IJCAI ECAI BREMEN 2026

Commonwealth Cyber Initiative

NSF National Science Foundation

*IJCAI-ECAI 2026*

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

**Lang Zhang[1], JinYi Yoon[1,2], Matthew Corbett[3], Abhijit Sarkar[1], and Bo Ji[1]**

[1]Virginia Tech [2]Inha University [3]Army Cyber Institute at West Point

VIRGINIA TECH

VIRGINIA TECH TRANSPORTATION INSTITUTE

INHA UNIVERSITY

ARMY CYBER INSTITUTE AT WEST POINT

**Project Resource**

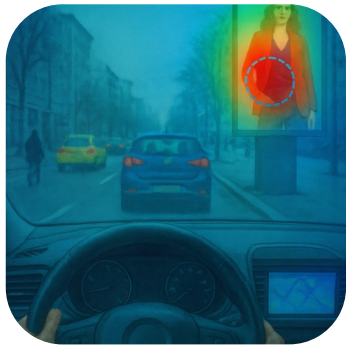
# Cognitive Distraction Is Hidden

- Manual distraction: hands off wheel
- Visual distraction: eyes look away from road-relevant areas
- **Cognitive distraction**: attention is diverted by thoughts unrelated to driving

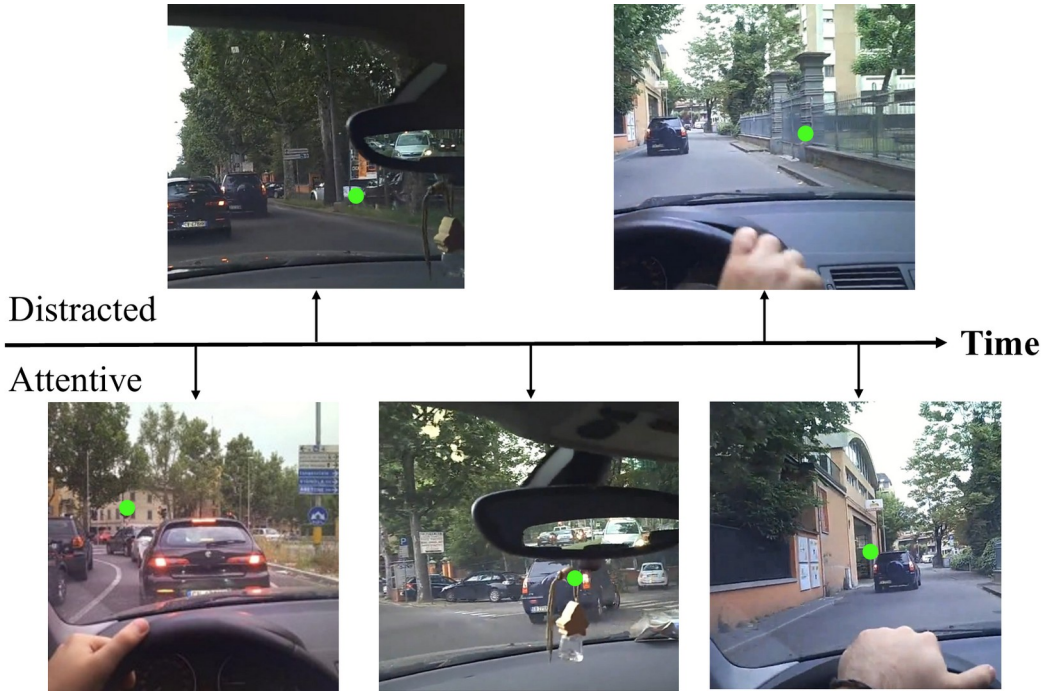
  ⇒ **Driver can look visually attentive, yet still be mentally distracted**

Hands

Eyes



Mind




**Key Question**

How can we detect cognitive distraction over time without disrupting normal driving?

Need a **non-intrusive, context-aware** signal for the driver's internal state

# Limitations of Existing Methods

**LIMITATION 1**

## Intrusive or disruptive methods

Body contact

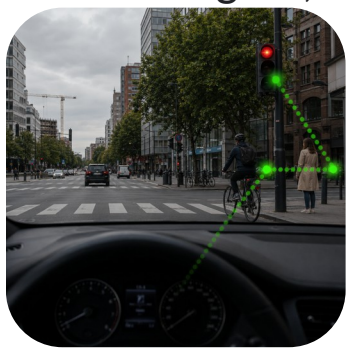
Post-hoc questionnaire

Interrupted driving task

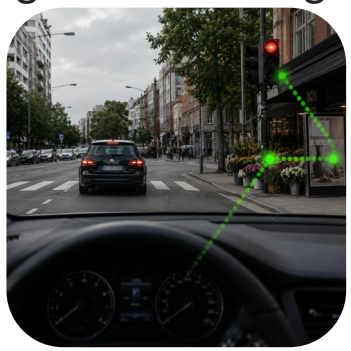
**LIMITATION 2**

## Gaze only is not enough

Same visible gaze, different cognitive meaning



**Attentive**: Gaze shifts from the signal to safety-relevant road pedestrians



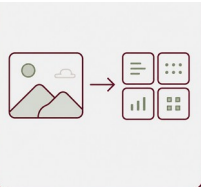
**Distracted**: Gaze shifts from the signal to irrelevant roadside objects

***Key Insight:*** **Cognitive distraction is reflected by how eye gaze interacts with the egocentric visual context over time**

# Research Challenges

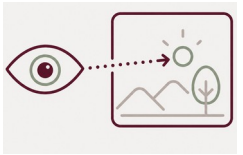
**1** **Multi-Scale Representation Learning**

How can we jointly capture global and fine-grained features from the driver's egocentric video and gaze data?



**2** **Gaze-Context Interaction Modeling**

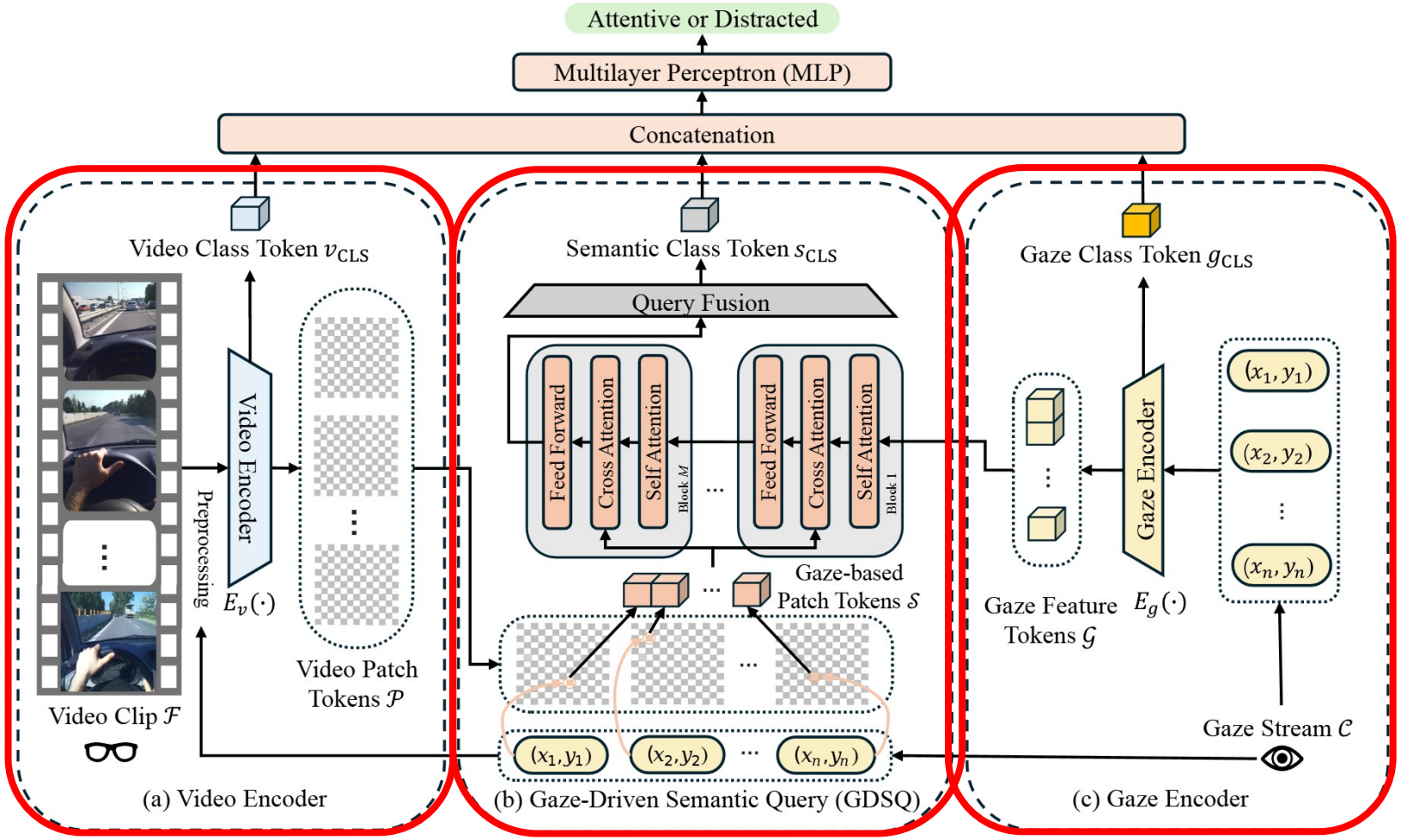
How can gaze be grounded in visual context to reveal whether attention supports the driving task?



**3** **Lack of Datasets for Cognitive Distraction Detection**

How can we construct reliable cognitive-distraction annotations across diverse real-world driving scenarios?

# EyeCue: Gaze-Empowered Egocentric Video Understanding




**Video Encoder**

Capture global driving context and visual cues across the video clip

**Gaze Encoder**

Model global attention patterns and frame-level gaze dynamics over time

**GDSQ**

*Core Innovation*

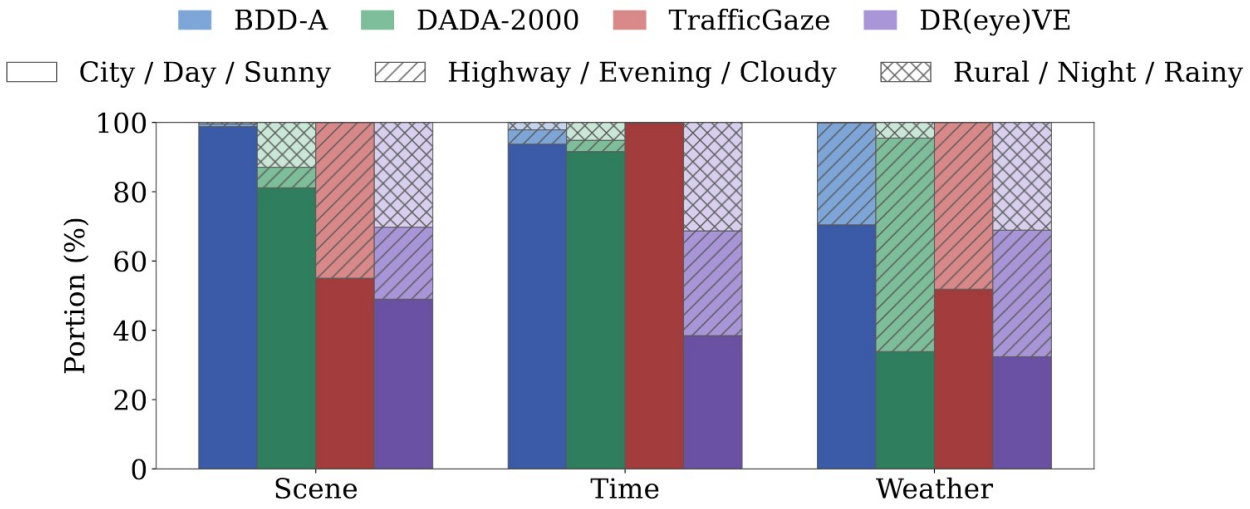
Aligns gaze-selected visual patches with gaze dynamics to infer whether attention supports driving or reflects cognitive distraction

# CogDrive: A Multi-Scenario Cognitive Distraction Dataset

## Dataset scale & composition

- Protocol: adapted from DR(eye)VE with domain-expert guidance
- Clips: videos segmented into fixed 16-frame clips
- Quality: two annotators w/ domain expert verification

| 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** |




## Scenario diversity

- Road type: city / highway / rural
- Time: day / evening / night
- Weather: sunny / cloudy / rainy

**CogDrive enables evaluation beyond a single restricted driving scenario**

# Overall Results: Accuracy

| 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** |

**EyeCue**

**74.38%**
Accuracy

**0.74**
F1 score

**6.61%**
absolute gain over baselines

**Gaze or video alone is insufficient; modeling their interaction yields the best performance**

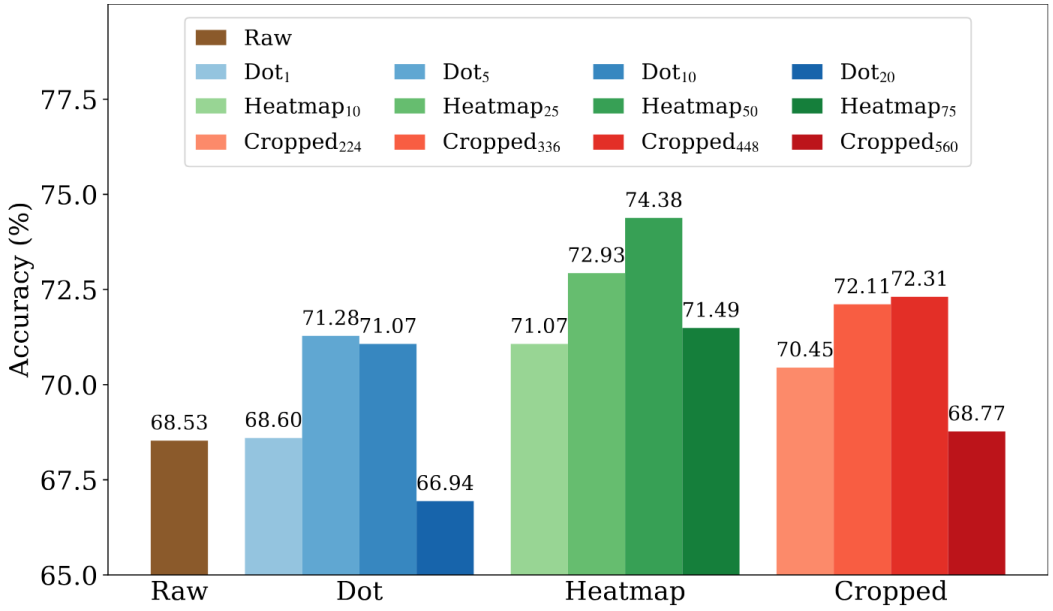
# Other Evaluations

| # 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 |

| | 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 |

C=City, H=Highway, R=Rural; D=Day, E=Evening, N=Night; Su=Sunny, Cl=Cloudy, Ra=Rainy.

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




***EyeCue Takeaways:***

- ✓ Longer context and one gaze-based video token perform best
- ✓ Heatmaps best balance gaze focus and scene context
- ✓ EyeCue generalizes across diverse driving conditions

# Conclusion

## “EyeCue: Gaze-Empowered Egocentric Video Understanding”

**Non-intrusive signal**

Egocentric video + gaze

**Gaze-context modeling**

Represent the driver’s interaction with environment

**Benchmark**

CogDrive: 3,662 video clips

**Future Work** Improve robustness in complex scenes and low-quality gaze data

VIRGINIA TECH®