DualScope: Capturing Critical Spatial and Temporal Cues for Distracted Driving Activity Recognition

Zhijie Qiu[1*], **Shuiabo Li**[1*†], **Laixin Zhang**[2], **Xuming Hu**[1,3‡], **Wei Ma**[2]
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]Beijing University of Technology
[3]The Hong Kong University of Science and Technology
{zqiu183, sli270}@connect.hkust-gz.edu.cn, s202374172@emails.bjut.edu.cn,
xuminghu@hkust-gz.edu.cn, mawei@bjut.edu.cn

Abstract

Accurately recognizing distracted driving activities in real-world scenarios is essential for improving road and pedestrian safety. However, existing approaches are prone to attend to irrelevant scene contexts and are susceptible to interference from redundant frames, compromising their robustness in complex driving environments. To overcome these limitations, we propose DualScope, a novel framework that captures behaviorally critical information from spatial and temporal perspectives. In the Spatial Scope, we introduce a Synergistic Behavior-Centric Distillation mechanism that leverages two key information sources: (1) position-aware knowledge derived from the SAM model that enhances the perception of critical regions and their semantic interaction structures and (2) fine-grained visual details obtained from cropped key regions that improve the model's ability to capture detailed patterns in behavior-relevant areas. In the Temporal Scope, we present the Saliency-Aware Fine-to-Coarse Temporal Modeling module comprising three components: a Fine-Grained Motion Encoder for capturing local inter-frame dependencies, a Dynamic Difference Extractor for extracting salient motion dynamics, and a Saliency-Aware Temporal Pyramid Mamba for integrating these features to enable multiscale temporal modeling. This design effectively captures short-term motions and long-term behavioral patterns. Furthermore, incorporating salient dynamics enhances the model's focus on substantial behavioral variations. Extensive experiments on seven publicly available distracted driving activity recognition datasets demonstrate that DualScope consistently outperforms state-of-the-art methods, validating its effectiveness in capturing behavioral cues across spatial and temporal dimensions.

Introduction

Distracted driving activity recognition (DDAR) aims to identify driver behaviors that lead to distraction in real-world driving scenarios, such as making phone calls or yawning. According to a report (National Center for Statistics and Analysis 2025), over 324,000 individuals in the United States were injured in motor vehicle accidents involving distracted driving in 2023. Consequently, developing reliable and effective DDAR models for driver monitoring and assistance systems is crucial to reducing distraction-related traffic accidents and enhancing overall road safety.

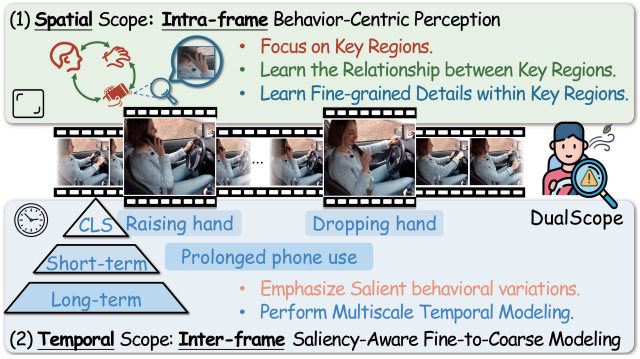

Figure 1: **Illustration of DualScope.** The Spatial Scope attends to key regions, inter-region relationships, and their internal fine-grained details within each frame. The Temporal Scope emphasizes salient behavioral variations and captures multiscale motion patterns across frames.

Existing DDAR methods focus on extracting discriminative features within individual frames. For example, (Zhao et al. 2021) employed Class Activation Maps generated by a convolutional neural network (CNN) to localize salient regions. This method emphasizes highly activated regions in pretrained networks rather than semantically relevant areas, limiting their ability to understand complex action semantics. Recently, vision foundation models (VFMs) represented by CLIP (Radford et al. 2021) have demonstrated strong generalization and transfer abilities through large-scale cross-modal pretraining. DriveCLIP (Hasan et al. 2024) applies CLIP to DDAR by encoding individual frames with a pretrained visual encoder, achieving promising results. However, although this encoding method can capture driving scene semantics, it cannot perceive behavior semantics limiting its ability to distinguish similar behaviors. To overcome this issue, (Chang et al. 2025) used human pose estimation to help distinguish subtle actions through skele-

*These authors contributed equally.
†Project lead.
‡Corresponding author.

tal motion. However, their approach focuses solely on body part movements and overlooks fine visual details within relevant regions; as a result, differentiating behaviors with similar motion involving different objects becomes difficult.

Temporal modeling is another major challenge in DDAR. DriveCLIP adopts a voting mechanism to obtain video-level outcomes but fails to aggregate holistic temporal information, making it challenging to recognize long-duration distraction behaviors. (Moslemi, Azmi, and Soryani 2019) utilized 3DCNN combined with RGB and optical flow information to model the spatiotemporal features of driving behaviors. However, this approache is susceptible to redundant frame information, leading to low recognition accuracy for brief and subtle distraction behaviors.

These findings highlight two core challenges in DDAR: (1) the need to accurately perceive behavior-relevant key information within frames, including the position, interaction, and visual details of key regions such as hands and heads, and (2) the need to effectively aggregate complete temporal information across frames while emphasizing dynamic changes that signal distracted behavior. To address these challenges, we propose DualScope, a framework that emphasizes the critical information in distracted driving behaviors from spatial and temporal perspectives. Specifically, DualScope consists of two primary components: the Spatial Scope and the Temporal Scope.

In the Spatial Scope, we employ a Synergistic Behavior-Centric Distillation (SBCD) mechanism that trains a Behavior-Centric Visual Encoder (BCVE) for frame-level encoding to enhance the model's ability to capture behavior-relevant features. SBCD contains the Cross-Region Contextual Distillation module (CRCD) and the Intra-Region Discriminative Distillation module (IRDD). CRCD aligns features of the original image with those enhanced by integrating position-aware knowledge from SAM (Kirillov et al. 2023) via a cross-attention mechanism, thereby strengthening the model's perception of key region positions and semantic interactions between regions. IRDD performs self-distillation to align the pooled regional features from the original image with the corresponding detail-enhanced featuress extracted from cropped key regions, improving the model's fine-grained understanding of intra-region details. CRCD and IRDD synergistically optimize BCVE, enhancing behavioral understanding through spatial-semantic context and fine-grained regional details. To resolve conflicts arising from their differing optimization objectives, we employ Gradient Surgery, which utilizes an orthogonal projection strategy during gradient updates. During training, to address the domain gap between pretraining data and driving scenes and overcome insufficient spatial modeling, we insert the Task-Aware Adapter and the Spatial Inductive Adapter into different layers for efficient adaptation.

In the Temporal Scope, we propose the Saliency-Aware Fine-to-Coarse Temporal Modeling module to guide the model's attention toward salient dynamics in behavior and effectively capture behavioral patterns across multiple temporal scales. For the video features encoded by BCVE, the Fine-Grained Motion Encoder models local inter-frame relationships via convolution. Simultaneously, the Dynamic Difference Extractor performs temporal saliency mining through feature-level central differencing to highlight remarkable behavior-related changes. We also introduce the Saliency-Aware Temporal Pyramid Mamba (SATPM) to further model distraction behaviors of varying durations. Salient dynamic information is integrated into SATPM blocks across multiple scales via attention mechanisms, enabling robust multiscale temporal modeling and enhancing the model's focus on critical behavioral changes.

Our contributions are summarized as follows:

- We propose DualScope, a novel framework for the DDAR task that captures distraction-related key information from spatial and temporal perspectives, leading to improved recognition performance.
- An innovative SBCD mechanism is presented to integrate the complementary strengths of CRCD and IRDD modules, thereby improving the model's understanding of key region positions, semantic interactions, and fine-grained intra-region details.
- We present the Saliency-Aware Fine-to-Coarse Temporal Modeling module, which integrates salient dynamics at multiple temporal scales, underscoring the value of salient dynamics and multiscale modeling in DDAR.
- Extensive experiments show that DualScope outperforms existing methods across seven public datasets.

Related Work

Distracted Driving Activity Recognition Methods. Many methods have been proposed for distracted driving recognition, and they can be broadly divided into conventional deep learning and VFM-based methods. (Moslemi, Azmi, and Soryani 2019) employs a two-stream 3DCNN architecture that combines RGB and optical flow for spatiotemporal modeling. (Yang et al. 2023b) applied a weakly supervised framework on the basis of Swin Transformer. (Akdag et al. 2023) integrates SlowFast and 2Dpose information for improved accuracy. (Pan et al. 2021) utilized LSTM for temporal modeling. (Zhang et al. 2024) improved performance by modeling the temporal context via VideoMAE. (Pizarro et al. 2024; Chang et al. 2025) employed VideoMAE and posture information to recognize distracted behaviors, enhancing detection performance. (Hasan et al. 2024) proposed a CLIP-based framework that leverages zero-shot inference and fine-tuning to detect distracted behaviors. However, most of these methods focus on capturing spatial information within frames and tend to disregard temporally salient clues across frames. This limited scope hinders their effectiveness in complex real-world scenarios. By contrast, our approach captures key behavioral cues by simultaneously attending to spatially behavior-relevant regions and temporally salient dynamics, resulting in the accurate recognition of distracted driving behaviors.

Vision Foundation Models. VFMs such as CLIP have demonstrated strong generalization in various downstream tasks, including synthetic image detection (Li et al. 2021,

2024b, 2025c; Zhang et al. 2025) and semantic segmentation (Li et al. 2025a,b). These applications highlight the broad applicability of VFMs, but their representations remain global. Recent advances in multimodal representation learning (Li et al. 2025e,d; Hao et al. 2025) further improve global and modality-level features but still do not provide region-level alignment. RegionCLIP (Zhong et al. 2022) introduces pseudo region–text pairs, and CLIPSelf (Wu et al. 2023) applies self-distillation to address this problem. Although effective, these approaches rely on object-centric region definitions and cannot capture behavior-critical areas in DDAR, such as the driver's head and hands, nor the semantic interactions among these regions. To address these limitations, we propose an SBCD mechanism that leverages the spatial priors from SAM and detailed information from region crops to enhance behavior-relevant perception.

State Space Models. State Space Models (SSMs), such as S4 (Yu et al. 2020), efficiently model long-range dependencies, with Mamba (Gu and Dao 2023) further enhancing temporal modeling through data-dependent selection and hardware-efficient algorithms. Although Mamba has been applied to various vision tasks (Li et al. 2024a,c), its effectiveness in DDAR is hindered by redundant frames and the difficulty of handling behaviors with varying durations. To address these issues, we propose the Saliency-Aware Fine-to-Coarse Temporal Modeling module, which leverages Mamba for multiscale temporal modeling and incorporates behavior-relevant salient dynamics to guide the model's attention toward remarkable motion changes. This design enhances the model's ability to recognize distracted behaviors across diverse temporal patterns.

Methodology

As illustrated in Figure 3, DualScope comprises two main components: the Spatial Scope and the Temporal Scope. The Spatial Scope employs a Synergistic Behavior-Centric Distillation mechanism to train a Behavior-Centric Visual Encoder, with SBCD comprising a Cross-Region Contextual Distillation module and an Intra-Region Discriminative Distillation module. Figure 2 shows their structure and training methodology. The Temporal Scope features the Saliency-Aware Fine-to-Coarse Temporal Modeling module that incorporates the Fine-Grained Motion Encoder, Dynamic Difference Extractor, and Saliency-Aware Temporal Pyramid Mamba. Each of these components is described in detail in the following sections.

Spatial Scope

In the DDAR task, the driver's head and hands are the primary regions associated with distracted behaviors (Xing et al. 2017), and their spatial locations, interactions, and internal details often serve as critical cues for identifying distraction. To enable ViT-based encoders to capture these cues, we employ the SBCD mechanism to train a BCVE. The encoder is trained on DriPE (Guesdon, Crispim-Junior, and Tougne 2021), a pose estimation dataset comprising N in-vehicle images captured in naturalistic driving scenarios, which closely matches our target DDAR setting while avoiding any prior knowledge or biased data distributions related to distraction labels.

Cross-Region Contextual Distillation Module. CRCD is designed to model the spatial positions of key regions and their semantic interaction structures. For each image i_j, we use RTMPose (Jiang et al. 2023) to extract 133 body keypoints and select the coordinates of the head, left hand, and right hand to generate three bounding boxes r_j^k, which form a paired dataset $\{i_j, r_j^k\}_{j=1,\dots,N}^{k=1,2,3}$.

Following (Yang et al. 2024), we input i_j and r_j^k into SAM model, extracting the first output tokens before the 3-layer MLP. These tokens are processed by a Key Region Contextual Encoder, which employs a self-attention mechanism to capture spatial contexts across regions and applies linear transformation to map them into the image-level feature space, resulting in position-aware tokens $\boldsymbol{p}_j \in \mathbb{R}^{3\times D}$.

Concurrently, i_j is encoded by a ViT-L/14 visual encoder to extract visual features $\boldsymbol{v}_j \in \mathbb{R}^{257\times D}$. We then apply a cross-attention mechanism, where $\boldsymbol{v}_j$ serves as the queries, and $\boldsymbol{p}_j$ acts as keys and values. This process yields enhanced features $\hat{\boldsymbol{v}}_j$ that incorporate information about key region positions and semantic interactions between regions. After discarding the $[CLS]$ tokens, $\boldsymbol{v}_j$ and $\hat{\boldsymbol{v}}_j$ are reshaped into spatial feature maps $\boldsymbol{f}_j$ and $\hat{\boldsymbol{f}}_j$ of the shape $16 \times 16 \times D$, representing the visual dense feature and context-enhanced features, respectively. This operation recovers the spatial layout of the image patches to support fine-grained region-wise alignment, which is enforced through a cosine similarity loss as follows:

$$\mathcal{L}_{\text{CRCD}} = \frac{1}{N}\sum_{j=1}^{N}\left(1 - \frac{\boldsymbol{f}_j \cdot \hat{\boldsymbol{f}}_j}{\|\boldsymbol{f}_j\| \cdot \left\|\hat{\boldsymbol{f}}_j\right\|}\right). \tag{1}$$

During training, the keypoints predicted by RTMPose are used only to provide weak pose cues around behavior-relevant regions. Notably, some noisy keypoints caused by occlusion or challenging viewpoints naturally act as a regularizer, improving robustness under occlusion and low-light conditions. During inference, our framework remains fully keypoint-free, and the pretrained BCVE directly processes frames without requiring any pose estimation.

Intra-Region Discriminative Distillation Module. The IRDD module enhances the model's ability to perceive fine-grained details within regions. We extract regional feature $\boldsymbol{x}_j^k \in \mathbb{R}^{1\times D}$ by applying RoIAlign (He et al. 2017) to visual dense feature $\boldsymbol{f}_j$ at location r_j^k. Moreover, we crop original image i_j in accordance with r_j^k. The cropped region i_j^k is then input to a fixed, pretrained teacher visual encoder with same initial weights, from which we extract the $[CLS]$ token as $\hat{\boldsymbol{x}}_j^k \in \mathbb{R}^{1\times D}$, serving as the detail-enhanced feature. Benefiting from the teacher encoder, $\hat{\boldsymbol{x}}_j^k$ captures rich fine-grained semantics and is a reliable target for supervising student feature $\boldsymbol{x}_j^k$. To enhance fine-grained detail learning, we maximize the cosine similarity between $\boldsymbol{x}_j^k$ and $\hat{\boldsymbol{x}}_j^k$ through

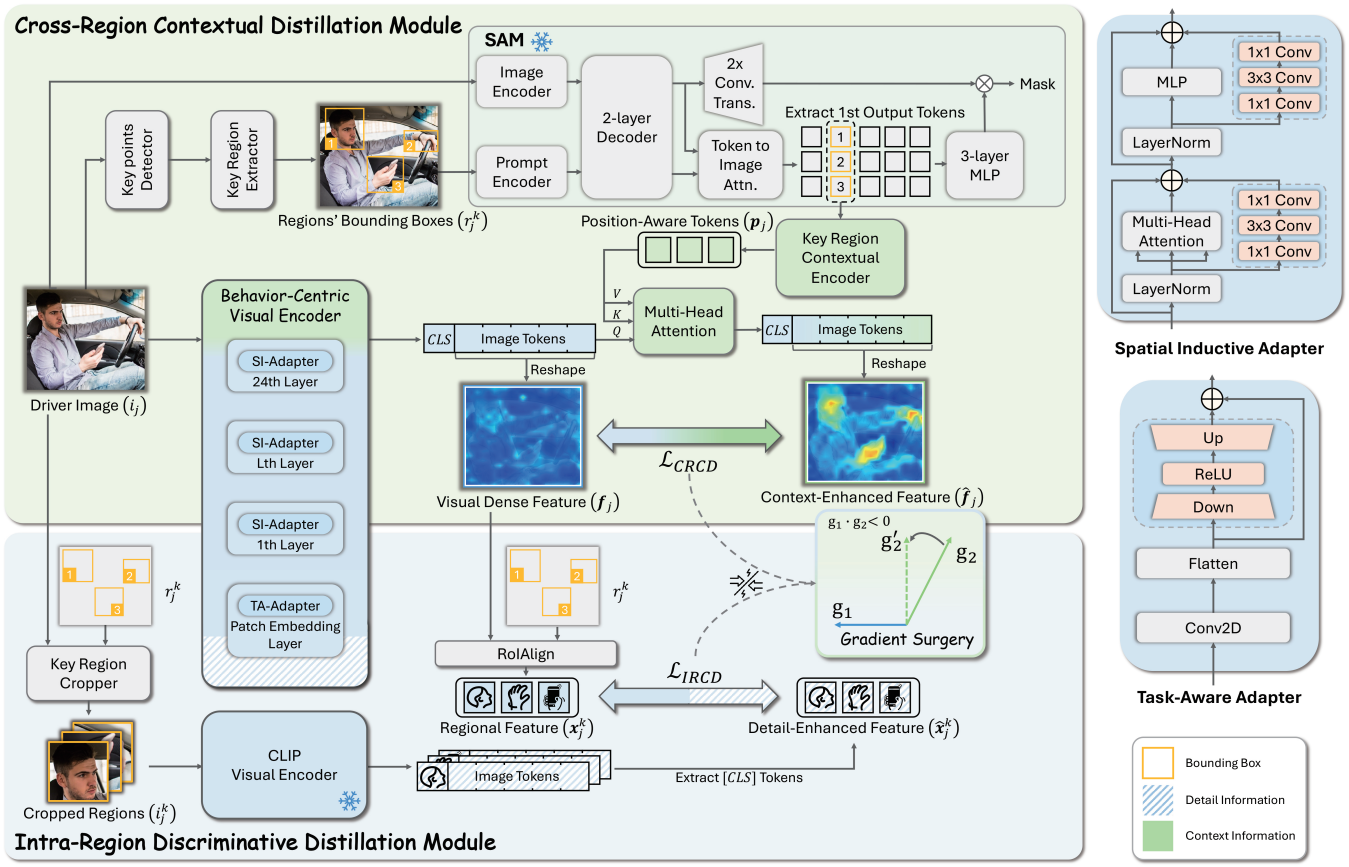

Figure 2: **Illustration of the Synergistic Behavior-Centric Distillation mechanism.** It consists of the Cross-Region Contextual Distillation module and the Intra-Region Discriminative Distillation module, which jointly enhance the behavioral understanding of the Behavior-Centric Visual Encoder (BCVE). Gradient Surgery is applied to mitigate the conflicts between their optimization objectives. To adapt BCVE effectively to driving-specific scenarios, the Task-Aware Adapter and the Spatial Inductive Adapter are inserted into different layers.

a self-distillation loss, which is defined as

$$\mathcal{L}_{\text{IRDD}} = \frac{1}{3N} \sum_{j=1}^{N} \sum_{k=1}^{3} \left(1 - \frac{\boldsymbol{x}_j^k \cdot \hat{\boldsymbol{x}}_j^k}{\left\| \boldsymbol{x}_j^k \right\| \cdot \left\| \hat{\boldsymbol{x}}_j^k \right\|} \right). \quad (2)$$

Conflict Mitigation Method. Although the CRCD and IRDD modules synergistically enhance behavioral understanding, their distinct optimization objectives may cause gradient conflicts during joint training. To resolve this issue, we apply Gradient Surgery (Yu et al. 2020), where $\mathbf{g}_1$ and $\mathbf{g}_2$ denote the gradients of the CRCD and IRDD modules, respectively. A conflict is detected when $\mathbf{g}_1 \cdot \mathbf{g}_2 < 0$, in which case each gradient is projected onto the normal plane of the other to remove the conflicting component: $\mathbf{g}_1' = \mathbf{g}_1 - \frac{\mathbf{g}_1 \cdot \mathbf{g}_2}{\|\mathbf{g}_2\|^2} \mathbf{g}_2$, $\mathbf{g}_2' = \mathbf{g}_2 - \frac{\mathbf{g}_1 \cdot \mathbf{g}_2}{\|\mathbf{g}_1\|^2} \mathbf{g}_1$. The final parameter update is computed as $\Delta\theta = \mathbf{g}_1' + \mathbf{g}_2'$. When no conflict is detected ($\mathbf{g}_1 \cdot \mathbf{g}_2 \geq 0$), the gradients are unaltered.

Parameter-Efficient Fine-Tuning. We employ a parameter-efficient fine-tuning (PEFT) strategy to enhance the model's perception of key regions while preserving its representational capacity. Inspired by (Jie et al. 2024; Pei, Huang, and Xu 2025), we insert the Task-Aware Adapter (TA-Adapter) into the Patch Embedding layer and the Spatial Inductive Adapter (SI-Adapter) into all Transformer layers, as depicted in Figure 2.

TA-Adapter uses a linear bottleneck structure to alleviate the domain gap between the pretraining data and driving scenarios. For input feature $\boldsymbol{t} \in \mathbb{R}^{n \times D}$, the adapter outputs $\boldsymbol{t}_{\text{out}} = \boldsymbol{t} + \text{ReLU}(\boldsymbol{t} \cdot \mathbf{W}_{\text{down}}) \cdot \mathbf{W}_{\text{up}}$, where $\mathbf{W}_{\text{up}} \in \mathbb{R}^{s \times D}$ and $\mathbf{W}_{\text{down}} \in \mathbb{R}^{D \times s}$, s denotes the bottleneck feature size, and n is the number of patches.

SI-Adapter is a convolutional bottleneck block inserted parallel to the Multi-Head Attention and MLP layers within each Transformer layer. It strengthens local spatial modeling through the inductive bias of convolutional layers. The adapter employs a $1\times1, 3\times3, 1\times1$ convolution sequence using GELU activations between convolution layers. Channel dimension D is reduced to h, subjected to spatial modeling, and restored to its original size. Before SI-Adapter is applied, image tokens are reshaped into a 2D spatial map; the $[CLS]$ token is treated as a 1×1 feature map and handled separately. Outputs are flattened and concatenated.

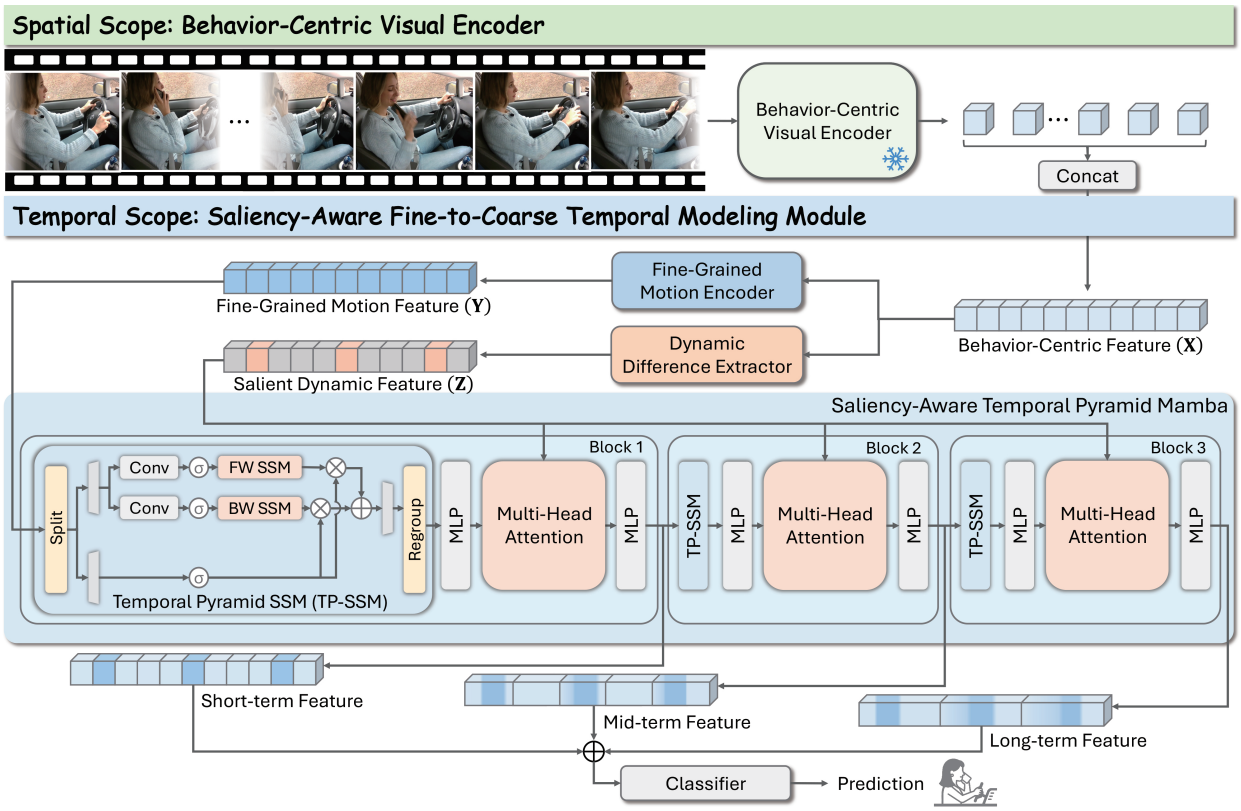

Figure 3: **Overall architecture of DualScope.** The Behavior-Centric Visual Encoder encodes each video frame and concatenates the resulting features, and the Fine-Grained Motion Encoder and Dynamic Difference Extractor model interframe relationships and highlight salient dynamics, respectively. Saliency-Aware Temporal Pyramid Mamba emphasizes salient behavioral variations and performs multiscale modeling. Finally, a summation operation aggregates multiscale features, and the classifier outputs the prediction.

Learning Objective. We introduce two loss functions. $\mathcal{L}_{\text{CRCD}}$ guides the model to learn key region positions and semantic interactions. $\mathcal{L}_{\text{IRDD}}$ improves the perception of fine-grained details within key regions. The two objectives jointly encourage the model to focus on distraction-relevant cues, and Gradient Surgery is applied to mitigate potential conflicts between their optimization objectives. The overall training objective is defined as $\mathcal{L} = \mathcal{L}_{\text{CRCD}} + \mathcal{L}_{\text{IRDD}}$.

Temporal Scope

As displayed in Figure 3, we propose a Saliency-Aware Fine-to-Coarse Temporal Modeling module that captures behavior patterns at different temporal resolutions and highlights salient dynamic changes. The pretrained BCVE encodes each frame and concatenates them, yielding a behavior-centric feature $\mathbf{X}$. Then, $\mathbf{X}$ is processed by a Fine-Grained Motion Encoder (FGME) and a Dynamic Difference Extractor (DDE) separately, generating features $\mathbf{Y}$ and $\mathbf{Z}$, which are input to the Saliency-Aware Temporal Pyramid Mamba (SATPM) for multi-scale modeling. Afterward, the classifier performs distracted behavior recognition.

Fine-Grained Motion Encoder. In DDAR, distracted driving behaviors involve subtle motion changes, such as hand raising and head turning, across consecutive frames within very short temporal windows. FGME models the frame-to-frame context to enhance local perception. For the independently modeled per-frame feature $\mathbf{X}$, FGME employs a 1D convolution with a kernel size of 3, a stride of 1, and a padding of 1 to aggregate local information. Subsequently, linear transformation is applied to unify the feature dimensions, resulting in fine-grained motion feature $\mathbf{Y} \in \mathbb{R}^{T \times D}$.

Dynamic Difference Extractor. In lengthy video frame sequences, focusing on key frames that contain remarkable behavioral changes (e.g., "looking down") is crucial for recognizing distraction behaviors. We introduce DDE for temporal saliency mining via differencing, which emphasizes salient interframe changes and suppresses redundant temporal information. Given feature sequence $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T]$, DDE computes central differences as $\boldsymbol{z}_t = \frac{1}{2}(\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t-1})$ for $t = 2, \ldots, T-1$, with boundary cases $\boldsymbol{z}_1 = \boldsymbol{x}_2 - \boldsymbol{x}_1$ and $\boldsymbol{z}_T = \boldsymbol{x}_T - \boldsymbol{x}_{T-1}$. Subsequently, linear transformation is applied to adjust the dimension,yielding the salient dynamic feature $\mathbf{Z} \in \mathbb{R}^{T \times D}$.

Saliency-Aware Temporal Pyramid Mamba. To effectively model distraction behaviors that exhibit diverse tem-

Method	Processing Type†	Video-based					Image-based	
		SAM-DD	SynDD1	AIDE	DMD	3MDAD	AUC	StateFarm
ResNet18 (He et al. 2016)	Image-only	95.91	47.55	66.52	85.95	-	95.54	82.68
VGG16 (Simonyan and Zisserman 2014)	Image-only	96.86	12.56	65.48	56.15	68.12	96.18	-
ShuffleNet-v2 (Ma et al. 2018)	Image-only	95.56	-	64.04	84.80	-	94.38	-
MobileNet-v2 (Sandler et al. 2018)	Image-only	94.78	36.10	61.74	93.90	65.13	94.74	80.04
MIFI (Kuang et al. 2023)	Video-only	96.98	80.43	66.17	93.20	83.70	-	-
DriveCLIP (Hasan et al. 2024)	Image&Video	97.86	81.85	66.01	98.44	83.13	96.58	83.15
DualScope	Image&Video	**98.56**	**89.66**	**79.64**	**98.63**	**87.72**	**96.86**	**86.85**

Table 1: **Performance comparison with state-of-the-art methods.** We report top-1 accuracy (%) across all datasets. The best results are **bolded**, while the second-best are underlined. Processing Type†: "Image-only" methods process video frame-by-frame with voting.

poral durations and salient motion changes, we design SATPM. SATPM consists of multiple consecutively stacked blocks, where each block gradually expands the temporal receptive field and incorporates salient dynamic cues extracted by DDE. Each SATPM block takes the output from the previous block $\mathbf{Y}_{b-1} \in \mathbb{R}^{T\times D}$ as its input, and the input of first block ($\mathbf{Y}_0$) is initialized using the fine-grained motion feature $\mathbf{Y}$ generated by FGME. Inspired by (Pei, Huang, and Xu 2025), we introduce Temporal Pyramid SSM (TP-SSM) to control the temporal resolution through a step size p, which is set to b for the b-th block. Specifically, $\mathbf{Y}_{b-1}$ is split into p interleaved, nonoverlapping subsequences:

$$\{\mathbf{Y}_{b-1}^{(i)}\}_{i=1}^{p} = \mathrm{Split}(\mathbf{Y}_{b-1}), \quad \mathbf{Y}_{b-1}^{(i)} \in \mathbb{R}^{\frac{T}{p}\times D}.$$

For example, when $p = 2$, the sequence $[t_0, t_1, \ldots, t_{2N}]$ is divided into $[t_0, t_2, \ldots, t_{2N}]$ and $[t_1, t_3, \ldots, t_{2N-1}]$. Each subsequence is independently modeled by BiMamba (Zhu et al. 2024),

$$\tilde{\mathbf{Y}}_b^{(i)} = \mathrm{BiMamba}(\mathbf{Y}_{b-1}^{(i)}), \quad i = 1, \ldots, p.$$

after which the processed subsequences are regrouped in their original temporal order to restore full-length temporal structure:

$$\hat{\mathbf{Y}}_b = \mathrm{Regroup}\big(\{\tilde{\mathbf{Y}}_b^{(i)}\}_{i=1}^{p}\big).$$

To inject salient dynamic cues, the salient feature $\mathbf{Z}$ obtained from DDE is used as queries, while $\hat{\mathbf{Y}}_b$ serves as keys and values. The cross-attention refinement is computed as

$$\mathbf{Y}_b = \mathrm{MLP}\left(\mathrm{Softmax}\left(\frac{\mathbf{Z}\hat{\mathbf{Y}}_b^{\top}}{\sqrt{D}}\right)\mathrm{MLP}(\hat{\mathbf{Y}}_b)\right).$$

By progressively increasing step size p across blocks, SATPM expands the temporal receptive field from fine to coarse, enabling comprehensive multiscale temporal modeling while emphasizing salient motion changes. This hierarchical structure substantially enhances the ability to recognize distraction behaviors that involve complex and variable temporal patterns.

Classifier. The classifier is designed to categorize input features. A simple summation operation effectively aggregates multiscale features from different SATPM blocks. We then apply temporal average pooling and perform linear projection into the distraction behavior category space. For optimization, we use cross-entropy loss as our objective.

Experiments

Experimental Settings

Dataset. To comprehensively evaluate the performance of our model on the DDAR task, we conduct experiments on five video datasets, namely, SAM-DD (Yang et al. 2023b), SynDD1 (Rahman et al. 2023), AIDE (Yang et al. 2023a), DMD (Ortega et al. 2020), and 3MDAD (Jegham et al. 2020), and two image datasets, namely, StateFarm (Montoya et al. 2016) and AUC-v1 (Eraqi et al. 2019). Following (Kuang et al. 2023; Hasan et al. 2024), we use side-view data for SAM-DD, dashboard-view data for SynDD1, and daytime data for 3MDAD. These datasets naturally cover a broad range of real-world challenges. In particular, several video datasets (e.g., SynDD1, AIDE, and 3MDAD) contain short-term occlusions, low-light environments, and complex viewing angles, making the DDAR task realistic and substantially challenging.

Evaluation Metric. This task is formulated as a multi-class video classification problem, and we follow previous studies (Hasan et al. 2024; Yang et al. 2023b; Kose et al. 2019; Moslemi, Azmi, and Soryani 2019) by adopting top-1 accuracy (Acc) as the evaluation metric. To ensure a fair, reliable evaluation, we employ the cross-validation and subject-level separation protocols from (Hasan et al. 2024; Chai et al. 2024). In particular, to avoid driver-specific bias and the inflated accuracy issue, we maintain same strict separation between training and testing subjects.

Implementation Details. During BCVE pretraining, the ViT-L/14 backbone is initialized with pretrained weights and with adapter parameters $s = 32$ and $h = 8$. Only the parameters in the adapters, key region contextual encoder, and multi-head attention are updated. The learning rate is set to $1e{-}4$ and adjusted via cosine annealing. In DualScope training, projected feature dimension D is 256, and three SATPM blocks are used. MLP is a two-layer feedforward block with an intermediate size of 1024 and GELU activation. We employ the AdamW optimizer with an initial learning rate of $1e-5$, a weight decay of 0.1, and cosine annealing scheduling with a 10-epoch warm-up. DualScope is implemented using PyTorch and trained on an NVIDIA A800 GPU.

Comparison with State-of-the-Art Methods

Quantitative Results. Table 1 compares DualScope with traditional deep learning models, namely, ResNet18, VGG16, ShuffleNet-v2, MobileNet-v2, and state-of-the-art DDAR methods, namely, MIFI and DriveCLIP. Evaluation results across seven distracted driving datasets show that DualScope achieves the higheset top-1 accuracy among the compared methods.

Compared with methods lacking adequate temporal modeling, DualScope achieves multiscale temporal modeling while focusing on behavioral salient changes, resulting in substantially improved recognition performance. DriveCLIP employs a simple majority voting mechanism to aggregate predictions and does not fully capture temporal information. As a result, it performs poorly on the AIDE dataset, which requires full temporal understanding of short clips. MIFI, affected by redundant frame noise, performs poorly on the long-duration SynDD1 dataset, which contains substantial redundancy. By contrast, DualScope consistently achieves top performance across multiple video datasets. On the image datasets, the SBCD mechanism enables DualScope to attend to key distraction-related regions precisely, enhancing fine-grained driver behavior recognition and ensuring superior performance.

Method	SynDD1	StateFarm	DMD	SAM-DD
DriveCLIP	54.00	58.22	48.12	66.52
Ours	**54.52**	**59.32**	**48.28**	**66.87**

Table 2: **Zero-shot Generalization comparison.** We report top-1 accuracy (%) on four datasets.

Method	#Params. (M)	FLOPs (G)	FPS
ResNet18	11.2	1.8	75.54
VGG16	13.8	15.5	69.21
ShuffleNet-v2	7.4	0.6	60.25
MobileNet-v2	3.4	0.3	64.27
MIFI	24.6	223	14.70
DriveCLIP	9.67	3.54	6.90
Ours	12.36	1.39	22.32

Table 3: **Computational Efficiency comparison.** We report Model size, computational complexity, and FPS, following the same evaluation protocol in DriveCLIP.

Zero-Shot Generalization Evaluation To verify that our PEFT-based pretraining strategy preserves the inherent generalization ability of the model, we conduct a zero-shot evaluation following the same protocol as DriveCLIP. This experiment is designed to confirm that the BCVE, trained with the SBCD mechanism, retains the generalization capacity of the original backbone. As shown in Table 2, BCVE achieves comparable or slightly improved performance compared with the DriveCLIP baseline, indicating that the proposed PEFT design maintains cross-domain generalization while enhancing the model's awareness of critical spatial cues in driving scenes.

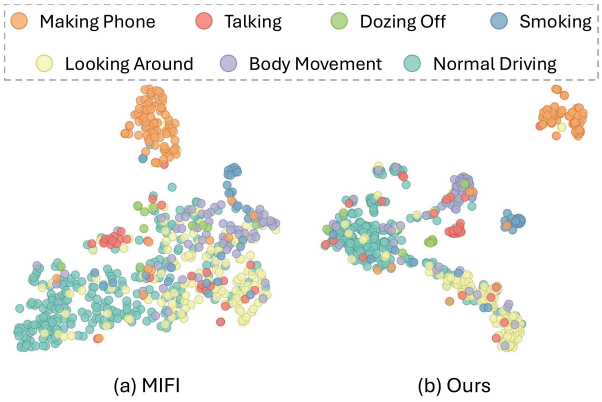

Figure 4: **Embedding space visualization.** t-SNE plots of MIFI and DualScope on the AIDE dataset. MIFI produces scattered and highly overlapping clusters across behavior categories, whereas DualScope forms compact clusters with well-separated boundaries between categories, demonstrating stronger discriminative capability.

Computational Efficiency. To evaluate the computational efficiency of DualScope in realistic deployment settings, Table 3 presents the model size, computational complexity, and end-to-end inference throughput measured on an RTX 3090 GPU. DualScope contains 12.36M parameters and requires only 1.39 GFLOPs, making it substantially lighter than high-overhead video models, such as MIFI (24.6M, 223 GFLOPs). The image-based architectures, such as ResNet18 and VGG16 achieve higher FPS values (75.54 and 69.21, respectively) because they operate on single frames, but their lack of temporal modeling limits their effectiveness on video-level DDAR benchmarks. DualScope attains 22.32 FPS, clearly surpassing video-based baselines, including MIFI (14.70 FPS) and DriveCLIP (6.90 FPS), while simultaneously achieving highest recognition accuracy across all the evaluated datasets. These results demonstrate that DualScope obtains an effective balance between temporal modeling capability and computational efficiency, enabling real-time inference and supporting practical deployment in resource-constrained driving environments.

Qualitative Results. As shown in Figure 4, we use t-SNE (Maaten and Hinton 2008) to visualize features from the AIDE dataset and qualitatively compare the discriminative abilities of DualScope and the baseline MIFI in modeling driver behaviors. Each point denotes a video sample, color-coded by behavior category. Notably, the features learned by MIFI are distributed sparsely, exhibiting considerable overlap among different categories. This phenomenon is particularly apparent for *Looking Around* and *Body Movement*, where the categories are highly entangled. This observation implies that MIFI's learned representations may lack sufficient granularity.

By contrast, the features learned by DualScope exhibit a distinct and structured clustering pattern in the embedding

Component	Settings	Acc
Visual Encoder	ViT-L/14	87.23
	BCVE	**90.84**
Temporal Modeling	FGME only	79.34
	w/o DDE	87.49
	w/o FGME	89.76
	Full	**90.84**
Number of SATPM blocks	1	87.53
	2	89.36
	3	**90.84**
	4	86.42

Table 4: **Ablation study of the main components.** We report average top-1 accuracy (%) across five video datasets.

space, with well-defined decision boundaries between multiple categories. For instance, the behaviors such as *Normal Driving*, *Smoking*, and *Making Phone* are effectively clustered and separated. This observation indicates that DualScope can capture discriminative behavior features and effectively distinguish fine-grained driver behaviors. These findings further validate the effectiveness of our approach in capturing behavior-relevant cues in temporal and spatial dimensions.

Ablation Analysis

Model components. We conduct ablation experiments to assess the contribution of each component in DualScope across five video datasets, and the results are summarized in Table 4. Replacing BCVE with pretrained ViT-L/14 leads to a 3.61% performance drop, indicating that ViT-L/14 lacks the ability to perceive distraction-relevant regions, whereas BCVE benefits from behavior-centric distillation and captures discriminative cues.

For temporal modeling, retaining only FGME substantially reduces accuracy to 79.34%, showing that local temporal cues alone are insufficient. Removing DDE results in a 3.35% decrease, demonstrating that salient dynamics are crucial for suppressing redundant motion information. Removing FGME produces a small drop (1.08%), suggesting that FGME mainly provides complementary short-range motion cues. These observations verify that FGME, DDE, and SATPM jointly form an effective temporal modeling pipeline.

We further vary the number of SATPM blocks. Increasing the depth enlarges the temporal receptive field and improves performance up to three blocks, achieving the highest accuracy of 90.84%. Using four blocks leads to a noticeable decline, which is likely due to excessive long-range modeling that weakens short-term behavioral cues. Overall, three blocks provide ideal balance between temporal granularity and contextual coverage.

Effect of SBCD components. To assess the contribution of the SBCD mechanism, we analyze two key aspects: conflict mitigation, and adapter design in the visual encoder. Given that SBCD focuses solely on spatial modeling, we conduct experiments on two image datasets (StateFarm and AUC) to isolate spatial effects. The results are summarized in Table 5.

For conflict mitigation, we analyze the role of Gradient Surgery in resolving the optimization inconsistency between CRCD and IRDD. Without this method, performance degrades to 89.31%, indicating that the two spatial objectives impose conflicting gradients during joint optimization. Applying Gradient Surgery yields a 2.55% improvement (91.86%) by orthogonalizing the gradients from CRCD and IRDD, enabling stable and cooperative multiobjective learning. These results highlight the necessity of explicitly addressing gradient conflicts introduced by the complementary yet divergent spatial supervision from CRCD and IRDD.

For adapter design, we compare various PEFT strategies. Full fine-tuning provides limited gains (90.33%), whereas the PEFT baselines, namely, LoRA (Hu et al. 2022) and AdaptFormer (Chen et al. 2022), achieve 90.76% and 91.09%, respectively. Our task-specific adapters obtain higher accuracy with fewer parameters (0.93M vs. 1.57M/3.15M). Using only TA-Adapter yields 90.67% accuracy, and using only SI-Adapter produces 91.21%. Combining both results in excellent performance (91.86%), confirming the efficiency and effectiveness of our spatially tailored adapter design.

Component	Settings	Acc(%)
Conflict Mitigation	w/o Gradient Surgery	89.31
	Gradient Surgery	**91.86**
Adapter Design	Full Fine-tuning	90.33
	LoRA	90.76
	AdaptFormer	91.09
	TA-Adapter	90.67
	SI-Adapter	91.21
	TA+SI-Adapter	**91.86**

Table 5: **Ablation study on the SBCD components.** The results show average top-1 accuracy (%) across StateFarm and AUC.

Conclusion

In this paper, we propose DualScope, a novel framework for DDAR that models critical behavioral cues from spatial and temporal perspectives. The Spatial Scope employs the Synergistic Behavior-Centric Distillation mechanism to guide attention to key regions, capturing their positions, interactions, and fine-grained details. The Temporal Scope introduces the Saliency-Aware Fine-to-Coarse Temporal Modeling module and effectively captures various behavioral patterns and salient dynamic changes. Experiments on seven public datasets show that DualScope consistently outperforms existing methods, demonstrating strong recognition accuracy and generalization in complex driving scenarios.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.62506318); Guangdong Provincial Department of Education Project (No.2024KQNCX028);

CAAI-Ant Group Research Fund; Scientific Research Projects for the Higher-educational Institutions (No.2024312096), Education Bureau of Guangzhou Municipality; Guangzhou-HKUST(GZ) Joint Funding Program (No.2025A03J3957), Education Bureau of Guangzhou Municipality; the National Natural Science Foundation of China (No.62176010); the Beijing Natural Science Foundation (No.4252029); the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No.VRLAB 2024B02).

References

Akdag, E.; Zhu, Z.; Bondarev, E.; et al. 2023. Transformer-based fusion of 2D-pose and spatio-temporal embeddings for distracted driver action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5453–5462.

Chai, W.; Wang, J.; Chen, J.; Velipasalar, S.; and Sharma, A. 2024. Rethinking the evaluation of driver behavior analysis approaches. *IEEE Transactions on Intelligent Transportation Systems*, 25(8): 9958–9966.

Chang, Q.; Dai, W.; Shuai, Z.; Yu, L.; and Yue, Y. 2025. Spatial-Temporal Perception with Causal Inference for Naturalistic Driving Action Recognition. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Chen, S.; Ge, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; and Luo, P. 2022. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35: 16664–16678.

Eraqi, H. M.; Abouelnaga, Y.; Saad, M. H.; and Moustafa, M. N. 2019. Driver distraction identification with an ensemble of convolutional neural networks. *Journal of advanced transportation*, 2019(1): 4125865.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Guesdon, R.; Crispim-Junior, C.; and Tougne, L. 2021. Dripe: A dataset for human pose estimation in real-world driving settings. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2865–2874.

Hao, P.; Li, S.; Wang, H.; Kou, Z.; Zhang, J.; Yang, G.; and Zhu, L. 2025. Surgery-R1: Advancing Surgical-VQLA with Reasoning Multimodal Large Language Model via Reinforcement Learning. *arXiv preprint arXiv:2506.19469*.

Hasan, M. Z.; Chen, J.; Wang, J.; Rahman, M. S.; Joshi, A.; Velipasalar, S.; Hegde, C.; Sharma, A.; and Sarkar, S. 2024. Vision-language models can identify distracted driver behavior from naturalistic videos. *IEEE Transactions on Intelligent Transportation Systems*, 25(9): 11602–11616.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Jegham, I.; Khalifa, A. B.; Alouani, I.; and Mahjoub, M. A. 2020. A novel public dataset for multimodal multiview and multispectral driver distraction analysis: 3MDAD. *Signal Processing: Image Communication*, 88: 115960.

Jiang, T.; Lu, P.; Zhang, L.; Ma, N.; Han, R.; Lyu, C.; Li, Y.; and Chen, K. 2023. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*.

Jie, S.; Deng, Z.-H.; Chen, S.; and Jin, Z. 2024. Convolutional bypasses are better vision transformer adapters. In *ECAI 2024*, 202–209. IOS Press.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Kose, N.; Kopuklu, O.; Unnervik, A.; and Rigoll, G. 2019. Real-time driver state monitoring using a CNN based spatio-temporal approach. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 3236–3242. IEEE.

Kuang, J.; Li, W.; Li, F.; Zhang, J.; and Wu, Z. 2023. Mifi: Multi-camera feature integration for robust 3d distracted driver activity recognition. *IEEE Transactions on Intelligent Transportation Systems*, 25(1): 338–348.

Li, B.; Dong, H.; Zhang, D.; Zhao, Z.; Gao, J.; and Li, X. 2025a. Exploring Efficient Open-Vocabulary Segmentation in the Remote Sensing. *arXiv preprint arXiv:2509.12040*.

Li, B.; Zhang, D.; Zhao, Z.; Gao, J.; and Li, X. 2025b. Stitchfusion: Weaving any visual modalities to enhance multimodal semantic segmentation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 1308–1317.

Li, K.; Li, X.; Wang, Y.; He, Y.; Wang, Y.; Wang, L.; and Qiao, Y. 2024a. Videomamba: State space model for efficient video understanding. In *European conference on computer vision*, 237–255. Springer.

Li, S.; Ma, W.; Guo, J.; Xu, S.; Li, B.; and Zhang, X. 2024b. Unionformer: Unified-learning transformer with multi-view representation for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12523–12533.

Li, S.; Xing, Z.; Wang, H.; Hao, P.; Li, X.; Liu, Z.; and Zhu, L. 2025c. Toward Medical Deepfake Detection: A Comprehensive Dataset and Novel Method. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 626–637. Springer.

Li, S.; Xu, S.; Ma, W.; and Zong, Q. 2021. Image manipulation localization using attentional cross-domain CNN features. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9): 5614–5628.

Li, Y.; Cao, Y.; He, H.; Cheng, Q.; Fu, X.; Xiao, X.; Wang, T.; and Tang, R. 2025d. M²IV: Towards Efficient and Fine-grained Multimodal In-Context Learning via Representation Engineering. In *Second Conference on Language Modeling*.

Li, Y.; Yang, J.; Yun, T.; Feng, P.; Huang, J.; and Tang, R. 2025e. Taco: Enhancing multimodal in-context learning via task mapping-guided sequence configuration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 736–763.

Li, Z.; Zhao, M.; Yang, X.; Liu, Y.; Sheng, J.; Zeng, X.; Wang, T.; Wu, K.; and Jiang, Y.-G. 2024c. STNMamba: Mamba-based spatial-temporal normality learning for video anomaly detection. *arXiv preprint arXiv:2412.20084*.

Ma, N.; Zhang, X.; Zheng, H.-T.; and Sun, J. 2018. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, 116–131.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov): 2579–2605.

Montoya, A.; Holman, D.; SF_data_science; Smith, T.; and Kan, W. 2016. State Farm Distracted Driver Detection. https://kaggle.com/competitions/state-farm-distracted-driver-detection. Kaggle.

Moslemi, N.; Azmi, R.; and Soryani, M. 2019. Driver distraction recognition using 3D convolutional neural networks. In *2019 4th International Conference on Pattern Recognition and Image Analysis (IPRIA)*, 145–151. IEEE.

National Center for Statistics and Analysis. 2025. Distracted Driving in 2023. Research Note DOT HS 813 703, National Highway Traffic Safety Administration.

Ortega, J. D.; Kose, N.; Cañas, P.; Chao, M.-A.; Unnervik, A.; Nieto, M.; Otaegui, O.; and Salgado, L. 2020. Dmd: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis. In *European Conference on Computer Vision*, 387–405. Springer.

Pan, C.; Cao, H.; Zhang, W.; Song, X.; and Li, M. 2021. Driver activity recognition using spatial-temporal graph convolutional LSTM networks with attention mechanism. *IET Intelligent Transport Systems*, 15(2): 297–307.

Pei, X.; Huang, T.; and Xu, C. 2025. Efficientvmamba: Atrous selective scan for light weight visual mamba. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6, 6443–6451.

Pizarro, R.; Valle, R.; Bergasa, L. M.; Buenaposada, J. M.; and Baumela, L. 2024. Pose-guided multi-task video transformer for driver action recognition. *arXiv preprint arXiv:2407.13750*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rahman, M. S.; Venkatachalapathy, A.; Sharma, A.; Wang, J.; Gursoy, S. V.; Anastasiu, D.; and Wang, S. 2023. Synthetic distracted driving (syndd1) dataset for analyzing distracted behaviors and various gaze zones of a driver. *Data in brief*, 46: 108793.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Wu, S.; Zhang, W.; Xu, L.; Jin, S.; Li, X.; Liu, W.; and Loy, C. C. 2023. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. *arXiv preprint arXiv:2310.01403*.

Xing, Y.; Lv, C.; Zhang, Z.; Wang, H.; Na, X.; Cao, D.; Velenis, E.; and Wang, F.-Y. 2017. Identification and analysis of driver postures for in-vehicle driving activities and secondary tasks recognition. *IEEE Transactions on Computational Social Systems*, 5(1): 95–108.

Yang, D.; Huang, S.; Xu, Z.; Li, Z.; Wang, S.; Li, M.; Wang, Y.; Liu, Y.; Yang, K.; Chen, Z.; et al. 2023a. Aide: A vision-driven multi-view, multi-modal, multi-tasking dataset for assistive driving perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20459–20470.

Yang, H.; Liu, H.; Hu, Z.; Nguyen, A.-T.; Guerra, T.-M.; and Lv, C. 2023b. Quantitative identification of driver distraction: A weakly supervised contrastive learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 25(2): 2034–2045.

Yang, H.; Ma, C.; Wen, B.; Jiang, Y.; Yuan, Z.; and Zhu, X. 2024. Recognize any regions. *Advances in Neural Information Processing Systems*, 37: 51312–51332.

Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; and Finn, C. 2020. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836.

Zhang, L.; Li, S.; Ma, W.; and Zha, H. 2025. TrueMoE: Dual-Routing Mixture of Discriminative Experts for Synthetic Image Detection. *arXiv preprint arXiv:2509.15741*.

Zhang, T.; Wang, Q.; Dong, X.; Yu, W.; Sun, H.; Zhou, X.; Zhen, A.; Cui, S.; Wu, D.; and He, Z. 2024. Augmented self-mask attention transformer for naturalistic driving action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7108–7114.

Zhao, L.; Yang, F.; Bu, L.; Han, S.; Zhang, G.; and Luo, Y. 2021. Driver behavior detection via adaptive spatial attention mechanism. *Advanced Engineering Informatics*, 48: 101280.

Zhong, Y.; Yang, J.; Zhang, P.; Li, C.; Codella, N.; Li, L. H.; Zhou, L.; Dai, X.; Yuan, L.; Li, Y.; et al. 2022. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16793–16803.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*.