EyeCue: Driver Cognitive Distraction Detection via Gaze-Empowered Egocentric Video Understanding

Lang Zhang[*] , **JinYi Yoon**[*,†] , **Matthew Corbett**[‡] , **Abhijit Sarkar**[*] and **Bo Ji**[*]
[*]Virginia Tech [†]Inha University [‡]Army Cyber Institute at West Point
langzhang@vt.edu, jinyiyoon@inha.ac.kr, matthew.corbett@westpoint.edu, asarkar@vtti.vt.edu, boji@vt.edu

Abstract

Driver cognitive distraction is a major cause of road collisions and remains difficult to detect. Unlike manual or visual distraction, cognitive distraction is diverted by thoughts unrelated to driving, even when the driver appears visually attentive and exhibits no explicit physical movements. In this work, we propose EyeCue, a gaze-empowered egocentric video understanding framework, to detect driver cognitive distraction. A key insight is that cognitive distraction manifests in the interaction between eye gaze and visual context. To capture this interaction, EyeCue integrates eye gaze with egocentric video to enable context-aware modeling of the driver's attention over time. Furthermore, to tackle the limited scale and diversity of existing datasets, we introduce CogDrive, a comprehensive multi-scenario dataset that augments four existing driving datasets with cognitive distraction annotations. Through extensive evaluations on CogDrive, we show that EyeCue achieves the highest accuracy of 74.38%, outperforming 11 baselines from 6 model families by over 7%. Notably, EyeCue can achieve an accuracy of over 70% across various driving scenarios (different road types, times of day, and weather conditions) with strong generalizability. These results highlight the importance of modeling gaze-context interactions and the effectiveness of cross-modal interaction modeling for multimodal cognitive distraction detection. Our codes and CogDrive dataset resources are available here.[1] The online technical report can be found here.[2]

1 Introduction

Distracted driving is a leading cause of traffic fatalities, accounting for approximately 30% of all such cases, claiming 3,275 lives in 2023 [NHTSA, 2023]. It also causes significant property damage, which requires $9.8 billion in direct infrastructure and road-maintenance costs [NHTSA, 2025]. Besides these economic burdens, distracted driving places heavy demands on both law enforcement and medical resources. Therefore, a timely investigation into the problem of driver distraction detection is crucial.

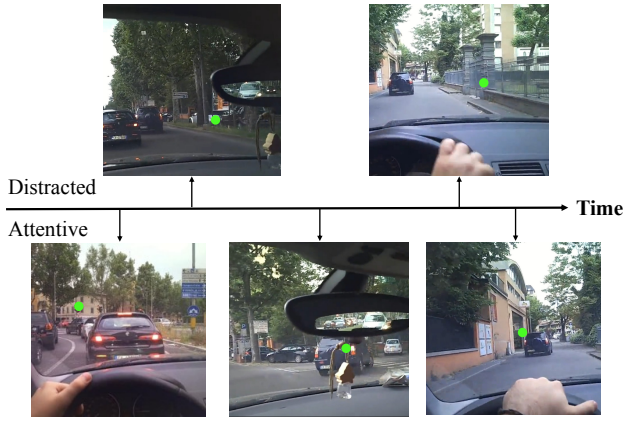

Figure 1: *Is the driver cognitively distracted?* This figure shows a driver's journey on the road, including distracted driving (top) and attentive driving (bottom). Video frames come from the DR(eye)VE dataset [Palazzi *et al.*, 2018], which records the driver's egocentric views and corresponding gaze points over time. We added a green dot to each raw frame to represent the driver's gaze point at that time.

Distractions during driving can be classified into three types: manual, visual, and cognitive distractions [Kashevnik *et al.*, 2021]: (i) *manual distraction* occurs when the driver's hands are off the wheel (e.g., secondary task: holding a phone or food while driving); (ii) *visual distraction* occurs when the driver looks away from the attention areas covered by the windshield or rear-view mirrors (e.g., looking at a navigation system); (iii) *cognitive distraction* occurs when the driver's attention is diverted by thoughts unrelated to driving (e.g., staring at a point while thinking about something else unrelated to driving), even though their gaze remains on the road. Both manual and visual distractions have been widely studied, whereas cognitive distraction remains underexplored because it is latent and non-observable [Guo *et al.*, 2024]. Detailed discussion of these distraction types is in the technical report [Zhang *et al.*, 2026].

Since cognitive distraction is reflected in the driver's internal mental state over time, it is often difficult to detect directly from external behaviors alone. Even when a driver appears

[1]Code: https://github.com/langzhang2000/EyeCue
[2]Technical report: https://arxiv.org/abs/2605.07859

visually attentive (looking at the road) and manually attentive (keeping their hands on the wheel), cognitive distraction may still happen. Existing approaches mostly rely on intrusive methods that require attaching physiological sensors to the driver or interrupting normal driving. For example, electroencephalography (EEG) provides precise neural signals for inferring cognitive state but requires direct physical coupling between sensors and the body [Li *et al.*, 2023]. Self-reported questionnaires can be used to evaluate a driver's perception of the driving environment [Cooper *et al.*, 2014], and detection response tasks are commonly used to assess their reaction time to unexpected events [Al-Mekhlafi *et al.*, 2024]. However, these methods may disrupt normal driving and only support post-hoc evaluations, limiting their practicality for continuous use. Therefore, *a key research question is how to design a non-intrusive method to detect driver cognitive distraction over time without disrupting driving.*

To answer this question, we leverage eye gaze as a behavioral cue to understand the driver's cognitive state, as eye gaze reflects how the internal cognitive process allocates attention [Oyama *et al.*, 2019]. Moreover, with the growing adoption of lightweight augmented reality glasses such as Meta Aria Gen 2, real-time eye-tracking data can be collected in a non-intrusive manner without additional user effort [Engel *et al.*, 2023]. However, relying on eye-tracking data alone for cognitive distraction detection is insufficient, as it lacks contextual scene information [Zhou *et al.*, 2025]. The same gaze pattern may reflect different cognitive states across driving scenarios. For example, as shown in Fig. 1, in the first attentive case, the driver fixates on the traffic light while waiting at the signal; in the first distracted case, the driver fixates on a parked vehicle along the roadside during straight-ahead driving. This naturally motivates a multimodal perspective that adopts both eye gaze and contextual scene information to interpret how attention is temporally allocated during driving.

To that end, we aim to infer the driver's cognitive state by analyzing temporal eye-gaze patterns and the visual targets of attention using egocentric video, which captures the spatio-temporal structure of the driving scene [Zhou *et al.*, 2024]. Specifically, we leverage a key insight: *detecting cognitive distraction requires understanding how the driver's gaze interacts with the surrounding egocentric visual context over time.* This interaction is reflected both over a short driving clip (i.e., global interaction) and in each frame (i.e., fine-grained interaction). Despite this useful insight, a lack of high-quality datasets suitable for modeling this interaction remains a challenge. Datasets that provide both eye-tracking information and egocentric driving videos with cognitive distraction annotations are very limited, primarily due to the difficulty of data collection and annotation. Hence, we are faced with three key challenges: *(C1) How to learn both global and fine-grained features from the driver's egocentric videos and eye gaze data, respectively? (C2) How to integrate eye gaze information with egocentric videos to detect driver cognitive distraction? (C3) Lack of scalable and generalizable datasets for cognitive distraction detection.*

To address C1 and C2, we propose EyeCue, a non-intrusive framework that integrates temporal eye gaze with egocentric video to detect driver cognitive distraction. EyeCue consists of three core components: (i) a *video encoder* that processes each egocentric video clip to understand the surrounding context of the driver; (ii) a *gaze encoder* that analyzes eye-tracking data to extract the driver's eye gaze patterns; and (iii) a *gaze-driven semantic query (GDSQ)* module that leverages gaze cues to dynamically select visual tokens from the egocentric video, reflecting how the driver's gaze is allocated across the context over time. Then, we fuse the outputs of these three modules to detect cognitive distraction.

As for C3, DR(eye)VE is the only publicly available dataset that contains eye gaze, egocentric videos, and cognitive distraction labels [Palazzi *et al.*, 2018]. However, it covers limited driving scenarios, and the annotated distracted samples may be insufficient. Hence, we explore three other existing driving datasets (along with DR(eye)VE) to create CogDrive, an egocentric cognitive distraction dataset. We select BDD-A [Xia *et al.*, 2018], DADA-2000 [Fang *et al.*, 2021], and TrafficGaze [Deng *et al.*, 2019] since they provide gaze data and egocentric videos. Following the DR(eye)VE's annotation procedure, we create a dataset consisting of 3,662 samples with cognitive distraction annotations.

To the best of our knowledge, *EyeCue is the first work that integrates temporal eye gaze information with egocentric videos to detect driver cognitive distraction.* Our main contributions are summarized as follows:

- We propose EyeCue, a gaze-empowered egocentric video framework for cognitive distraction detection that explores the interaction between eye gaze and the driving context. Specifically, EyeCue jointly learns representations from egocentric video and eye gaze through two encoders: a video encoder captures scene context and local visual cues, while a gaze encoder models temporal gaze behavior patterns (addressing Challenge C1). Moreover, we use gaze to guide video preprocessing and introduce a GDSQ module that directs cross-attention toward gaze-relevant visual regions, modeling gaze-context interactions for cognitive understanding (addressing Challenge C2).
- We introduce CogDrive, a cognitive distraction dataset consisting of 3,662 annotated egocentric video clips with gaze signals (addressing Challenge C3). It covers various driving scenarios, including diverse road scenes and driving events.
- Through extensive experiments on CogDrive, we show that EyeCue achieves an accuracy of 74.38%, outperforming 11 baselines from 6 model families, including gaze-only, classical video classification, egocentric, foundation, gaze-with-image, and gaze-with-video models, by more than 7% in absolute gain. These findings highlight the importance of modeling gaze-context interactions and the effectiveness of cross-modal interaction modeling for multimodal cognitive distraction detection.

2 Related Work

In this section, we first provide the background on driver distraction detection, then discuss gaze-based analysis techniques and video understanding models, and finally, examine recent advances in gaze-empowered vision models.

Driver Distraction Detection. There are three main types of driver distraction: manual, visual, and cognitive [Kashevnik *et al.*, 2021]. For manual and visual distractions, recent approaches commonly use in-car cameras to detect the driver's observable body movements [Sonth *et al.*, 2023]. Since cognitive distraction manifests in an individual's mental state, most research focuses on intrusive solutions to measure their cognitive loads. For example, Figalová et al. [2023] use questionnaires to evaluate the driver's cognitive state. However, these methods could disrupt normal driving. This motivates us to design a non-intrusive method to detect cognitive distractions.

Gaze-based Analysis Techniques. The driver's eye gaze cues are valuable for driver attention assessments [Huang *et al.*, 2025]. For example, Maralappanavar et al. [2016] uses the driver's pupils to estimate the driver's state, but it misses the environmental context. Zhou et al. [2025] predict the driver's gaze area and generate text explaining why drivers should focus on these areas. Bhagat et al. [2023] show that driver gaze patterns and saliency can vary by driving tasks. Recently, DCDD combines a single frame with eye-tracking data for distraction detection [Qiao *et al.*, 2025], but it lacks temporal attention modeling.

Video Understanding Models. Video understanding models can understand the context perceived by the driver during driving [Min *et al.*, 2024]. Models like TimeSformer [Bertasius *et al.*, 2021] and VideoMAE [Tong *et al.*, 2022] capture the spatio-temporal relationship of visual content. Video foundation models have enhanced reasoning ability when integrated with other modalities (e.g., text) [Wang *et al.*, 2024]. This inspires us to consider the integration of eye gaze with video understanding models.

Gaze-Empowered Vision Models. Current multimodal models illustrate the superiority of integrating eye gaze with visual content. For example, Voila-A [Yan *et al.*, 2024] and GazeGPT [Konrad *et al.*, 2024] use eye gaze to help visual models recognize which objects the user is focusing on in an image. GazeLLM reduces the computational load of the model by using only the user's fixation area [Rekimoto, 2025]. Egovideo captures the egocentric video for a better perception [Pei *et al.*, 2024]. However, these works focus on the visual content in an image rather than the first-person state understanding in a video. Recently, egoEmotion recognize eye gaze as a critical perceptual modality to reflect human states [Jammot *et al.*, 2025]. Inspired by this, we design a method that can fuse eye gaze and egocentric video to detect driver cognitive distraction.

3 Key Insights and Challenges

In this section, we introduce the key insights and discuss the main challenges in model design and dataset construction.

3.1 Key Insights

Research on safe driving shows that the driver's perception is reflected in how they allocate attention to different objects over time [Du *et al.*, 2020]. Musabini and Chetitah [2020] find that during normal driving, the driver fixates on a wide area of the environment, whereas when cognitively distracted, their attention remains confined to specific regions. In addition, Ojstersek and Topolsek [2019] observe that the driver looks at different objects in various contexts to maintain an attentive cognitive state. From these studies, we draw the following key insight: *detecting cognitive distraction requires understanding how the driver's gaze interacts with the surrounding egocentric visual context over time.* When the driver is cognitively attentive, they exhibit more frequent interactions with task-related objects [Bhagat *et al.*, 2023]. In contrast, when they are distracted, their gaze is less likely to support driving task-related objects in the scene [Sarkar, 2022].

3.2 Main Challenges

Building on the aforementioned insight, designing a non-intrusive driver cognitive distraction detection system requires addressing two model design challenges and one data construction challenge.

(C1) How to learn both global and fine-grained features from the driver's egocentric view and eye gaze? Specifically, global features help us grasp the overall information about the environment and the driver's gaze throughout the entire video clip. Fine-grained features reveal object features in the scene and the gaze pattern at each moment. However, global and fine-grained features operate at different temporal and spatial scales. Therefore, we need a way to model these features of gaze and video modalities separately.

(C2) How to integrate gaze information with egocentric videos to detect driver cognitive distraction? We need to leverage the driver's eye gaze information to understand how they perceive the environment. For example, the driver might move their gaze from a traffic light to a pedestrian when they are waiting for the red light. In this case, we should fuse gaze information with egocentric video representations to capture how the driver attends to different regions in the scene. How to effectively integrate gaze cues with egocentric video representations remains a major challenge.

(C3) Lack of scalable and generalizable datasets for cognitive distraction detection. Designing datasets for driver cognitive distraction detection remains challenging due to limited data availability. Currently, DR(eye)VE is the only dataset that provides egocentric driving videos, eye gaze, and cognitive distraction labels. However, DR(eye)VE is collected under restricted driving scenarios and geographic regions, limiting its generalizability to other scenarios. Besides, cognitive distraction is implicit, which makes reliable annotation difficult without a concrete protocol. These challenges highlight both the necessity and difficulty of constructing a comprehensive dataset for cognitive distraction detection.

4 Our Design: EyeCue

In this section, we state our goal and discuss how our solution addresses the aforementioned two design challenges.

Design Goal. Our goal is to detect the driver's cognitive distraction. Based on the driver's egocentric video, we obtain a sequence of preprocessed video clips $\mathcal{F}$ together with temporally aligned gaze coordinates $\mathcal{C}$. The task is formulated

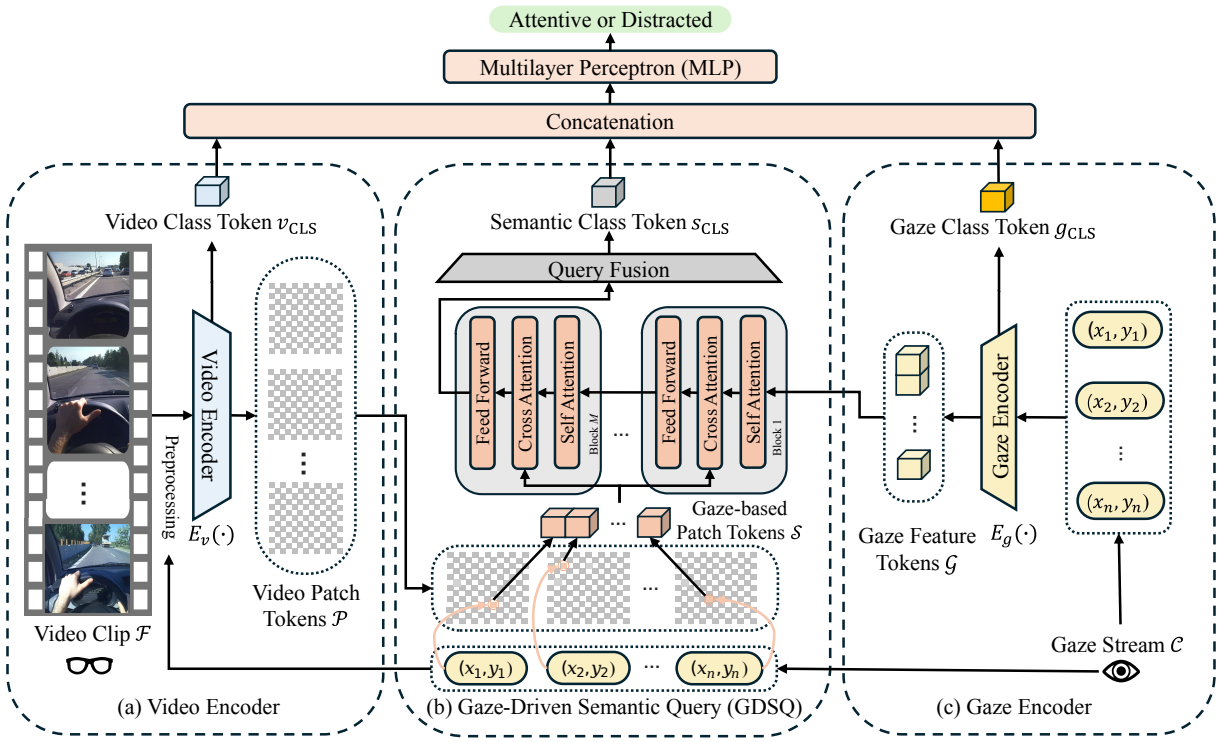

Figure 2: **EyeCue architecture.** (a) The video encoder extracts the global contextual information and fine-grained visual details. (b) The GDSQ module captures the relationship between visual context and eye gaze. (c) The gaze encoder obtains global attention patterns and frame-level gaze details. Finally, these three types of class tokens are concatenated and fed into a multilayer perceptron for classification.

as a binary classification problem, where the model predicts whether the driver is *cognitively distracted* or *attentive*.

Architecture Overview. We propose EyeCue, a multimodal framework that jointly leverages egocentric video and eye gaze to detect driver cognitive distraction. As illustrated in Fig. 2, EyeCue comprises three core components specialized to address the design challenges outlined above. To address Challenge C1, EyeCue employs two modality-specific encoders: (i) a *video encoder* that extracts both global and fine-grained spatio-temporal features from the egocentric videos, and (ii) a *gaze encoder* that models global attention patterns as well as frame-level gaze dynamics. These encoders enable the model to represent the driving scene and the driver's gaze behavior across the entire video clip. To address Challenge C2, we further introduce (iii) a *Gaze-Driven Semantic Query* module, which models the interaction between the gaze and the context. Finally, we concatenate the video class token, gaze class token, and semantic class token to form a unified representation. This representation is then fed into a multilayer perceptron for classification.

In EyeCue, gaze information is utilized in three ways. First, in the video encoder, gaze is used to guide video preprocessing by emphasizing regions attended by the driver. Second, the gaze encoder models both global attention patterns over the entire clip and frame-level gaze dynamics. Third, the GDSQ module aligns gaze with fine-grained visual content through gaze-guided patch selection and cross-attention. Together, these designs enable EyeCue to fully exploit both gaze and video modalities for cognitive distraction detection. Among these strategies, video preprocessing and the GDSQ are critical for modeling the interaction between the driver's gaze and the surrounding context.

Video Encoder. Given an egocentric video clip consisting of n frames, we first apply gaze-based video preprocessing, such as dot overlays, heatmap masks, and cropped frames, to enhance the input videos. Let f_t be the t-th frame. Then, we can denote the preprocessed video clip consisting of n frames by $\mathcal{F} := \{f_t\}_{t=1}^{n}$. Next, we employ a pre-trained video encoder $E_v(\cdot)$ to obtain rich spatio-temporal representations of the driving scene[Bertasius *et al.*, 2021]. We prefer a video encoder that can extract features from fine-grained regions within each frame. The video encoder processes the driver's egocentric video to generate a single video class token v_{CLS} and a sequence of video patch tokens $\mathcal{P} := \{p_i\}_{i=1}^{l}$, where l is the total number of video patch tokens, and p_i denotes the i-th video patch token. The video class token learns the global representation of the driving environment. Also, each image frame is divided into multiple patches, and each patch is encoded into a patch token to represent the fine-grained feature of that small area in the environment scene.

Gaze Encoder. Let (x_t, y_t) be the aligned eye gaze coordinates in frame f_t. Then, for the entire video clip, we have a sequence of coordinates $\mathcal{C} := \{(x_t, y_t)\}_{t=1}^{n}$. We use a gaze encoder $E_g(\cdot)$ to embed these gaze coordinates into a gaze class token g_{CLS} and a sequence of gaze feature tokens $\mathcal{G} := \{g_t\}_{t=1}^{n}$, where g_t denotes the t-th gaze feature to-

ken. Specifically, in the gaze encoder $E_g(\cdot)$, we first project each fixation coordinate to the same embedding space as the video patch tokens through a learnable linear layer. This produces an embedded coordinates token sequence. Then, we add a class token to this sequence. Next, we process them through a customized lightweight transformer encoder to obtain the gaze class token and the sequence of gaze feature tokens [Vaswani *et al.*, 2017]. The gaze class token summarizes the driver's gaze pattern throughout the entire clip. The gaze feature tokens capture the gaze pattern at the frame level.

Gaze-Driven Semantic Query. This module integrates the driver's gaze with egocentric videos, which is critical for gaze-context interaction modeling. We aim to extract the driver's visual perception feature. First, we project a gaze point to the corresponding area in the frame and select h video patch tokens around this area. Specifically, for each frame, we select either a single patch (i.e., $h = 1$) where the gaze point is located or multiple patches (i.e., $h > 1$) that contain the surrounding area. We discuss gaze-based patch token selection strategies in the technical report [Zhang *et al.*, 2026]. Then, we repeat this operation for all frames, and can get $h \times n$ video patch tokens in total. Gaze-based patch tokens can be denoted by $\mathcal{S} := \{s_m\}_{m=1}^{h \cdot n}$, where s_m is the m-th gaze-based patch token.

To capture the interaction between gaze and environment at the frame level, we inject gaze feature tokens $\mathcal{G}$ and gaze-based patch tokens $\mathcal{S}$ as input into the cross-attention block. This cross-attention block allows the system to model the interaction between the driver's eye gaze and the environment. We draw on the intuition that the driver uses eye gaze to perceive and query the surrounding environment. Hence, we use gaze feature tokens $\mathcal{G}$ as query inputs, and the gaze-based patch tokens $\mathcal{S}$ serve as key and value inputs in the cross-attention block. We repeat the cross-attention block M times to iteratively refine the relationship between the gaze and the environment features. Then, we apply a pooling strategy to the output of the M-th cross-attention block to obtain a single semantic class token s_{CLS}. The class token can represent the overall interaction between the driver and the context.

Finally, we concatenate the three types of class tokens (v_{CLS}, s_{CLS}, and g_{CLS}), which capture global information for classification, and feed them into a multilayer perceptron that predicts whether the driver is cognitively distracted.

5 Evaluations

In this section, we describe the implementation details and experimental setup, introduce the CogDrive dataset, and present evaluation results for EyeCue.

5.1 Implementation and Experimental Setup

Model Configuration. For the *video encoder*, we evaluate two pre-trained backbones: $\text{TimeSformer}_{\text{K600}}$ [Bertasius *et al.*, 2021] and $\text{VideoMAE}_{\text{K400}}$ [Tong *et al.*, 2022]. The *gaze encoder* consists of a Transformer with 8 attention heads and 1 encoder block to embed raw gaze information. The *GDSQ* module has 2 cross-attention blocks (i.e., $M = 2$), followed by a two-layer MLP. The hyperparameters of the model configuration are chosen empirically, as this lightweight setting achieves strong performance. We discuss video encoder selection strategies in the technical report [Zhang *et al.*, 2026].

Dataset	Attentive	Distracted	Total
DR(eye)VE	1,485 (78.41%)	409 (21.59%)	**1,894**
BDD-A	424 (66.88%)	210 (33.12%)	**634**
DADA-2000	463 (73.84%)	164 (26.16%)	**627**
TrafficGaze	481 (94.87%)	26 (5.13%)	**507**
CogDrive	**2,853 (77.91%)**	**809 (22.09%)**	**3,662**

Table 1: Clip statistics of CogDrive datasets.

Baselines. We compare EyeCue with 11 baselines from 6 model families (i) *Gaze only*: Heatmap-based SVM [Musabini and Chetitah, 2020]; (ii) *Classical backbones*: TimeSformer [Bertasius *et al.*, 2021] and VideoMAE [Tong *et al.*, 2022]; (iii) *Egocentric video understanding models*: EgoVideo [Pei *et al.*, 2024]; (iv) *Video foundation models*: InternVideo2 [Wang *et al.*, 2024], Video-LLaVA [Lin *et al.*, 2023], and VideoLLaMA3 [Zhang *et al.*, 2025]; (v) *Gaze with images*: GazeGPT [Konrad *et al.*, 2024] and Voila-A [Yan *et al.*, 2024]; (vi) *Gaze with videos*: GazeVQA [Ilaslan *et al.*, 2023] and GazeLLM [Rekimoto, 2025].

Training and Inference. We conduct training and inference on an NVIDIA L40S GPU, fine-tuning the video encoder, and training the gaze encoder and GDSQ module from scratch for 15 epochs. The remaining training configuration settings follow the default specifications of the video encoder. All trainable baseline methods are fine-tuned on the CogDrive dataset for fair comparison, while training-free methods (e.g., GazeGPT and GazeLLM) are evaluated without fine-tuning. We discuss prompt designs for language models in the technical report. Trainable baselines are averaged over five runs, while training-free models are evaluated once.

5.2 CogDrive Dataset

To address Challenge C3, we introduce CogDrive, a comprehensive dataset built by integrating four existing driving video datasets: DR(eye)VE [Palazzi *et al.*, 2018], BDD-A [Xia *et al.*, 2018], DADA-2000 [Fang *et al.*, 2021], and TrafficGaze [Deng *et al.*, 2019]. We select these datasets because they provide both gaze and egocentric videos. We follow a unified protocol derived from DR(eye)VE's labels with domain expert guidance for consistency. Detailed annotation procedures are provided in the technical report [Zhang *et al.*, 2026]. The protocol defines attentive vs. distracted criteria based on gaze-context interaction. All clips are independently labeled by two annotators and subsequently reviewed by a domain expert, achieving an inter-annotator agreement of over 98%. Following prior observations that cognitive distraction clips are typically short, we segment each video into fixed-length clips consisting of 16 frames.

Clip-Level Statistics. Table 1 reports the clip statistics of CogDrive. The dataset contains 3,662 clips, including 2,853 attentive and 809 distracted samples. Attentive clips dominate across all source datasets, which is in line with reality.

Methods		Backbone	Leave-one-dataset-out						Aggregated	
			BDD-A		DADA-2000		DR(eye)VE		CogDrive	
			Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Gaze-only	Heatmap-based	SVM	61.95	0.57	65.85	0.64	49.39	0.14	66.12	0.64
Classical	TimeSformer$_{K400}$	ViT-B/16	60.71	0.62	62.63	0.63	55.40	0.55	65.35	0.70
	TimeSformer$_{K600}$	ViT-B/16	60.52	0.67	61.89	0.53	55.41	0.61	66.80	0.71
	VideoMAE$_{K400}$	ViT-B/16	57.38	0.62	61.28	0.60	58.85	0.64	67.21	0.66
Egocentric	EgoVideo	ViT-B/14	58.57	0.65	59.15	0.63	55.65	0.56	65.29	0.68
Foundation	InternVideo2$_{s1\text{-}1B}$	ViT-B/14	50.16	0.45	50.51	0.45	59.31	0.58	56.61	0.53
	Video-LLaVA	OpenCLIP-L/14	52.43	0.37	52.09	0.47	53.75	0.51	55.06	0.55
	VideoLLaMA3	ViT-B/16	54.86	0.47	53.26	0.38	51.17	0.39	53.17	0.45
Gaze w/ images	GazeGPT*	GLM-4.5V-AWQ	51.00	0.47	51.81	0.53	51.29	0.42	51.69	0.48
	Voila-A	CLIP ViT-L/14	58.81	0.57	67.68	0.67	54.91	0.45	62.81	0.58
Gaze w/ videos	GazeVQA	CLIP ViT-B/32	59.15	0.49	59.45	0.47	51.97	0.65	67.77	0.68
	GazeLLM*	Gemini-2.5-Pro	44.53	0.02	49.09	0.28	49.14	0.08	48.35	0.12
	EyeCue (Ours)	VideoMAE$_{K400}$	60.95	0.59	62.50	0.62	54.67	0.62	70.83	0.68
	EyeCue (Ours)	TimeSformer$_{K600}$	**65.24**	**0.71**	**68.29**	**0.68**	**60.20**	**0.67**	**74.38**	**0.74**

Table 2: Comparison of accuracy (%) and F1 score between EyeCue and baseline methods. Methods marked with $*$ are training-free models. **Bold** and underlined values indicate the best and the second-best results, respectively.

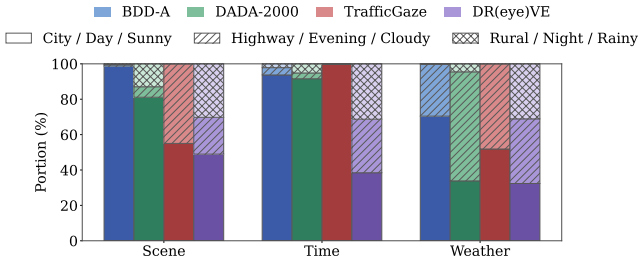

Figure 3: Clip distribution of CogDrive across various scenarios.

(a) Dot overlay (20 pixels) (b) Heatmap mask (75 pixels) (c) Cropped frame (448×448 pixels)

Figure 4: Example of gaze-integrated video preprocessing methods.

For training, we use all 809 distracted clips and randomly select 809 attentive clips following the sub-dataset distributions, then split the data into 70% for training and 30% for testing.

Scenario-Wise Distribution. Fig. 3 illustrates the clip distribution of CogDrive across four datasets under three complementary conditions: road type, time of day, and weather. This visualization highlights both intra-dataset composition and cross-dataset heterogeneity.

5.3 Evaluation Results

Comparison Study. Table 2 reports the accuracy and F1 score (for the distracted class) of EyeCue and 11 competitive baselines from 6 model families evaluated on the full CogDrive dataset and its sub-datasets. We adopt a leave-one-dataset-out protocol to assess cross-dataset generalization (e.g., training on DADA-2000, DR(eye)VE, and TrafficGaze when evaluating on BDD-A). We do not do leave-one-dataset-out on TrafficGaze due to its severe class imbalance. Overall, EyeCue achieves the best performance across all settings. On the full CogDrive dataset, EyeCue with a pre-trained TimeSformer$_{K600}$ backbone achieves the highest accuracy of 74.38% and an F1 score of 0.74, outperforming all baselines; it also demonstrates strong generalization under the leave-one-dataset-out setting. Notably, models trained with DR(eye)VE included in the training set generalize better than those trained on BDD-A or DADA-2000, suggesting that DR(eye)VE provides more informative samples for cognitive distraction modeling. In contrast, gaze-only, classical video methods, egocentric video models, video foundation models, and gaze-integrated vision approaches all show worse performance, indicating that spatio-temporal visual modeling or gaze cues alone are insufficient. These results highlight the importance of modeling gaze-context interactions and demonstrate the strong generalizability of EyeCue.

Gaze-Guided Video Preprocessing. Beyond raw videos, we explore several gaze-based video preprocessing strategies (Fig. 4), including dot overlays, heatmap masks, and gaze-centered cropping. As shown in Fig. 5, heatmap masks perform best, achieving the highest accuracy with a medium kernel radius, as they emphasize gaze-attended regions while preserving surrounding context. Dot overlays offer limited gains and degrade accuracy when overly salient. Cropping improves accuracy only within a narrow range of crop sizes,

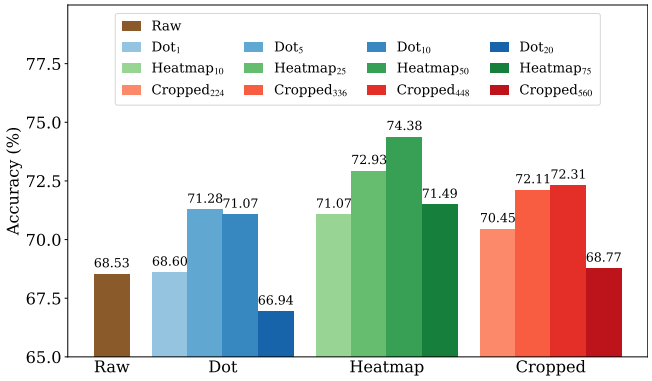

Figure 5: Accuracy (%) under different preprocessing methods.

# of Frames	Gaze-based Video Tokens			
	$h = 1$	$h = 5$	$h = 9$	$h = 25$
8	74.09	72.93	70.04	68.60
16	**74.38**	73.35	72.11	70.04

Table 3: Accuracy (%) under different number of input frames and gaze-based video tokens.

reflecting a trade-off between local focus and global context.

Hyperparameter Analysis. We analyze two key hyperparameters in EyeCue: clip length and the number of gaze-based video tokens. We compare clip lengths of 8 and 16 frames, and vary the number of gaze-based tokens as $h \in \{1, 5, 9, 25\}$ to control regional coverage around each fixation. As shown in Table 3, EyeCue achieves the highest accuracy at 74.38% with 16-frame clips and a single gaze-based token per frame (i.e., $h = 1$). Overall, 16-frame inputs get higher accuracy results, indicating the benefit of longer temporal context. In contrast, increasing h leads to a gradual accuracy drop, reaching its lowest value at $h = 25$. We attribute this phenomenon to the foveated nature of human vision. When $h = 1$, the fixation-centered token corresponds to the single video patch that contains the gaze point in each frame, providing a precise representation of the driver's foveal visual focus. In contrast, larger h selects multiple surrounding patches, expanding the representation beyond the fovea and introducing spatial overlap across frames, which dilutes the eye cue and weakens gaze-context interactions.

Ablation Study. We conduct an ablation study to analyze the contribution of each component in EyeCue. The results are shown in Table 4. Using only the gaze encoder yields limited performance as 54.13%, while the video encoder alone achieves 67.53%, highlighting the importance of scene context. Using the GDSQ module alone improves accuracy to 68.80%, suggesting that explicitly modeling gaze–context interactions captures useful information for cognitive distraction recognition. Direct fusion of the video and gaze encoders further improves accuracy to 72.31%, indicating that jointly modeling the two modalities provides complementary cues. Hence, both GDSQ and direct fusion highlight the importance of incorporating video and gaze information. Finally,

Gaze	✓			✓		✓	✓
Video		✓			✓	✓	✓
GDSQ			✓	✓	✓		✓
Acc.	54.13	67.53	68.80	69.36	70.25	72.31	**74.38**

Table 4: Accuracy (%) of ablation study for EyeCue components.

	Scene			Time			Weather		
	C	H	R	D	E	N	Su	Cl	Ra
Clips	368	50	66	366	44	74	253	147	84
Acc.	73.64	76.00	78.79	74.32	61.36	79.73	74.70	72.79	72.62

Table 5: Scenario-wise accuracy (%). "Clips" denotes the number of samples. C=City, H=Highway, R=Rural; D=Day, E=Evening, N=Night; Su=Sunny, Cl=Cloudy, Ra=Rainy.

the full EyeCue model, which combines direct multimodal fusion with gaze-context interaction modeling through GDSQ, achieves the best accuracy of 74.38%. In all settings with GDSQ, both video and gaze encoders are executed, but their class tokens may not be used. Overall, these results highlight the importance of interactions between gaze and context.

Scenario Analysis. We report the scenario-wise accuracy of EyeCue in Table 5 across different driving scenes, times of day, and weather conditions. EyeCue performs more reliably in rural scenarios. This may be because city driving involves more distracting factors, while the monotonous highway environment makes the driver's mind wander. Regarding time of day, night driving yields higher accuracy, likely because the scene generally includes fewer visible objects. As for weather, sunny scenes yield the highest accuracy due to its better visual quality. Overall, results suggest that the model generalizes well across diverse real-world scenarios. See failure cases analysis in the technical report [Zhang *et al.*, 2026].

Additional Evaluation Results. We provide additional detailed confusion matrices, Receiver Operating Characteristic (ROC), Area Under the Curve (AUC), computational complexity, and gaze noise robustness analysis for EyeCue in the technical report [Zhang *et al.*, 2026].

6 Conclusion

We propose EyeCue, a gaze-empowered egocentric video understanding model, to detect the driver's cognitive distraction in a non-intrusive way. By explicitly integrating eye gaze with egocentric video, EyeCue models the interaction between the driver and the surrounding context. Extensive experiments show that EyeCue achieves strong performance. Besides, the CogDrive dataset establishes a new benchmark for cognitive distraction across diverse driving scenarios.

Despite these advantages, EyeCue has several limitations. First, the additional datasets used to construct CogDrive are not originally designed for cognitive distraction analysis, which may lead to less consistent cognitive cues. Second, EyeCue may fail in complex driving scenes and poor gaze data quality situation. Addressing these limitations will further improve robustness in real-world driving scenarios.

Ethical Statements

All data used in this work is obtained from publicly available datasets. We do not collect any new human-subject data. Our method only uses gaze coordinates as input and does not involve biometric information such as iris patterns, pupil images, or other personally identifiable ocular features.

Acknowledgments

We are grateful to anonymous reviewers for providing valuable feedback that helps us improve the paper. We also thank Mr. Heesang Han from the Virginia Tech Transportation Institute (VTTI) for his help with dataset annotation. This work was supported in part by the Inha University Research Grant.

References

[Al-Mekhlafi *et al.*, 2024] Al-Baraa Abdulrahman Al-Mekhlafi, Ahmad Shahrul Nizam Isha, Nicholas Chileshe, Ahmed Farouk Kineber, Muhammad Ajmal, Abdullah O Baarimah, and Al-Hussein MH Al-Aidrous. Risk assessment of driver performance in the oil and gas transportation industry: Analyzing the relationship between driver vigilance, attention, reaction time, and safe driving practices. *Heliyon*, 10(6), 2024.

[Bertasius *et al.*, 2021] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[Bhagat *et al.*, 2023] Hirva Bhagat, Sandesh Jain, Lynn Abbott, Akash Sonth, and Abhijit Sarkar. Driver gaze fixation and pattern analysis in safety critical events. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, pages 1–8. IEEE, 2023.

[Cooper *et al.*, 2014] Simon Cooper, Joanne Porter, and Linda Peach. Measuring situation awareness in emergency setting: a systematic review of tools and outcomes. *Open Access Emergency Medicine*, pages 1–7, 2014.

[Deng *et al.*, 2019] Tao Deng, Hongmei Yan, Long Qin, Thuyen Ngo, and BS Manjunath. How do drivers allocate their potential attention? driving fixation prediction via convolutional neural networks. *IEEE Transactions on Intelligent Transportation Systems*, 21(5):2146–2154, 2019.

[Du *et al.*, 2020] Na Du, X Jessie Yang, and Feng Zhou. Psychophysiological responses to takeover requests in conditionally automated driving. *Accident Analysis & Prevention*, 148:105804, 2020.

[Engel *et al.*, 2023] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023.

[Fang *et al.*, 2021] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, and Hongkai Yu. Dada: Driver attention prediction in driving accident scenarios. *IEEE transactions on intelligent transportation systems*, 23(6):4959–4971, 2021.

[Figalová *et al.*, 2023] Nikol Figalová, Jürgen Pichen, Vanchha Chandrayan, Olga Pollatos, Lewis L Chuang, and Martin Baumann. Manipulating drivers' mental workload: Neuroergonomic evaluation of the speed regulation n-back task using nasa-tlx and auditory p3a. In *Adjunct Proceedings of the 15th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 145–149, 2023.

[Guo *et al.*, 2024] Zizheng Guo, Qing Liu, Lin Zhang, Zhenning Li, and Guofa Li. L-tla: A lightweight driver distraction detection method based on three-level attention mechanisms. *IEEE Transactions on Reliability*, 73(4):1731–1742, 2024.

[Huang *et al.*, 2025] Yexin Huang, Yongbin Lin, Lishengsa Yue, Zhihong Yao, and Jie Wang. From gaze to movement: Predicting visual attention for autonomous driving human-machine interaction based on programmatic imitation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 26146–26155, 2025.

[Ilaslan *et al.*, 2023] Muhammet Ilaslan, Chenan Song, Joya Chen, Difei Gao, Weixian Lei, Qianli Xu, Joo Lim, and Mike Shou. Gazevqa: A video question answering dataset for multiview eye-gaze task-oriented collaborations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10462–10479, 2023.

[Jammot *et al.*, 2025] Matthias Jammot, Björn Braun, Paul Streli, Rafael Wampfler, and Christian Holz. egoemotion: Egocentric vision and physiological signals for emotion and personality recognition in real-world tasks. *arXiv preprint arXiv:2510.22129*, 2025.

[Kashevnik *et al.*, 2021] Alexey Kashevnik, Roman Shchedrin, Christian Kaiser, and Alexander Stocker. Driver distraction detection methods: A literature review and framework. *IEEE Access*, 9:60063–60076, 2021.

[Konrad *et al.*, 2024] Robert Konrad, Nitish Padmanaban, J Gabriel Buckmaster, Kevin C Boyle, and Gordon Wetzstein. Gazegpt: Augmenting human capabilities using gaze-contingent contextual ai for smart eyewear. *arXiv preprint arXiv:2401.17217*, 2024.

[Li *et al.*, 2023] Guofa Li, Yufei Yuan, Delin Ouyang, Long Zhang, Bangwei Yuan, Xiaoyu Chang, Zizheng Guo, and Gang Guo. Driver distraction from the eeg perspective: A review. *IEEE Sensors Journal*, 24(3):2329–2349, 2023.

[Lin *et al.*, 2023] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

[Maralappanavar *et al.*, 2016] Shweta Maralappanavar, ReenaKumari Behera, and Uma Mudenagudi. Driver's distraction detection based on gaze estimation. In *2016 international conference on advances in computing, communications and informatics (icacci)*, pages 2489–2494. IEEE, 2016.

[Min *et al.*, 2024] Chen Min, Dawei Zhao, Liang Xiao, Jian Zhao, Xinli Xu, Zheng Zhu, Lei Jin, Jianshu Li, Yulan Guo, Junliang Xing, et al. Driveworld: 4d pre-trained scene understanding via world models for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15522–15533, 2024.

[Musabini and Chetitah, 2020] Antonyo Musabini and Mounsif Chetitah. Heatmap-based method for estimating drivers' cognitive distraction. In *2020 IEEE 19th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, pages 179–186. IEEE, 2020.

[NHTSA, 2023] NHTSA. Budget estimates: Fiscal year 2024. Technical report, U.S. Department of Transportation, National Highway Traffic Safety Administration, Washington, D.C., March 2023.

[NHTSA, 2025] NHTSA. Research note: Distracted driving in 2023. Technical Report DOT HS 813 703, U.S. Department of Transportation, National Highway Traffic Safety Administration, April 2025.

[Ojstersek and Topolsek, 2019] Tina Cvahte Ojstersek and Darja Topolsek. Eye tracking use in researching driver distraction: A scientometric and qualitative literature review approach. *Journal of eye movement research*, 12(3):10–16910, 2019.

[Oyama *et al.*, 2019] Akane Oyama, Shuko Takeda, Yuki Ito, Tsuneo Nakajima, Yoichi Takami, Yasushi Takeya, Koichi Yamamoto, Ken Sugimoto, Hideo Shimizu, Munehisa Shimamura, et al. Novel method for rapid assessment of cognitive impairment using high-performance eye-tracking technology. *Scientific reports*, 9(1):12932, 2019.

[Palazzi *et al.*, 2018] Andrea Palazzi, Davide Abati, Francesco Solera, Rita Cucchiara, et al. Predicting the driver's focus of attention: the dr (eye) ve project. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1720–1733, 2018.

[Pei *et al.*, 2024] Baoqi Pei, Guo Chen, Jilan Xu, Yuping He, Yicheng Liu, Kanghua Pan, Yifei Huang, Yali Wang, Tong Lu, Limin Wang, et al. Egovideo: Exploring egocentric foundation model and downstream adaptation. *arXiv preprint arXiv:2406.18070*, 2024.

[Qiao *et al.*, 2025] Yu Qiao, Xiaohui Yang, Jing Wang, Tongzhen Si, and Qingbei Guo. Driver cognitive distraction detection based on eye movement behavior and integration of multi-view space-channel feature. *Expert Systems with Applications*, 266:125975, 2025.

[Rekimoto, 2025] Jun Rekimoto. Gazellm: Multimodal llms incorporating human visual attention. *arXiv preprint arXiv:2504.00221*, 2025.

[Sarkar, 2022] Abhijit Sarkar. A comprehensive safety analysis for gaze fixation of drivers to outside scene. *Human Factors in Transportation*, 60(60), 2022.

[Sonth *et al.*, 2023] Akash Sonth, Abhijit Sarkar, Hirva Bhagat, and Lynn Abbott. Explainable driver activity recognition using video transformer in highly automated vehicle. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, pages 1–8. IEEE, 2023.

[Tong *et al.*, 2022] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2024] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024.

[Xia *et al.*, 2018] Ye Xia, Danqing Zhang, Jinkyu Kim, Ken Nakayama, Karl Zipser, and David Whitney. Predicting driver attention in critical situations. In *Asian conference on computer vision*, pages 658–674. Springer, 2018.

[Yan *et al.*, 2024] Kun Yan, Zeyu Wang, Lei Ji, Yuntao Wang, Nan Duan, and Shuai Ma. Voila-a: Aligning vision-language models with user's gaze attention. *Advances in Neural Information Processing Systems*, 37:1890–1918, 2024.

[Zhang *et al.*, 2025] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

[Zhang *et al.*, 2026] Lang Zhang, JinYi Yoon, Matthew Corbett, Abhijit Sarkar, and Bo Ji. Eyecue: Driver cognitive distraction detection via gaze-empowered egocentric video understanding. *arXiv preprint arXiv:2605.07859*, 2026.

[Zhou *et al.*, 2024] Yunsong Zhou, Linyan Huang, Qingwen Bu, Jia Zeng, Tianyu Li, Hang Qiu, Hongzi Zhu, Minyi Guo, Yu Qiao, and Hongyang Li. Embodied understanding of driving scenarios. In *European Conference on Computer Vision*, pages 129–148. Springer, 2024.

[Zhou *et al.*, 2025] Yuchen Zhou, Jiayu Tang, Xiaoyan Xiao, Yueyao Lin, Linkai Liu, Zipeng Guo, Hao Fei, Xiaobo Xia, and Chao Gou. Where, what, why: Towards explainable driver attention prediction. *arXiv preprint arXiv:2506.23088*, 2025.