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

Lang Zhang[1], JinYi Yoon[1,2], Matthew Corbett[3], Abhijit Sarkar[1], Bo Ji[1]
[1]Virginia Tech, USA; [2]Inha University, Korea; [3]Army Cyber Institute at West Point, USA

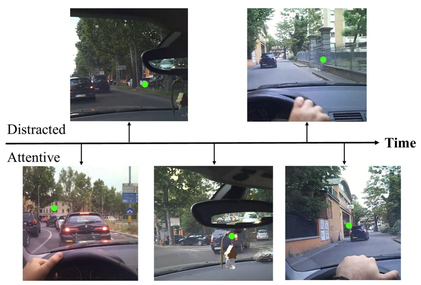
Project Resource

My LinkedIn

Background and Motivation

Manual and *visual* distractions usually have observable behaviors. But ***Cognitive*** distraction can occur even when the driver is looking at the road and keeping hands on the wheel. Therefore, cognitive distraction is latent and difficult to detect.

The driver may look at the road but are they cognitively attentive?



Manual

Hands off wheel

Visual

Looks away

Cognitive

Mind off task

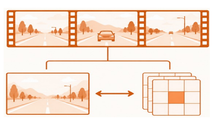
Limitation of Existing Approaches

- Intrusive measurements may disrupt normal driving and are unsuitable for continuous monitoring.
- Gaze alone model cannot reveal the driver's cognitive state without understanding the surrounding visual context.
- Existing gaze-vision models lack temporal modeling of gaze–context interactions in egocentric videos.

Key Insights and Challenges

Cognitive distraction is reflected in how the driver's gaze interacts with the surrounding visual context over time.

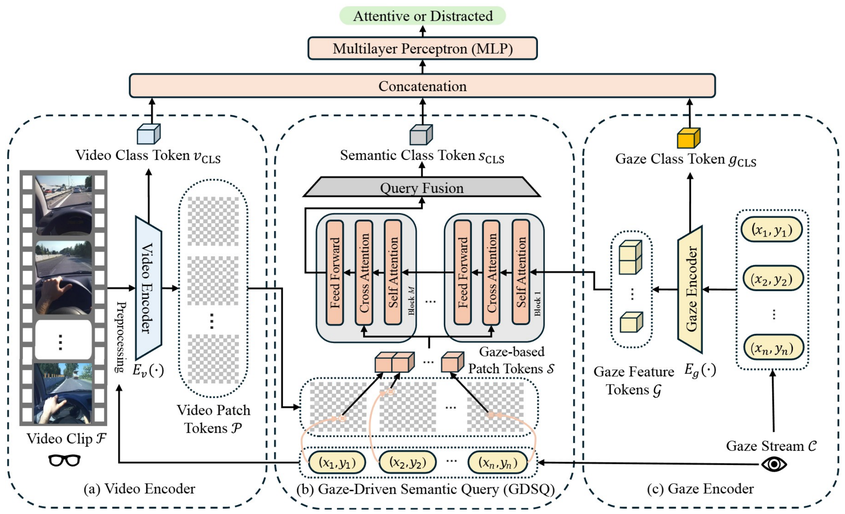
C1: Multiscale Representation
How can we capture global and fine-grained features from the driver's egocentric video and gaze?



C2: Gaze-Context Interaction
How can gaze dynamically query relevant visual regions?

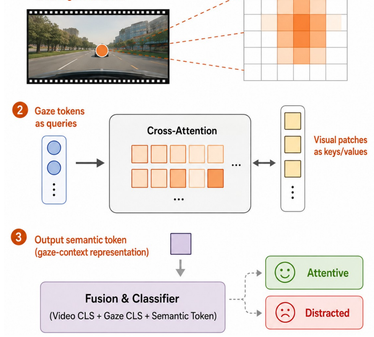
C3: Limited Dataset
How can we build a diverse and generalizable benchmark?

EyeCue Framework



How to Model Gaze-Context Interaction

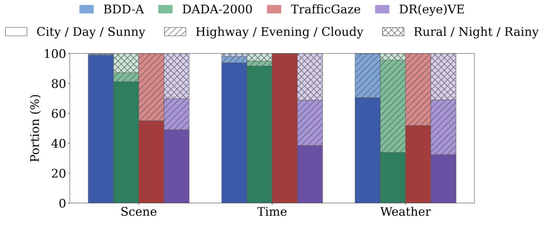
- EyeCue first selects *gaze-centered visual patches* from each video frame.
- Gaze feature tokens then query these patches through cross-attention, producing a semantic token that captures *gaze-context interactions* over time.
- Finally, the semantic token is fused with video and gaze class tokens to classify the driver as attentive or cognitively distracted.

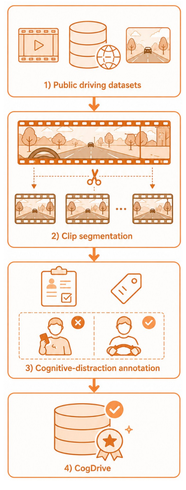


CogDrive Dataset

CogDrive Dataset: We integrate four public gaze-aligned driving datasets, segment videos into clips, and annotate cognitive distraction. The resulting CogDrive benchmark contains 3,662 clips and spans diverse road scenes, times of day, and weather conditions.

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**

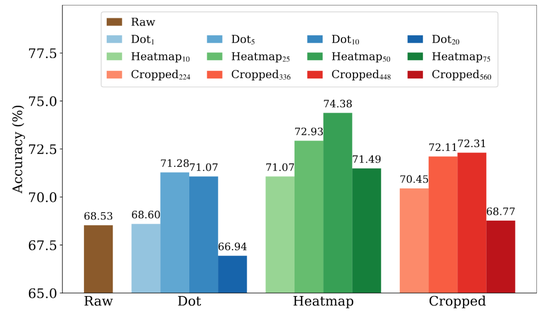




Main Results

Performance and Generalization: We compare EyeCue with 11 baselines across 6 model families on CogDrive. EyeCue achieves the best overall performance while maintaining strong robustness across different preprocessing strategies, model components, road scenes, times of day, and weather conditions.

Methods		Backbone	Leave-one-dataset-out: BDD-A Acc.	BDD-A F1	DADA-2000 Acc.	DADA-2000 F1	DR(eye)VE Acc.	DR(eye)VE F1	Aggregated: CogDrive Acc.	CogDrive F1
Gaze-only	Heatmap-based	SVM	61.95	0.57	65.85	0.64	49.39	0.14	66.12	0.64
Classical	$\text{TimeSformer}_{K400}$	ViT-B/16	60.71	0.62	62.63	0.63	55.40	0.55	65.35	0.70
	$\text{TimeSformer}_{K600}$	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	$\text{InternVideo2}_{s1\text{-}1B}$	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)	$\text{TimeSformer}_{K600}$	**65.24**	**0.71**	**68.29**	**0.68**	**60.20**	**0.67**	**74.38**	**0.74**



# 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

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

	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

Additional Findings

- *Gaze preprocessing:* Heatmap masks outperform raw video, dot overlays, and cropping.

- *Temporal context:* 16-frame clips perform better than 8-frame clips.

- *Precise gaze localization:* One gaze-centered patch performs better than selecting broad surrounding regions.

- *Scenario generalization:* Accuracy remains above 70% across different scenes, times, and weather conditions.

Take Away Message

- *Cognitive distraction requires understanding gaze-context interactions over time, beyond gaze or video alone.*
- *EyeCue achieves accurate and generalizable detection by jointly modeling gaze and egocentric visual context.*
- *CogDrive provides a diverse benchmark with 3,662 annotated clips across real-world driving scenarios.*