaluate across all five drivers (P01–P05), testing was conducted on both the historical $N = 66$ balanced benchmark (33 Drowsy/33 Alert sequences) and the expanded $N = 270$ benchmark (60 Drowsy/210 Alert sequences across all five drivers). The initial baseline experiments with a partially unfrozen MobileNetV3-Small spatial encoder (Table 3, Part A) demonstrated a severe performance drop across cross-domain settings. In particular, during 5-fold ensemble validation on the real-world driving logs, the partially unfrozen MobileNetV3-Small setup predicts zero True Negatives (TN = 0). This translates to a specificity collapse of 0.00% and an accuracy of 43.94% with a macro F_1-score of 0.3053. This failure mode indicates that the partially unfrozen encoder memorized the training distribution, including the homogeneous laboratory background.

As demonstrated in the 2 × 3 controlled freezing matrix (Table 3, Part A), partially unfrozen backbones (MobileNetV3-Small with 385,282 trainable parameters and MobileViT-v2 with 9,401,269 trainable parameters) both collapse to 0.00% Specificity (TN = 0, 100% FPR) on $N = 66$ and ≤2.86% Specificity on $N = 270$. Conversely, 100% fully unfrozen backbones suffer severe sensitivity collapse, achieving only 33.33% TPR for MobileNetV3-Small (100% trainable) and 18.18% TPR for MobileViT-v2 (100% trainable; ROC-AUC = 0.3494 on $N = 270$), missing up to 81.8% of active drowsy episodes.

By contrast, the proposed approach with a fully frozen, domain-aligned MobileViT-v2 spatial encoder attenuates shortcut learning from the background (Table 3, Part A and Part B). Specifically, the frozen TradAug configuration achieves a cross-domain accuracy of 71.21% [95% CI: 60.61–81.82%] and a 0.6917 macro F_1-score [95% CI: 0.5696–0.8062], recovering 15 True Negatives (TN = 15, Specificity 45.45% [95% CI: 27.27–63.64%]). Evaluated via 10,000 stratified bootstrap resamples, this 100% frozen MobileViT-v2 FP32 Ensemble achieves a high ROC-AUC of 0.8838 [95% CI: 0.7906–0.9596] and 96.97% Sensitivity [95% CI: 90.91–100.00%] (TP = 32/33) on $N = 66$, while the single-fold TradAug Fold 2 FP32 checkpoint achieves 84.85% accuracy [95% CI: 75.76–92.42%] (84.85% TNR [95% CI: 72.73–96.97%], 84.85% TPR [95% CI: 72.73–96.97%], 0.8485 Macro F_1 [95% CI: 0.7556–0.9242], 0.9036 ROC-AUC [95% CI: 0.8182–0.9697]) on $N = 66$ and improves from 34.44% uncalibrated FP32 accuracy to 80.37% under Selective FX INT8 on $N = 270$.

Although uncalibrated FP32 inference ($\tau = 0.50$) exhibits an upward cross-domain logit shift across Folds 1, 3, 4, and 5 that lowers raw uncalibrated specificity on the Alert-dominated $N = 270$ benchmark (210 Alert vs. 60 Drowsy sequences; 24.07% TradAug and 24.81% No-Aug FP32 Ensemble accuracy), calibrating the decision threshold on UTA-RLDD ($\tau^* = 0.99$) elevates $N = 66$ accuracy to 80.30% (0.8027 Macro F_1, 78.79% TPR, 81.82% TNR), and post-hoc Youden's J^* ($\tau = 0.9996$) reaches 87.88% accuracy (0.8788 Macro F_1, 87.88% TPR, 87.88% TNR, 12.12% FPR, TN = 29, FP = 4).

Crucially, Selective FX INT8 quantization automatically recenters this cross-domain logit bias without threshold tuning to achieve 74.24% Ensemble accuracy [95% CI: 63.64–84.85%] (0.7395 Macro F_1 [95% CI: 0.6251–0.8447], 63.64% TPR [95% CI: 45.45–78.79%], 84.85% TNR [95% CI: 72.73–96.97%], 0.7824 ROC-AUC [95% CI: 0.6657–0.8861]) on $N = 66$ and 76.30% Ensemble accuracy [95% CI: 71.11–81.11%] (0.6885 Macro F_1 [95% CI: 0.6226–0.7505], 61.67% TPR [95% CI: 48.33–73.33%], 80.48% TNR [95% CI: 75.24–85.71%], 0.7608 ROC-AUC [95% CI: 0.6961–0.8221]; per-fold mean 62.74 ± 21.14%) on $N = 270$, though the remaining 15.15–19.52% window-level FPR (84.85% and 80.48% Specificity) indicates that downstream multi-window temporal hysteresis is needed before triggering driver-warning alerts in production vehicles.

Furthermore, controlled stream-isolation ablations (Table 3, Part B) confirm that masking the Motion Stream (Spatial-Only) drops specificity from 45.45% to 15.15% (TN = 5, 57.58% Ensemble accuracy), whereas isolating the Motion Stream (Motion-Only, where $\Delta I_{\text{bg}} \approx 0$) yields 69.70% specificity (TN = 23), confirming that motion cues suppress false alarms. Finally, the cross-domain analysis unveils a critical insight that naive synthetic diversity strategies have limited practical value without temporal consistency: the complex GenAIAug variant (60.61% accuracy, 0.5568 F_1-score, TN = 9) severely underperforms against the No-Aug (68.18% accuracy) and TradAug (71.21% accuracy) baselines. Therefore, expensive VFX-style synthetic augmentations are outperformed by architectural inductive bias: strictly locking a domain-aligned spatial backbone limits the Lab-to-Vehicle spatial domain gap in a native, reliable, and far more computationally efficient manner than computationally expensive generative pipelines for edge-ADAS vigilance monitoring.

4.5. Post-Training Quantization and Edge Deployment Evaluation

To evaluate the viability of edge deployment for automotive Electronic Control Units via CPU edge-simulation benchmarks, static PTQ was applied to the trained network architectures. Using PyTorch FX Graph Mode and the `fbgemm` engine, the continuous FP32 representations in the early convolutional stages and bottleneck/head projections were converted into discrete INT8 registers. In this manner, the optimized spatio-temporal fatigue features were evaluated to determine whether they can withstand numerical truncation and satisfy the resource constraints of embedded automotive hardware (Table 4).

Table 3. Cross-Domain Generalization on DriverMVT Dataset. Controlled 2 × 3 backbone × freezing factorial ablation, augmentation comparison, and stream-isolation matrix on the $N = 66$ and expanded $N = 270$ benchmarks.

Architecture & Pipeline	Freezing/ Mode	Trainable/ Total Params	Historical Benchmark ($N = 66$)				Expanded 5-Driver Benchmark ($N = 270$)			
			Mean Acc (%)	Ens Acc (%) [Macro F_1]	TPR/ TNR (%)	ROC-AUC	Mean Acc (%)	Ens Acc (%) [F_1]	TPR / TNR (%)	ROC-AUC
Part A: Controlled 2 × 3 Backbone × Freezing Strategy Ablation (UTA-RLDD 5-Fold Models, FP32, $\tau = 0.50$)										
MobileNetV3-Small (Frozen Baseline)	100% Frozen	34,738/ 961,746 (3.61%)	56.97 ± 8.22	46.97 [0.4545]	63.64/ 30.30	0.5303	34.30 ± 7.60	28.52 [0.2852]	65.00/18.10	0.4790
MobileNetV3-Small (TradAug)	Partially Unfrozen	385,282/ 961,746 (40.06%)	51.21 ± 8.04	43.94 [0.3053]	87.88/**0.00**	0.5794	37.26 ± 8.49	21.48 [0.1913]	86.67/**2.86**	0.6206
MobileNetV3-Small (TradAug)	100% Unfrozen	961,746/ 961,746 (100.0%)	51.52 ± 6.98	56.06 [0.5367]	**33.33**/78.79	0.5702	54.52 ± 20.23	58.52 [0.4872]	**33.33**/65.71	0.5144
MobileViT-v2 (TradAug, Proposed)	**100% Frozen**	**41,906/ 17,466,235 (0.24%)**	**68.48 ± 14.45**	**71.21 [0.6917]**	**96.97/45.45**	**0.8838**	27.19 ± 4.56	24.07 [0.2170]	93.33/4.29	0.5237
MobileViT-v2 (TradAug)	Partially Unfrozen	9,401,269/ 17,466,235 (53.83%)	56.06 ± 7.67	50.00 [0.3333]	100.00/**0.00**	0.6786	41.19 ± 22.54	22.96 [0.1924]	100.00/**0.95**	0.7481
MobileViT-v2 (TradAug)	100% Unfrozen	17,466,235/ 17,466,235 (100.0%)	57.27 ± 6.53	59.09 [0.5087]	**18.18**/100.00	0.7181	50.07 ± 17.52	52.22 [0.4118]	**20.00**/61.43	0.3494
Part B: Augmentation Pipelines, Stream-Isolation, and Threshold/INT8 Recentering (100% Frozen MobileViT-v2)										
MobileViT-v2 (No-Aug Baseline)	100% Frozen (FP32)	41,906/ 17,466,235 (0.24%)	67.27 ± 10.17	68.18 [0.6592]	93.94/42.42	0.8760	26.15 ± 3.00	24.81 [0.2267]	93.33/5.24	0.5935
MobileViT-v2 (GenAIAug VFX)	100% Frozen (FP32)	41,906/ 17,466,235 (0.24%)	64.55 ± 7.82	60.61 [0.5568]	93.94/27.27	0.8861	29.11 ± 5.90	22.96 [0.2023]	93.33/2.86	0.6721
MobileViT-v2 (TradAug Spatial-Only)	Stream 2 Masked	41,906/ 17,466,235 (0.24%)	61.82 ± 11.03	57.58 [0.4826]	100.00/15.15	0.8838	—	—	—	—
MobileViT-v2 (TradAug Motion-Only)	Stream 1 Masked	41,906/ 17,466,235 (0.24%)	50.61 ± 4.85	39.39 [0.3327]	9.09/69.70	0.5298	—	—	—	—
MobileViT-v2 (TradAug J^* [$N = 66$]/τ^* [$N = 270$])	**Calibrated FP32**	**41,906/ 17,466,235 (0.24%)**	—	**87.88 [0.8788]**	**87.88/87.88**	**0.8838**	—	**28.52 [0.2780]**	**86.67/11.90**	**0.5237**
MobileViT-v2 (TradAug Selective INT8)	**100% Frozen (INT8)**	**41,906/ 17,466,235 (53.1 MB)**	**60.61 ± 8.78**	**74.24 [0.7395]**	**63.64/84.85**	**0.7824**	**62.74 ± 21.14**	**76.30 [0.6885]**	**61.67/80.48**	**0.7608**

Note: Italic headings denote sub-table sections (Parts A and B); bold rows highlight the proposed framework configurations; bold standalone values highlight specificity or sensitivity collapse modes under unfrozen backbones.

The empirical comparison performed on this cross-domain edge classification benchmark highlighted the difference in model robustness to the quantization-induced distribution shifts. According to Table 4, the TradAug champion checkpoint (Fold 1) demonstrated a positive accuracy gain (+16.67%) that ultimately improved the baseline from 53.03% to 69.70% after quantization (with an accompanying increase in the F_1-score from 0.3973 to 0.6967). This result suggests that conventional training-time augmentations act as an implicit denoiser, shaping weight distributions into quantization-friendly topologies.

Mechanistically, evaluating 5-fold paired checkpoints reveals that Selective FX INT8 PTQ (53.10 MB) does not synthesize new discriminative features or improve underlying rank-order separability on $N = 66$—where 8-bit discretization lowers ROC-AUC from 0.9343 to 0.6850 on Fold 1 and from 0.8838 to 0.7824 on the 5-fold ensemble (and where threshold-calibrated FP32 Youden's J^* still achieves the highest balanced accuracy at 87.88%). Instead, quantization attenuates positive logit bias by inducing a negative shift ($\Delta z_0 = -10.878$ on Alert logits and $\Delta z_1 = -7.880$ on Drowsy logits) through L2 convolutional activation damping in perturbation-hardened TradAug layers.

Table 4. Quantitative Edge Evaluation Matrix of Post-Training Static Quantization INT8 on the DriverMVT Test Set ($N = 66$ and Expanded $N = 270$ Benchmarks), including 5-fold paired statistics, PTQ granularity ablations, and 4-thread/1-thread streaming CPU latency (`fbgemm` backend). Part A reports single-fold checkpoints, whereas Part B reports 5-fold ensembles across all five drivers on $N = 270$.

Model Variant & Mode	PTQ Granularity	Size (MB) FP32 → INT8	DriverMVT ($N = 66$ Benchmark)			DriverMVT ($N = 270$ Benchmark)		Streaming CPU Latency 4T|1T ms (FPS)
			FP32 → INT8 Acc (%) (Δ|Ens)	FP32 → INT8 Macro F_1	INT8 TNR/FPR (%)	FP32 → INT8 Acc (%) [F_1]	INT8 TNR/ROC-AUC	
Part A: Single-Fold Champion Checkpoints (Paired FP32 vs. Selective FX INT8 Evaluation)								
No-Aug Baseline (Fold 4)	Selective FX INT8	66.87 → 53.10 (20.6%)	72.73 → 50.00 (−22.73)	0.7144 → 0.3333	0.00/100.00	—	—	200.42 (4.99)|430.65 (2.32)
TradAug (Fold 1 Champion)	**Selective FX INT8**	**66.87 → 53.10 (20.6%)**	**53.03 → 69.70 (+16.67)**	**0.3973 → 0.6967**	**66.67/33.33**	**24.07 → 68.52**	**—**	**200.42 (4.99)|430.65 (2.32)**
TradAug (Fold 2 FP32 Best)	Selective FX INT8	66.87 → 53.10 (20.6%)	84.85 → 68.18 (−16.67)	0.8485 → 0.6460	100.00/0.00	34.44 → 80.37	—	200.42 (4.99)|430.65 (2.32)
GenAIAug (Fold 1)	Selective FX INT8	66.87 → 53.10 (20.6%)	69.70 → 50.00 (−19.70)	0.6780 → 0.3333	0.00/100.00	—	—	200.42 (4.99)|430.65 (2.32)
Part B: 5-Fold Paired Mean ± SD, 5-Fold INT8 Ensembles, and Quantization Granularity Ablation								
No-Aug (5-Fold)	Head-Only INT8	66.87 → 66.84 (0.04%)	67.27 ± 10.17 → 67.58 ± 10.26 (Ens: 68.18 → 68.18)	0.6229 → 0.6240 (0.6592)	42.42/57.58	—	—	210.43 (4.75)|478.22 (2.09)
No-Aug (5-Fold)	Selective FX INT8	66.87 → 53.10 (20.6%)	67.27 ± 10.17 → 56.67 ± 9.47 (Ens: 68.18 → 48.48)	0.6229 → 0.4641 (0.3265)	0.00/100.00	—	—	200.42 (4.99)|430.65 (2.32)
No-Aug (5-Fold)	Full-Backbone INT8	66.87 → 17.83 (73.3%)	67.27 ± 10.17 → 50.00 ± 0.00 (Ens: 68.18 → 50.00)	0.6229 → 0.3333 (0.3333)	0.00/100.00	—	—	342.61 (2.92)|886.27 (1.13)
TradAug (5-Fold)	Head-Only INT8	66.87 → 66.84 (0.04%)	68.48 ± 14.45 → 68.48 ± 14.45 (Ens: 71.21 → 72.73)	0.6222 → 0.6222 (0.7102)	48.48/51.52	—	—	210.43 (4.75)|478.22 (2.09)
TradAug (5-Fold Proposed)	**Selective FX INT8**	**66.87 → 53.10 (20.6%)**	**68.48 ± 14.45 → 60.61 ± 8.78 (Ens: 71.21 → 74.24)**	**0.6222 → 0.5272 (0.6917 → 0.7395)**	**84.85/15.15**	**24.07 [0.2170] → 76.30 [0.6885]**	**80.48%/0.7608**	**200.42 (4.99)|430.65 (2.32)**
TradAug (5-Fold)	Full-Backbone INT8	66.87 → 17.83 (73.3%)	68.48 ± 14.45 → 50.00 ± 0.00 (Ens: 71.21 → 50.00)	0.6222 → 0.3333 (0.3333)	0.00/100.00	—	—	342.61 (2.92)|886.27 (1.13)
GenAIAug (5-Fold)	Head-Only INT8	66.87 → 66.84 (0.04%)	64.55 ± 7.82 → 64.55 ± 7.82 (Ens: 60.61 → 60.61)	0.5930 → 0.5930 (0.5568)	27.27/72.73	—	—	210.43 (4.75)|478.22 (2.09)
GenAIAug (5-Fold)	Selective FX INT8	66.87 → 53.10 (20.6%)	64.55 ± 7.82 → 51.82 ± 3.64 (Ens: 60.61 → 50.00)	0.5930 → 0.3775 (0.3333)	0.00/100.00	—	—	200.42 (4.99)|430.65 (2.32)
GenAIAug (5-Fold)	Full-Backbone INT8	66.87 → 17.83 (73.3%)	64.55 ± 7.82 → 50.00 ± 0.00 (Ens: 60.61 → 50.00)	0.5930 → 0.3333 (0.3333)	0.00/100.00	—	—	342.61 (2.92)|886.27 (1.13)
MobileViT-v2 (5-Fold TradAug)	FP32 Baseline	66.87 MB (0.0%)	68.48 ± 14.45 (Ens: 71.21)	0.6222 (0.6917)	45.45/54.55	24.07 [0.2170]	4.29%/0.5237	211.19 (4.74)|529.03 (1.89)
MobileNetV3-Small (5-Fold)	FP32 Baseline	3.81 MB (0.0%)	56.97 ± 8.22 (Ens: 46.97)	0.4545 (Ens)	30.30/69.70	28.52 [0.2852]	18.10%/0.4790	15.22 (65.72)|16.29 (61.38)
MobileNetV3-Small (5-Fold)	Static FX INT8	3.81 → 1.18 (69.0%)	56.97 ± 8.22 (FP32)/—	—	—	—	—	20.56 (48.64)|24.73 (40.44)

Note: Italic headings denote sub-table sections (Parts A and B); bold rows highlight the proposed TradAug Selective FX INT8 single-fold champion and 5-fold ensemble configurations.

On TradAug Fold 1, this operating-point recentering shifts True Alert logits below the $z = 0.0$ ($\tau = 0.50$) decision boundary (-2.325 ± 7.565), converting 20 False Positives into True Negatives (FP $\rightarrow$ TN $= 20$) and elevating uncalibrated Fold 1 accuracy from 53.03% to 69.70% (68.52% on $N = 270$). Extending this to the 5-fold ensemble elevates TradAug INT8 Ensemble accuracy to 74.24% [95% CI: 63.64–84.85%] (0.7395 F_1, 15.15% FPR) on $N = 66$ and 76.30% [95% CI: 71.11–81.11%] (0.6885 F_1, 80.48% Specificity, 19.52% FPR) on $N = 270$, achieving 100.0% Specificity (0.0% FPR) on unseen drivers P03 (42/42) and P05 (42/42).

In contrast, the No-Aug (Fold 4) and GenAIAug (Fold 1 and 5-fold ensemble) variations had their accuracy decay to the chance baseline (achieving only 50.00% accuracy and an F_1-score of 0.3333 with 0.00% Specificity). While being able to achieve competitive full-precision results, unperturbed laboratory features in No-Aug Fold 4 and the synthetic boundary artifacts and static-plate textures added by the VFX pipeline in GenAIAug resulted in calibration range mismatches and brittle learned activations. After quantization, these activations were binned to 8-bit integer levels (256 discrete bins), inducing shared-bottleneck zero-point saturation (`zero_point = 80`) and shifting Alert logits further positive ($\Delta z_0 =$ +9.252 and +7.793), which flips residual True Negatives into False Positives and renders uncalibrated cross-domain predictions ineffective (50.00% chance accuracy).

Alongside increased prediction stability for TradAug, the memory footprint of the Selective FX INT8 checkpoint was reduced by 20.6% to 53.10 MB (compared to 66.87 MB for full-precision FP32) across all variations. This substantial reduction was achieved by using the partial quantization recipe, which allows specifying individual layers that will undergo the conversion procedure. Specifically, to achieve the maximum memory savings without sacrificing the temporal fatigue prediction accuracy, the FX graph was modified to only quantize the early convolutional stages and bottleneck/head projections while keeping `MobileViTv2Block` linear self-attention and temporal transformer blocks in FP32.

Execution latency and RAM usage were profiled in a streaming FIFO ring-buffer setup on an x86 CPU edge-simulation benchmark using the `fbgemm` engine across both 1-thread and 4-thread configurations, where caching the 44 historical backbone feature pairs reduces per-frame inference latency by 40× to 46× compared to cold-start 45-frame sequence re-encoding (8045.2 ± 71.8 ms to 9265.5 ± 42.3 ms on 4 threads). On 1 CPU thread, Selective FX INT8 reduces MobileViT-v2 streaming latency by 18.6% from 529.03 ± 19.03 ms (1.89 FPS, 1377.4 MB RSS) in FP32 to 430.65 ± 12.18 ms (2.32 FPS, 1538.9 MB RSS; vs. 478.22 ± 11.41 ms [2.09 FPS, 1538.1 MB RSS] for Head-Only INT8 and 886.27 ± 14.60 ms [1.13 FPS, 1494.6 MB RSS] for Full-Backbone INT8).

On 4 CPU threads, Selective FX INT8 MobileViT-v2 achieves a streaming latency of 200.42 ± 11.85 ms per frame-pair (4.99 FPS, 1589.8 MB peak RSS RAM, compared to 211.19 ± 13.25 ms [4.74 FPS, 1510.8 MB RSS] in FP32, 210.43 ± 9.83 ms [4.75 FPS, 1588.9 MB RSS] in Head-Only INT8, and 342.61 ± 18.05 ms [2.92 FPS, 1570.2 MB RSS] in Full-Backbone INT8), matching the native 5.0 FPS sensor acquisition rate (200 ms budget) for real-time edge execution, while the lightweight MobileNetV3-Small baseline (3.81 MB FP32 $\rightarrow$ 1.18 MB INT8) achieves 16.29 ± 1.64 ms (61.38 FPS, 1452.5 MB RSS) in FP32 vs. 24.73 ± 2.58 ms (40.44 FPS, 1358.3 MB RSS) in Static FX INT8 on 1 thread and 15.22 ± 1.68 ms (65.72 FPS, 1497.3 MB RSS) in FP32 vs. 20.56 ± 2.28 ms (48.64 FPS, 1501.7 MB RSS) in Static FX INT8 on 4 threads.

5. Discussion

5.1. The Data Leakage Illusion and Baseline Validity

A critical review of the recent literature using the UTA-RLDD dataset exhibits a common methodological flaw—a phenomenon which is termed as the "data leakage illusion". Several works report unrealistically high frame-level classification accuracies from heavily

leaking train-test splits in the order of 95% [13,25]. However, as Ghoddoosian et al. [24] and Magán et al. [18] demonstrated, these values arise from a validation procedure involving randomized frame-level train-test splits, which encode a significant "memory effect" into the network: the convolutional network learns to ignore fatigue semantics completely, instead encoding trivial features unique to faces, accessories, and the background of the individual in the training frame as a proxy for the label.

To break this data leakage illusion and establish an honest benchmark for ADAS, a subject-independent, video-level 5-fold cross-validation scheme is imposed, obtaining an internal validation accuracy of 59.96% using the optimal architecture. Rather than being indicative of poor performance, this accuracy should instead be contrasted against the only other zero-leakage baselines in the literature, namely 65.2% by Ghoddoosian et al. [24] and 63.0% by Magán et al. [18], which both use the UTA-RLDD dataset. In fact, this strict validation scheme is essential for rigorously evaluating generalization performance, as it guarantees that no driver in the validation set ever appeared in the model's training set. Consequently, this study establishes that the proposed architecture serves as a rigorous, generalization-focused baseline optimized for real-world edge cases.

To ensure rigorous baseline validity, this study evaluated key literature architectures on the exact same 5-fold subject-independent splits: Magán et al.'s EfficientNetB0_GRU baseline [18] achieved 51.09% ± 4.29% internal accuracy (0.4207 F_1) on UTA-RLDD and collapsed to 53.03% accuracy (87.88% FPR) on DriverMVT, while Single-Stream BottleneckTransformer (50.00% accuracy, 0.2819 ROC-AUC) and Single-Stream BottleneckTCN (53.03% accuracy, 0.3949 ROC-AUC) both experienced cross-domain collapse. This confirms that two-stream motion isolation paired with backbone freezing is required for true cross-domain generalization.

5.2. The Laboratory-to-Vehicle (Sim-to-Real) Domain Shift and Significant Spatial Collapse

The transition from constrained laboratory datasets to real-world vehicular environments presents a significant Lab-to-Vehicle domain shift that exposes the fragility of conventional DL fine-tuning strategies. Direct cross-domain evaluation on the unseen DriverMVT vehicular cabin dataset clearly demonstrates this vulnerability. When a partially unfrozen MobileNetV3-Small baseline is deployed across domains, it suffers a complete loss of specificity, yielding zero True Negatives (TN = 0) during out-of-distribution testing.

The mechanism driving this severe structural failure lies within the gradient update process. Allowing upper convolutional layers to update gradients during training on a homogeneous, static dataset causes the spatial stream to overfit to environmental invariants. Specifically, the network memorizes the static laboratory background as a primary spatial anchor for the "Alert" class. Unseen vehicular environments introduce novel visual elements, including steering wheels, dynamic ambient lighting, and complex cabin interiors. Upon encountering these elements, the ungrounded spatial stream undergoes a severe representation collapse, flagging novel background contexts as anomalies and defaulting to "Drowsy" predictions regardless of driver vigilance.

This failure demonstrates the importance of architectural inductive bias in cross-domain driver monitoring. Directly contrasting the collapsed unfrozen baselines with the proposed framework highlights the effectiveness of locking 100% of parameters within a domain-aligned spatial backbone (MobileViT-v2). The explicit complete parameter freeze acts as a strong inductive bias, insulating the spatial feature extractor from background memorization and identity overfitting during the training phase. By forcing the network to optimize solely the low-dimensional information bottleneck and subsequent temporal transformer layers, the architecture attenuates contextual background noise and focuses on temporal fatigue dynamics.

This structural intervention substantially suppresses the spatial shortcut learning phenomenon, recovering True Negatives and curbing the Lab-to-Vehicle spatial domain shift without relying on computationally expensive generative pipelines. The 2 × 3 controlled freezing matrix (Table 3) reveals a consistent freezing law: partially unfreezing either MobileNetV3-Small or MobileViT-v2 causes an identical 0.00% Specificity collapse (TN = 0, 100% FPR on $N = 66$; ≤2.86% on $N = 270$). Conversely, 100% fully unfreezing either backbone triggers severe sensitivity collapse (33.33% TPR for MobileNetV3-Small and 18.18% TPR for MobileViT-v2; 0.3494 ROC-AUC on $N = 270$). Enforcing 100% spatial backbone freezing reduces, though does not completely eliminate, this cross-domain degradation, recovering TN = 10–14 (56.97% mean accuracy) in MobileNetV3-Small and achieving 71.21% Ensemble accuracy (0.8838 ROC-AUC) in MobileViT-v2.

Controlled stream-occlusion (stream-isolation) ablations provide direct physical proof that downstream trainable layers do not exploit residual background or identity shortcuts: occluding the frame-differencing Motion Stream (Spatial-Only, $m = 0$) causes Specificity on DriverMVT to drop precipitously from 45.45% (TN = 15) to 15.15% (TN = 5), dropping Ensemble accuracy from 71.21% to 57.58% ($F_1 = 0.4826$). Conversely, occluding the Spatial Stream to evaluate Motion-Only ($s = 0$) achieves 69.70% Ensemble Specificity (81.82% in Fold 2), proving that downstream temporal layers rely on the frame-differencing motion stream ($\Delta I_{\text{bg}} \approx 0$) to suppress background false alarms rather than exploiting static spatial shortcuts.

To address the 54.55% False Positive Rate ($\text{TN} = 15, \text{FP} = 18$) of the uncalibrated FP32 Ensemble at default threshold $\tau = 0.50$, logit analysis shows that this high FPR stems from an upward logit bias in Folds 1 and 3 under cross-domain transfer, despite strong class separability (ROC-AUC = 0.8838; Fold 2 independently achieves 84.85% accuracy and 15.15% FPR at $\tau = 0.50$). Calibrating the decision threshold to Youden's J^* (τ^*) slashes the FPR to 12.12% ($\text{TN} = 29, \text{FP} = 4$), boosting accuracy to 87.88% (0.8788 Macro F_1, 87.88% TPR, 87.88% TNR).

Alternatively, applying Selective FX INT8 quantization inherently recenters logits ($\Delta z_0 = -10.878$), reducing FPR to 15.15% (74.24% Ensemble accuracy on $N = 66$) and 19.52% FPR (76.30% accuracy, 80.48% TNR across 5 drivers on $N = 270$), substantially attenuating false alarms relative to uncalibrated FP32 inference, although commercial ADAS pipelines should still pair window-level inference with multi-window temporal hysteresis or confirmation filtering to suppress residual nuisance alerts (12.12–19.52% FPR) during extended driving.

5.3. Architecture vs. Augmentation: The Volatility of Generative VFX

A critical analysis of the empirical results provides several insights into the practicalities of using generative AI for domain adaptation in temporal driver monitoring. First, the generative visual effects pipeline, though intended to bridge the Lab-to-Vehicle gap, exhibited structural and temporal limitations that degraded its effectiveness for sequential domain adaptation.

The first structural factor was that targeted scarcity augmentations were applied to 2 of the 60 participants (Participants 26 and 55), which localized synthetic variations to specific training folds. Crucially, while augmenting 2 of 60 participants (Participant 26 Drowsy in Fold 3 and Participant 55 Alert in Fold 5; 3.3% of the cohort) introduces fold-level distribution asymmetry relative to No-Aug (+6.29% sequence volume), both TradAug and GenAIAug augmented the exact same 2 participants with the exact same sample volume (+1584 sequences each, expanding the training corpus from 25,178 to 26,762 sequences). Therefore, the head-to-head superiority of TradAug (71.21% FP32/74.24% INT8 Ensemble accuracy) over GenAIAug (60.61% FP32/50.00% INT8 Ensemble accuracy) is 100% participant-

and density-matched, isolating generative spatio-temporal artifacts—rather than participant count alone—as the primary cause of cross-domain performance degradation.

This fold-level asymmetry manifested as elevated variance during cross-validation. To illustrate this volatility, the performance differences across folds trained with synthetically enriched data are analyzed below.

While in Fold 1 the GenAIAug variant reached a high internal validation accuracy of 72.03%, this specialization did not transfer consistently across folds. When evaluated on Fold 2 (which was also trained on the synthetically enriched distribution), the model's internal validation accuracy dropped to 48.04%. This indicates that the GenAIAug model failed to learn a temporally stable representation across validation splits, oscillating between fold-specific specialization and severe generalization degradation depending on the validation partition.

The DL mechanic that is responsible for this unstable behavior is related to the requirements that the temporal sequence modeling imposes. While the VFX-style compositing pipeline excellently injected robust spatial priors, the frame-by-frame alpha-matting process unintentionally introduced a number of microscopic edge artifacts. Over the course of a 45-frame video sequence, these per-frame inconsistencies severely disrupted the optical flow and coherent temporal continuity that are implicitly expected by the subsequent transformer layers. As such, the temporal attention mechanism was forced to deal with artificial flickering rather than realistic fatigue dynamics.

To quantitatively prove the physical mechanisms corrupting the motion stream ($\Delta I_t = |I_t - I_{t-1}|$), an optical-flow temporal continuity audit was conducted across $N = 20$ 45-frame sequences per augmentation regime (900 frame pairs per regime; 2700 total frame pairs across all three regimes, Table 5).

Physical measurements in Table 5 reveal two distinct distortion mechanisms in GenAIAug that degrade motion stream representations. First, active-foreground alpha-matting introduces severe boundary flicker, elevating Laplacian edge variance to 27.20±3.95 (+63.0% vs. No-Aug [$p = 1.00 \times 10^{-7}$]; +77.2% vs. TradAug [$p = 1.93 \times 10^{-7}$]), inflating temporal spike ratio R_{spike} by +25.5% (2.212 vs. 1.763, $p = 1.23 \times 10^{-3\,M}$, $p_{\text{Welch}} = 0.0340^{W}$), concentrating +16.9% more motion on matte contours (2.369 vs. 2.026, $p = 0.0198$), and increasing Farnebäck optical-flow spatial shear S_{shear} by +22.7% vs. TradAug (1.5495 vs. 1.2633, $p = 3.68 \times 10^{-4}$).

Second, compositing onto static 2D LoRA+ car plates ($\Delta I_{\text{bg}} = 0$) deprives the background of natural motion, inflating zero-motion pixels by +107.8% (81.42% ± 5.04% vs. 39.19% in No-Aug, $p = 3.27 \times 10^{-15}$) and depressing global motion energy by −59.8% (0.926 vs. 2.302, $p = 1.15 \times 10^{-6}$). When deployed to real moving vehicles (DriverMVT) where passing scenery and cabin vibration create continuous optical flow, models trained on this synthetic background-replacement regime misclassify normal background dynamics as anomalous fatigue, collapsing to 27.27% Specificity (9/33 TN, 72.73% FPR) under FP32 and 0.00% Specificity (0/33 TN, 100% FPR) under INT8 quantization.

It is proposed that future works on sequence-level domain adaptation should abandon frame-wise VFX compositing pipelines in favor of native temporal video generation models that are architected to naturally produce fluid, temporally consistent driving environments [57]. Ultimately, however, empirical evidence has demonstrated that until generative video synthesis becomes computationally viable, architectural inductive bias is of paramount importance. Specifically, freezing a domain-aligned spatial backbone provides the most globally stable and computationally efficient solution for reducing the Lab-to-Vehicle gap in edge-ADAS applications.

Table 5. Quantitative Optical-Flow Temporal Continuity and Spatio-Temporal Artifact Audit ($N = 20$ 45-frame sequences per regime, 2700 total frame pairs $\Delta I_t = |I_t - I_{t-1}|$), with statistical significance assessed via two-sided Welch's t-test (^{W}) and two-sided Mann-Whitney U test (^{M}).

Augmentation Regime	Motion Energy ($\bar{E}_{\Delta I}$)	Active Laplacian Flicker (σ^2_{Lap})	Zero-Motion Ratio (P_0 %)	Spike Ratio (R_{spike})	Edge Motion Ratio (R_{edge})	Spatial Shear (S_{shear})
No-Aug (Original Lab)	2.302 ± 0.895	16.69 ± 5.75	39.19 ± 10.76	1.763 ± 0.795	2.026 ± 0.552	1.4151 ± 0.3321
TradAug (Photometric/Geometric)	2.211 ± 0.568	15.35 ± 6.86	43.78 ± 5.75	2.160 ± 0.552	2.203 ± 0.460	1.2633 ± 0.2259
GenAIAug (VFX Composited)	$\mathbf{0.926 \pm 0.281}$	$\mathbf{27.20 \pm 3.95}$	$\mathbf{81.42 \pm 5.04}$	$\mathbf{2.212 \pm 0.429}$	$\mathbf{2.369 \pm 0.285}$	$\mathbf{1.5495 \pm 0.2368}$
Ratio & p-value (GenAI vs. No-Aug)	-59.8% ($p = 1.15 \times 10^{-6\,W}$)	$+63.0\%$ ($p = 1.00 \times 10^{-7\,W}$)	$+107.8\%$ ($p = 3.27 \times 10^{-15\,W}$)	$+25.5\%$ ($p = 1.23 \times 10^{-3\,M}$)	$+16.9\%$ ($p = 0.0198^W$)	$+9.5\%$ ($p = 0.150^W$)
Ratio & p-value (GenAI vs. TradAug)	-58.1% ($p = 8.70 \times 10^{-10\,W}$)	$+77.2\%$ ($p = 1.93 \times 10^{-7\,W}$)	$+86.0\%$ ($p = 5.36 \times 10^{-23\,W}$)	$+2.4\%$ ($p = 0.740^W$)	$+7.6\%$ ($p = 0.179^W$)	$+22.7\%$ ($p = 3.68 \times 10^{-4\,W}$)

Note: Bold formatting highlights the GenAIAug regime exhibiting temporal artifacts; italic rows denote pairwise statistical comparisons.

5.4. Low-Precision Lossy Regularization and Edge Feasibility

Contrary to the conventional expectation of accuracy degradation under standard post-training quantization, out-of-distribution evaluation on the DriverMVT dataset revealed that the TradAug Fold 1 checkpoint achieved a significant positive accuracy gain (+16.67%), rising from 53.03% FP32 accuracy to 69.70% INT8 accuracy.

The phenomenon can be justified by analyzing the lossy regularization mechanism imposed by the Selective FX Static PTQ procedure. Photometric shift, flipping, and Gaussian blurring regularize the network weight distributions in the gradient descent update procedure, forming wide, high noise-tolerance subspaces during the optimization. The subsequent application of a static 8-bit uniform quantization becomes a powerful tool for constraining the network by further regularizing the distributions.

Specifically, the hard integer truncation during the mapping attenuates unstable, overfitted floating-point activations in the latent space of the spatio-temporal bottleneck. These high-magnitude activations in the continuous space represent domain-shifted logit offsets relative to the default $\tau = 0.50$ threshold. Crucially, this mechanism separates threshold-independent rank separability (governed by the frozen backbone and training augmentation, measured via ROC-AUC) from uncalibrated threshold alignment at $\tau = 0.50$: L2 activation damping in quantized early convolutional layers shifts inflated cross-domain logits downward across the $z = 0.0$ decision boundary without target-domain tuning, rather than generating new semantic features or universally boosting arbitrary architectures.

Mechanistically profiling activations and decision margins (Figure 5) reveals the mathematical cause of this gain. In FP32, TradAug Fold 1 possessed the highest rank separability across all trained checkpoints (ROC-AUC = 0.9343, PR-AUC = 0.9558), but its raw logits were shifted upward under Lab-to-Vehicle transfer: True Alert logits averaged $+8.553 \pm 5.151$ and True Drowsy averaged $+10.790 \pm 1.730$ relative to the $z = 0.0$ boundary (31/33 false alarms).

Under Selective FX INT8, early 1024-D backbone activations undergo L2-norm damping (13.257 → 12.299 spatial; 11.794 → 11.198 motion) with 22.6–26.3% zero-clamping while preserving motion bottleneck directional fidelity (cos = 0.6207, SQNR = +0.61 dB). This induces a uniform negative logit recentering ($\Delta z_0 = -10.878$ on Alert, $\Delta z_1 = -7.880$ on Drowsy), shifting True Alert logits below $z = 0.0$ to -2.325 ± 7.565 while keeping True Drowsy positive ($+2.910 \pm 7.239$). This converts 20 False Positives directly into True Negatives (FP → TN = 20) at the cost of 9 TP → FN flips, yielding a net gain of +11 sequences (+16.67%).

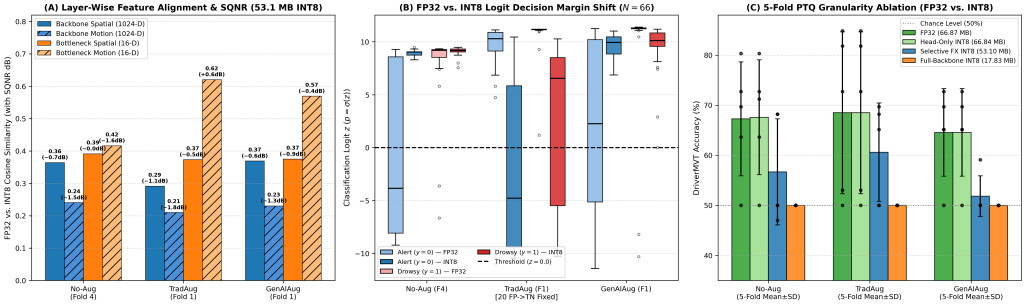

Figure 5. Mechanistic analysis of Selective FX INT8 Post-Training Quantization on DriverMVT ($N = 66$). (**A**) Layer-wise FP32-to-INT8 cosine similarity and SQNR (dB) across spatial and motion streams (53.10 MB), showing preserved motion bottleneck fidelity (cos = 0.6207, SQNR = +0.61 dB) on TradAug Fold 1; (**B**) decision margin ($z_1 - z_0$) distribution shift showing uniform negative logit recentering ($\Delta z_0 = -10.878$, FP → TN = 20) on TradAug Fold 1 versus positive saturation on No-Aug Fold 4 and GenAIAug Fold 1; and (**C**) 5-fold PTQ granularity ablation comparing FP32 (66.87 MB), Selective INT8 (53.10 MB), Full-Backbone INT8 (17.83 MB, 50.0% accuracy), and the 5-fold Selective INT8 Ensemble (74.24%).

Conversely, for No-Aug Fold 4 and GenAIAug Fold 1, where weights lack photometric/geometric noise conditioning or inherit alpha-matting boundary flicker, calibrating the shared bottleneck (`zero_point = 80`) induces positive logit saturation ($\Delta z_0 = +9.252$ for No-Aug Fold 4; $\Delta z_0 = +7.793$ for GenAIAug Fold 1), flipping all True Negatives into False Positives (50.00% accuracy) and confirming that positive PTQ recentering is strictly regime-dependent rather than a universal property of INT8 quantization. Crucially, when extended across all 5 folds, the 5-Fold TradAug INT8 Ensemble (74.24% Accuracy, 0.7395 F_1, 15.15% FPR on $N = 66$; 76.30% Accuracy, 80.48% TNR across 5 drivers on $N = 270$) strictly outperforms the equivalent 5-Fold FP32 Ensemble (71.21% Accuracy, 54.55% FPR), exhibiting 0.00% variance across 3 independent calibration seeds (42, 100, 2026).

Perhaps more importantly, the practical value of the Selective FX INT8 procedure is evident across both single-checkpoint and ensemble deployment modes. When deploying a single quantized checkpoint on memory-constrained edge hardware, the quantized TradAug Fold 1 model (69.70% accuracy) approaches the performance of the 5-fold TradAug FP32 ensemble (71.21% accuracy, 334.35 MB combined footprint), while under an equivalent 5-fold ensemble evaluation, the 5-fold Selective FX INT8 ensemble (74.24% on $N = 66$ and 76.30% on $N = 270$) strictly surpasses the 5-fold FP32 ensemble by +3.03 percentage points (pp) on $N = 66$ (71.21% → 74.24%) and +52.23 pp on $N = 270$ (24.07% → 76.30% at uncalibrated $\tau = 0.50$).

This demonstrates that the logit recentering and activation damping imposed by Selective FX INT8 weight mapping both approximate ensemble regularization in single-model deployment and compound synergistically with 5-fold ensemble averaging.

By contrast, the severe 8-bit representation collapse exhibited by the GenAIAug variant highlights its poor robustness to integer activation truncation. Specifically, the GenAIAug model, despite achieving competitive full-precision sensitivity, collapses to a chance-level 50.00% accuracy (0.00% Specificity) following static quantization. Thus, traditional photometric and geometric perturbations more effectively shape the network's weight and activation topology to remain resilient to integer casting. Such numerical resilience is critical for sustaining spatio-temporal discrimination after precision reduction in embedded automotive safety systems.

5.5. Edge Benchmarks, Study Limitations, and Future Directions

Benchmarking the proposed Selective FX INT8 model in a streaming FIFO ring-buffer setup on 4 CPU threads (`fbgemm` engine) yields a desktop CPU simulation throughput of 4.99 FPS (200.42 ± 11.85 ms per frame-pair, 1589.8 MB peak RSS RAM; 2.32 FPS [430.65 ms] on 1 thread), establishing algorithmic and memory feasibility against the native 5.0 FPS camera acquisition rate rather than physical automotive ECU qualification. However, two explicit study limitations should be addressed in future work. First, hardware-in-the-loop (HIL) deployment profiling—measuring active power dissipation (Watts), thermal throttling, multi-window temporal alert hysteresis, and hardware NPU/TensorRT latency on physical automotive System-on-Chip (SoC) hardware such as NVIDIA Jetson Orin or TI TDA4VM—remains to be conducted prior to commercial vehicle integration. Second, generative data augmentation should be expanded across all 60 cohort participants using temporally coherent video diffusion models rather than frame-wise compositing.

6. Conclusions

Advancing driver vigilance monitoring requires balancing cross-domain generalization with embedded hardware constraints, as existing SOTA methods often rely on non-differentiable preprocessing or frame-level "data leakage" splits that favor background memorization. The proposed framework addresses these challenges: by enacting a strict video-level, subject-independent cross-validation paradigm, this study demonstrates that high-level spatio-temporal fatigue dynamics can be captured and classified by a parameter-efficient architecture (41,906 trainable temporal parameters [0.24%] out of 17,466,235 total inference parameters [66.87 MB FP32 → 53.10 MB Selective INT8] for MobileViT-v2, compared against the 961,746-parameter [34,738 trainable] MobileNetV3-Small baseline). As such, the framework achieves competitive zero-leakage accuracy tailored for edge-hardware constraints.

A key insight of this study concerns reducing the Lab-to-Vehicle spatial domain shift. Empirical evaluation indicates that allowing spatial backbones to be optimized on constrained laboratory datasets consistently leads to background memorization, resulting in a severe spatial representation collapse when the model is exposed to novel real-world vehicle cabins. It is demonstrated that this limitation is mitigated by recognizing that architectural inductive bias outperforms naive frame-wise synthetic data compositing. Specifically, completely freezing a domain-aligned spatial backbone (MobileViT-v2) curbed the Lab-to-Vehicle gap and recovered True Negatives in cross-domain evaluations. This structural intervention avoids the need for volatile and computationally expensive frame-wise VFX augmentation. Furthermore, empirical analysis confirms that such frame-wise compositing actively disrupts the inter-frame optical-flow continuity required by temporal transformer layers for sequential fatigue modeling.

Finally, algorithmic and memory feasibility are demonstrated through desktop CPU edge-simulation benchmarks. Application of Selective FX INT8 PTQ revealed a notable regularization effect, wherein traditional data augmentations coupled with partial INT8 mapping acted as an activation-damping and logit-recentering operator. The resulting numerical truncation effectively filtered out unstable, domain-shifted latent activations, yielding a hardware-resilient INT8 model.

Notably, the 5-fold Selective INT8 Ensemble elevated cross-domain accuracy to 74.24% (15.15% FPR) on $N = 66$ and 76.30% (80.48% Specificity across 5 drivers; 100% Specificity on unseen subjects P03 and P05) on $N = 270$, while achieving a 20.6% memory footprint reduction (53.10 MB) and 4.99 FPS (200.42 ms) streaming throughput on CPU edge simulation. While direct hardware-in-the-loop profiling of active power draw, thermal dissipation, multi-window temporal hysteresis filtering, and hardware NPU acceleration on physical

automotive SoCs (such as NVIDIA Jetson Orin or TI TDA4VM) remains an essential direction for future work, this study establishes that combining strict architectural inductive biases with targeted partial quantization provides a reproducible, quantitatively grounded foundation for next-generation embedded ADAS research.

Author Contributions: Conceptualization, M.B., B.R.H. and G.A.E.; Data Curation, M.B.; Formal Analysis, M.B.; Investigation, M.B., B.R.H. and G.A.E.; Methodology, M.B., B.R.H. and G.A.E.; Project Administration, B.R.H. and G.A.E.; Resources, M.B.; Software, M.B.; Supervision, B.R.H. and G.A.E.; Validation, M.B., B.R.H. and G.A.E.; Visualization, M.B.; Writing—Original Draft Preparation, M.B.; Writing—Review and Editing, M.B., B.R.H. and G.A.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable. No new human subjects were enrolled by the authors. All human facial images analyzed in the experiments and depicted in Figures 1 and 2 are unmodified foreground frames sourced exclusively from the preexisting, publicly distributed academic benchmark datasets UTA-RLDD [24] and DriverMVT [33], for which participant consent for academic research and publication was established by the original dataset creators.

Data Availability Statement: The benchmark datasets analyzed in this study are publicly available in permanent online repositories: (1) the University of Texas at Arlington Real-Life Drowsiness Dataset (UTA-RLDD) [24] is publicly accessible at https://sites.google.com/view/utarldd/home (accessed on 13 October 2026) (reference DOI: https://doi.org/10.1109/CVPRW.2019.00027); and (2) the Driver Monitoring Dataset with Video and Telemetry (DriverMVT) [33] is publicly archived on Zenodo at https://doi.org/10.5281/zenodo.6542011 (Part 1) and https://doi.org/10.5281/zenodo.6584185 (Part 2), with the corresponding dataset descriptor available at https://doi.org/10.3390/data7050062. The training, evaluation, and generative compositing source code, empirical audit logs, and compressed Selective FX INT8 model checkpoints are available from the corresponding author upon reasonable request.

Acknowledgments: During the preparation of this study, the authors used Stable Diffusion v1.5 (`runwayml/stable-diffusion-v1-5`, via Hugging Face `diffusers` v0.25.0 and `transformers` v4.36.0) coupled with custom Low-Rank Adaptation (`peft` v0.7.1, LoRA+ and T-LoRA), manual Pen Tool masking in GIMP, and U^2-Net (`rembg`, `u2net_human_seg`) for the purposes of generating synthetic in-cabin vehicle background plates (`plate_1.png`–`plate_4.png`) and compositing real UTA-RLDD driver foregrounds with GIMP-masked steering wheel overlays (`plate_1_driving_wheel.png`–`plate_4_driving_wheel.png`) for the GenAIAug ($\text{Dataset}_{\text{GenAI}}$) experimental corpus illustrated in Figure 2C. The prompts used were: positive prompt "interior of a car, empty driver seat, steering wheel, dashboard camera perspective, daytime, highly detailed, photorealistic, no driver"; negative prompt "person, human, face, body, hands, driver, ghost, silhouette, cartoon, illustration, distorted, blurry, low quality"; and LoRA+ adapter training prompt "a photo of sks driver inside a car, low vigilance, realistic, 4k". No generative AI tools were used to generate or alter the real human driver faces (which originate strictly from the public UTA-RLDD dataset). The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Smoliński, A.; Forczmański, P.; Nowosielski, A. Processing and Integration of Multimodal Image Data Supporting the Detection of Behaviors Related to Reduced Concentration Level of Motor Vehicle Users. *Electronics* **2024**, *13*, 2457. [CrossRef]
2. Yang, G.; Ridgeway, C.; Miller, A.; Sarkar, A. Comprehensive Assessment of Artificial Intelligence Tools for Driver Monitoring and Analyzing Safety Critical Events in Vehicles. *Sensors* **2024**, *24*, 2478. [CrossRef] [PubMed]
3. Duan, C.; Liu, Z.; Xia, J.; Zhang, M.; Liao, J.; Cao, L. Enhancing Cross-Dataset Performance of Distracted Driving Detection With Score Softmax Classifier and Dynamic Gaussian Smoothing Supervision. *IEEE Trans. Intell. Veh.* **2025**, *10*, 282–295. [CrossRef]
4. Rimal, R.; Herceg, M.; Vranješ, M.; Grbić, R. Driver Monitoring System using an Embedded Computer Platform. In Proceedings of the 2023 International Conference on Software, Telecommunications and Computer Networks (SoftCOM), Split, Croatia, 21–23 September 2023; pp. 1–7. [CrossRef]
5. Li, T.; Zhang, Y.; Li, Q.; Zhang, T. AB-DLM: An Improved Deep Learning Model Based on Attention Mechanism and BiFPN for Driver Distraction Behavior Detection. *IEEE Access* **2022**, *10*, 83138–83151. [CrossRef]
6. Yang, D.; Li, X.; Dai, X.; Zhang, R.; Qi, L.; Zhang, W.; Jiang, Z. All In One Network for Driver Attention Monitoring. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 2258–2262. [CrossRef]
7. Fathy, M.; Ramadan, A.; Emad, A.; Atia, N.; Mamdouh, A.; Ghareeb, M.; Aboelkheir, A.; Salah, M. Optimizing Driver Monitoring System Performance on TI-TDA4VM Edge Device. In Proceedings of the 2024 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), Cairo, Egypt, 13–14 November 2024; pp. 121–127. [CrossRef]
8. M, G.; Asha, V.; Katingeri, M.M.; L J, M.; Ch, M.R.P.; Mathew, J.C. Deep Learning for Real-Time Drowsiness Onset Detection. In Proceedings of the 2024 International Conference on Integrated Circuits and Communication Systems (ICICACS), Raichur, India, 23–24 February 2024; pp. 1–7. [CrossRef]
9. Ahuja, H.; Saurav, S.; Srivastava, S.; Shekhar, C. Driver Drowsiness Detection using Knowledge Distillation Technique for Real Time Scenarios. In Proceedings of the 2020 IEEE 17th India Council International Conference (INDICON), New Delhi, India, 10–13 December 2020; pp. 1–5. [CrossRef]
10. Khalil, H.A.; Hammad, S.A.; Abd El Munim, H.E.; Maged, S.A. Low-Cost Driver Monitoring System Using Deep Learning. *IEEE Access* **2025**, *13*, 14151–14164. [CrossRef]
11. Awachat, S.; Mundada, K.; Dewangan, P.; Ranjan, C. DrowseGuard—DeepAlert Driver Vigilance System. In Proceedings of the 2024 2nd World Conference on Communication & Computing (WCONF), Raipur, India, 12–14 July 2024; pp. 1–6. [CrossRef]
12. El Zein, H.; Harb, H.; Delmotte, F.; Zahwe, O.; Haddad, S. VAS-3D: A Visual-Based Alerting System for Detecting Drowsy Drivers in Intelligent Transportation Systems. *World Electr. Veh. J.* **2024**, *15*, 540. [CrossRef]
13. Majeed, F.; Shafique, U.; Safran, M.; Alfarhood, S.; Ashraf, I. Detection of Drowsiness among Drivers Using Novel Deep Convolutional Neural Network Model. *Sensors* **2023**, *23*, 8741. [CrossRef] [PubMed]
14. Khandare, H.; Rajak, A.; Tripathi, R. A deep learning framework for non-intrusive driver drowsiness detection. In Proceedings of the 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 6–8 July 2023; pp. 1–5. [CrossRef]
15. Suguna, R.; Priya, M.S.; Hema, S.; Aravind, K.D.; Gowtham, A.M.; Vignesh, S. Deep Learning Based Driver Drowsiness Monitoring and Alert System. In Proceedings of the 2023 International Conference on Circuit Power and Computing Technologies (ICCPCT), Kollam, India, 10–11 August 2023; pp. 241–246. [CrossRef]
16. Kim, W.; Lee, Y.K.; Jung, W.S.; Yoo, D.; Kim, D.H.; Jo, K.H. An Adaptive Batch-Image Based Driver Status Monitoring System on a Lightweight GPU-Equipped SBC. *IEEE Access* **2020**, *8*, 206074–206087. [CrossRef]
17. Florez, R.; Palomino-Quispe, F.; Alvarez, A.B.; Coaquira-Castillo, R.J.; Herrera-Levano, J.C. A Real-Time Embedded System for Driver Drowsiness Detection Based on Visual Analysis of the Eyes and Mouth Using Convolutional Neural Network and Mouth Aspect Ratio. *Sensors* **2024**, *24*, 6261. [CrossRef] [PubMed]
18. Magán, E.; Sesmero, M.P.; Alonso-Weber, J.M.; Sanchis, A. Driver Drowsiness Detection by Applying Deep Learning Techniques to Sequences of Images. *Appl. Sci.* **2022**, *12*, 1145. [CrossRef]
19. Anagnostou, P.; Mitianoudis, N. A lightweight ConvGRU network for Distracted Driving detection. In Proceedings of the 2023 24th International Conference on Digital Signal Processing (DSP), Rhodes (Rodos), Greece, 11–13 June 2023; pp. 1–5. [CrossRef]
20. Borrego-Carazo, J.; Castells-Rufas, D.; Biempica, E.; Carrabina, J. Resource-Constrained Machine Learning for ADAS: A Systematic Review. *IEEE Access* **2020**, *8*, 40573–40598. [CrossRef]
21. Tanama, C.; Peng, K.; Marinov, Z.; Stiefelhagen, R.; Roitberg, A. Quantized Distillation: Optimizing Driver Activity Recognition Models for Resource-Constrained Environments. In Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Detroit, MI, USA, 1–5 October 2023; pp. 5479–5486. [CrossRef]
22. Alajlan, N.N.; Ibrahim, D.M. DDD TinyML: A TinyML-Based Driver Drowsiness Detection Model Using Deep Learning. *Sensors* **2023**, *23*, 5696. [CrossRef] [PubMed]

23. Ramesh, G.B.; Chamas, M.; Wang, A.; Sax, E. Auto-Adaptive Diffusion Model-Based Uniformization of Automotive Camera Sensor Data. In Proceedings of the 2024 8th International Conference on Advances in Artificial Intelligence, London, UK, 17–19 October 2024; pp. 246–255. [CrossRef]
24. Ghoddoosian, R.; Galib, M.; Athitsos, V. A Realistic Dataset and Baseline Temporal Model for Early Drowsiness Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–17 June 2019; pp. 178–187. [CrossRef]
25. Ganguly, B.; Dey, D.; Munshi, S. An Attention Deep Learning Framework-Based Drowsiness Detection Model for Intelligent Transportation System. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 4517–4527. [CrossRef]
26. Qu, F.; Dang, N.; Furht, B.; Nojoumian, M. Comprehensive study of driver behavior monitoring systems using computer vision and machine learning techniques. *J. Big Data* **2024**, *11*, 32. [CrossRef]
27. Barrios, F.A.; Emadi, A. A Developer's Guide to Compressing Pre-Trained Transformer Neural Networks Across Different Domains. *IEEE Access* **2025**, *13*, 166873–166891. [CrossRef]
28. Li, Z.; Li, H.; Meng, L. Model Compression for Deep Neural Networks: A Survey. *Computers* **2023**, *12*, 60. [CrossRef]
29. Shafique, M.A.; Munir, A.; Kong, J. Deep Learning Performance Characterization on GPUs for Various Quantization Frameworks. *AI* **2023**, *4*, 926–948. [CrossRef]
30. Li, M.; Huang, Z.; Chen, L.; Ren, J.; Jiang, M.; Li, F.; Fu, J.; Gao, C. Contemporary Advances in Neural Network Quantization: A Survey. In Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN), Yokohama, Japan, 30 June–5 July 2024; pp. 1–10. [CrossRef]
31. Jelić, B.; Grbić, R.; Vranješ, M.; Mijić, D. Can We Replace Real-World With Synthetic Data in Deep Learning-Based ADAS Algorithm Development? *IEEE Consum. Electron. Mag.* **2023**, *12*, 32–38. [CrossRef]
32. Peng, M.; Chen, K.; Guo, X.; Zhang, Q.; Zhong, H.; Zhu, M.; Yang, H. Diffusion Models for Intelligent Transportation Systems: A Survey. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 21526–21543. [CrossRef]
33. Othman, W.; Kashevnik, A.; Ali, A.; Shilov, N. DriverMVT: In-Cabin Dataset for Driver Monitoring including Video and Vehicle Telemetry Information. *Data* **2022**, *7*, 62. [CrossRef]
34. Wani, A.; Singh, J.; Kumari, D.; Ithape, A.; Rapanwad, G. FEV's 'CogniSafe': An Innovative Deep Learning-Based AI Driver Monitoring System for the Future of Mobility. In Proceedings of the SAE Technical Paper Series, Detroit, MI, USA, 16–18 April 2024; Product Code 2024-01-2012 . [CrossRef]
35. Li, T.; Li, C. A Deep Learning Model Based On Multi-granularity Facial Features And LSTM Network For Driver Drowsiness Detection. *J. Appl. Sci. Eng.* **2023**, *27*, 2869–2881. [CrossRef]
36. Civik, E.; Yuzgec, U. Real-time driver fatigue detection system with deep learning on a low-cost embedded system. *Microprocess. Microsystems* **2023**, *99*, 104851. [CrossRef]
37. Su, L.; Sun, C.; Cao, D.; Khajepour, A. Efficient Driver Anomaly Detection via Conditional Temporal Proposal and Classification Network. *IEEE Trans. Comput. Soc. Syst.* **2023**, *10*, 736–745. [CrossRef]
38. Li, Z.; Li, M.; Shi, L.; Li, D. A novel fatigue driving detection method based on whale optimization and Attention-enhanced GRU. *Intell. Robot.* **2024**, *4*, 230–243. [CrossRef]
39. Das, S.; Pratihar, S.; Pradhan, B.; Jhaveri, R.H.; Benedetto, F. IoT-Assisted Automatic Driver Drowsiness Detection through Facial Movement Analysis Using Deep Learning and a U-Net-Based Architecture. *Information* **2024**, *15*, 30. [CrossRef]
40. Dantas, P.V.; Sabino Da Silva, W.; Cordeiro, L.C.; Carvalho, C.B. A comprehensive review of model compression techniques in machine learning. *Appl. Intell.* **2024**, *54*, 11804–11844. [CrossRef]
41. Singh, S.; Sharma, K.; Karna, B.K.; Raj, P. Pruning and Quantization for Deeper Artificial Intelligence (AI) Model Optimization. In *Proceedings of the Intelligent Control, Robotics, and Industrial Automation*; Sharma, S., Subudhi, B., Sahu, U.K., Eds.; Springer : Singapore, 2023; Volume 1066, pp. 933–945. [CrossRef]
42. Zhao, M.; Li, M.; Peng, S.L.; Li, J. A Novel Deep Learning Model Compression Algorithm. *Electronics* **2022**, *11*, 1066. [CrossRef]
43. Zhang, Z.; Li, J. A Review of Artificial Intelligence in Embedded Systems. *Micromachines* **2023**, *14*, 897. [CrossRef] [PubMed]
44. Merone, M.; Graziosi, A.; Lapadula, V.; Petrosino, L.; d'Angelis, O.; Vollero, L. A Practical Approach to the Analysis and Optimization of Neural Networks on Embedded Systems. *Sensors* **2022**, *22*, 7807. [CrossRef] [PubMed]
45. Wang, T.; Li, Y.; Tian, H.; Xie, J. Selective knowledge distillation-based domain adaptation framework towards edge computing fault Diagnosis for high-speed train bogie. *Green Energy Intell. Transp.* **2026**, *5*, 100373. [CrossRef]
46. Zhuang, T.; Liang, X.; Xue, B.; Tang, X. An in-vehicle real-time infrared object detection system based on deep learning with resource-constrained hardware. *Intell. Robot.* **2024**, *4*, 276–292. [CrossRef]
47. Ou, C.; Karray, F. Enhancing Driver Distraction Recognition Using Generative Adversarial Networks. *IEEE Trans. Intell. Veh.* **2020**, *5*, 385–396. [CrossRef]
48. Yang, Y.; Sun, L.; Mao, X.; Zhao, M. Data Augmentation Based on Generative Adversarial Network with Mixed Attention Mechanism. *Electronics* **2022**, *11*, 1718. [CrossRef]

49. Wang, M.; Su, T.; Chen, S.; Yang, W.; Liu, J.; Wang, Z. Automatic Model-Based Dataset Generation for High-Level Vision Tasks of Autonomous Driving in Haze Weather. *IEEE Trans. Ind. Inform.* **2023**, *19*, 9071–9081. [CrossRef]
50. Sai, S.; Mittal, U.; Chamola, V. DMDAT: Diffusion Model-Based Data Augmentation Technique for Vision-Based Accident Detection in Vehicular Networks. *IEEE Trans. Veh. Technol.* **2025**, *74*, 2241–2250. [CrossRef]
51. Li, B.; Liu, P.; Fu, L.; Li, J.; Fang, J.; Xu, Z.; Yu, H. Enhancing vehicle Re-identification by pair-flexible pose guided vehicle image synthesis. *Green Energy Intell. Transp.* **2025**, *4*, 100269. [CrossRef]
52. Ji, H.; Tian, L.; Wang, J.; Yao, Y.; Wang, J. Face Desensitization for Autonomous Driving Based on Identity De-Identification of Generative Adversarial Networks. *Electronics* **2025**, *14*, 2843. [CrossRef]
53. Cortes, A.; Rodriguez, C.; Velez, G.; Barandiaran, J.; Nieto, M. Analysis of Classifier Training on Synthetic Data for Cross-Domain Datasets. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 190–199. [CrossRef]
54. Dargahi Nobari, K.; Bertram, T. A multimodal driver monitoring benchmark dataset for driver modeling in assisted driving automation. *Sci. Data* **2024**, *11*, 327. [CrossRef] [PubMed]
55. Qin, X.; Zhang, Z.; Huang, C.; Dehghan, M.; Zaiane, O.R.; Jagersand, M. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognit.* **2020**, *106*, 107404. [CrossRef]
56. Nousias, S.; Pikoulis, E.V.; Mavrokefalidis, C.; Lalos, A.S. Accelerating Deep Neural Networks for Efficient Scene Understanding in Multi-Modal Automotive Applications. *IEEE Access* **2023**, *11*, 28208–28221. [CrossRef]
57. Hu, Y.; Lu, M.; Xie, C.; Lu, X. Driver Drowsiness Recognition via 3D Conditional GAN and Two-Level Attention Bi-LSTM. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 4755–4768. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.