A Staged Workflow for Driver Glance and Hand-State Annotation

Zhixiong Wang
Thrust of Robotics and Autonomous Systems
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, Guangdong, China
zwang477@connect.hkust-gz.edu.cn

Zeyun Du
Thrust of Data Science and Analysis
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, Guangdong, China
zdu881@connect.hkust-gz.edu.cn

Yuwen Chen
Thrust of Robotics and Autonomous Systems
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, Guangdong, China
ychen623@connect.hkust-gz.edu.cn

Dengbo He*
Thrust of Intelligent Transportation
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, Guangdong, China
Division of Emerging Interdisciplinary Areas
The Hong Kong University of Science and Technology
Hong Kong, Hong Kong
dengbohe@ust.hk

Abstract

Naturalistic driving video captures how drivers supervise under real road conditions, but converting long, multi-camera in-cabin footage into behavioral features still relies on manual coding, which is costly and hard to scale. We present a human-in-the-loop workflow that converts raw in-cabin video into two reviewable supervisory features: area-of-interest (AOI) gaze labels and hand-on-wheel state. Instead of a single end-to-end model, the workflow keeps region selection, gaze labeling, participant-specific calibration, temporal stabilization, and window-level metric extraction as distinct, observable stages, so an error can be localized. Specifically, the gaze module pairs Sample and Computation Redistribution for Efficient Face Detection (SCRFD) face detection with a YOLOv8-cls classifier, while the hand module uses GroundingDINO with a temporal state stabilization strategy. The workflow was validated in a naturalistic driving study with millions of video frames from 15 drivers. The source code of the workflow has been released at https://github.com/zdu881/autodri.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**.

Keywords

Naturalistic driving, Driver monitoring, Glance behavior, Hand-on-wheel state, Annotation

*Corresponding author.



This work is licensed under a Creative Commons Attribution 4.0 International License.
AutomotiveUI Adjunct '26, Gothenburg, Sweden
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2815-0/26/09
https://doi.org/10.1145/3828158.3838248

ACM Reference Format:
Zhixiong Wang, Zeyun Du, Yuwen Chen, and Dengbo He. 2026. A Staged Workflow for Driver Glance and Hand-State Annotation. In *18th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI Adjunct '26), September 20–23, 2026, Gothenburg, Sweden.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3828158.3838248

1 Introduction

Naturalistic driving studies record drivers' operations in their own vehicles in real-road environments and naturally occurring traffic, preserving the traffic, cabin layouts, lighting, and behavior that simulator and controlled field studies cannot fully reproduce [11, 25]. To minimize interference with driving behavior, an in-cabin data collection device that requires no pre-calibration before each drive is preferred. Thus, an in-cabin camera system is a good choice, as it can record drivers' manual and gaze behaviors at low cost and unobtrusively [1, 27]. Unlike eye-tracking systems, though in-cabin cameras do not require per-driver calibration [10], they do not automatically yield accurate fixation estimates. Fortunately, for many driving-related questions, coarser area-of-interest (AOI) gaze labels suffice. The challenging part is how to efficiently label a large number of frames.

This challenge scales with dataset size. Large naturalistic studies usually generate long, multi-view video; the MIT Advanced Vehicle Technology (MIT-AVT) study, for example, has accumulated billions of frames from more than 100 participants [2]. Traditional video reduction relies on human coders, a process that is time-consuming, costly, and dependent on coder training, quality control, and agreement checking [10], making exhaustive coding infeasible and annotation typically confined to a sampled subset [12]. A recent naturalistic analysis of Autopilot use, for instance, drew on this corpus but manually coded gaze and steering-wheel state from only a sampled set of short epochs from 16 drivers [16]. The challenge is even greater for multi-camera in-cabin video, where

researchers must identify not only where the driver is looking but also the driver's manual behaviors.

A fully automated process based on computer vision (CV) is an attractive alternative [2, 27]. Prior work has used in-vehicle cameras to analyze driver gaze, head pose, distraction, hand activity, and maneuver intention [27]. Specifically, studies have estimated driver gaze regions without eye tracking and detected hands, steering wheels, or phones to recognize hands-on-wheel and engagement in non-driving-related tasks (NDRTs) [3, 7, 21]. Recent workflows combine video processing, GPS synchronization, and head-pose estimation to extract task-specific behaviors such as head scans at intersections [17]. However, CV is a black-box classifier, which can raise concerns. For example, a prediction can be invalid when a driver crop is incorrect, a passenger enters the selected region, a face detector fails, or a brief flicker is misread as a behavioral event. Thus, they also only reported aggregate accuracy without revealing where an individual prediction came from. As a result, invalid outputs cannot be localized or reviewed. Further, most previous work targeted a single behavior and used a fixed camera view. To the best of our knowledge, no studies have examined how long, non-standardized, multi-camera in-cabin video can be converted into auditable and scalable behavioral features for studying driver supervision.

To address this gap, we present a human-in-the-loop workflow for extracting driver behavior from multi-camera naturalistic driving video, focusing on two observable dimensions of supervision: AOI-level visual attention and hand-on-wheel state. Rather than detecting a single behavior category, it converts ordinary in-cabin video into frame-level and segment-level features that support later analysis of driver behavior. The workflow starts from a small set of manually labeled samples, allowing subject-specific calibration to handle differences in drivers, interiors, camera placement, and lighting, and human review at key stages. In general, it converts annotation from long-duration manual coding to small-sample human calibration with model-assisted inference, and ensures that every feature in the workflow is traceable at the review stage. Finally, a participant-holdout protocol was used to validate our workflow.

2 Method

Figure 1 illustrates the proposed four-stage annotation and extraction pipeline designed to convert raw in-cabin video into measurable behavioral features. The workflow proceeded as follows:

(1) **Multi-Subject Sample Extraction and Preprocessing**: Raw video frames were first processed with the Sample and Computation Redistribution for Efficient Face Detection (SCRFD) face detector [4] to generate driver-region clips. These clips underwent sample extraction, followed by manual annotation into four categories (Forward, In-Vehicle, Non-Forward, and Others) to construct a multi-subject labeled dataset.
(2) **Base Gaze Model Training**: A base gaze model was pretrained on the multi-subject labeled dataset to establish a generalized feature representation for driver gaze.
(3) **Subject-Specific Calibration and Inference**: For each subject, a subject-specific small sample was annotated. Then, the model was validated against this sample; if the performance was acceptable, a YOLOv8-cls classifier was fine-tuned and exported as an optimized ONNX model for inference. If validation failed, the samples were reviewed and relabeled to refine the calibration.
(4) **Study Integration**: In parallel, GroundingDINO [13] detected the hands and the steering wheel in each frame, and the overlap between their bounding boxes was used to classify each frame as hands ON, OFF, or UNCERTAIN (when a frame provided no reliable detection evidence). The per-frame outputs of the gaze and hand modules were then converted into behavioral metrics in two steps. First, temporal stabilization was implemented, which required a predicted label to persist over several consecutive frames before it replaced the current one, so that a single misclassified or missed frame was not marked as a change in behavior. Then, the stabilized frame sequences were summarized over fixed-length time windows into metrics such as the proportion of time spent glancing at each AOI and the proportion of time with hands on the wheel.

2.1 Gaze Detection Module

Figure 2 illustrates our AOI classification scheme. Rather than the binary "forward vs. non-forward" approach used in prior work [16], which further subdivided non-forward gazes into driving-related (e.g., instrument cluster) and non-driving-related (e.g., smartphone) categories, we defined three distinct areas of interest (AOIs). Forward covered attention toward the forward roadway. In-Vehicle covered gazes toward the instrument cluster, center display, mirrors, or vehicle controls. Non-Forward captured attention directed away from both the forward roadway and in-vehicle targets. A fourth label, Others, handled invalid or non-interpretable frames (e.g., occlusion, extreme lighting). We separated In-Vehicle from Non-Forward because gazes to the instrument cluster, center display, or mirrors can reflect legitimate supervision of the vehicle and automation status; collapsing them into a single class would conflate legitimate supervisory glances with attention directed away from driving. Note that our AOIs were defined by gaze direction toward spatial regions of the cabin rather than by the semantic target of a gaze. This scheme avoided visually indistinguishable regions and would allow higher-accuracy estimation. For analyses requiring a binary measure, both classes (non-forward and in-vehicle) can be combined as off-path; they may also be examined separately when the target AOI can be identified reliably. Ground-truth AOI labels were produced by three trained coders working from a shared annotation manual and a set of exemplar frames. A subset of 3000 frames was independently labeled by three coders, with an initial agreement of 96.57%; all disagreements were resolved through joint discussion before the labels were accepted.

As mentioned, the gaze module used a YOLOv8-cls classifier for compact training and deployment, calibrated per participant from a small set of labeled samples for participant-related inference. Specifically, SCRFD [4] first detected the face regions. Then, the gaze estimation models were trained and fine-tuned with Ultralytics version 8.4.9 [24], and initialized from the official ImageNet-pretrained `yolov8n-cls.pt` checkpoint. As required by the base models, our

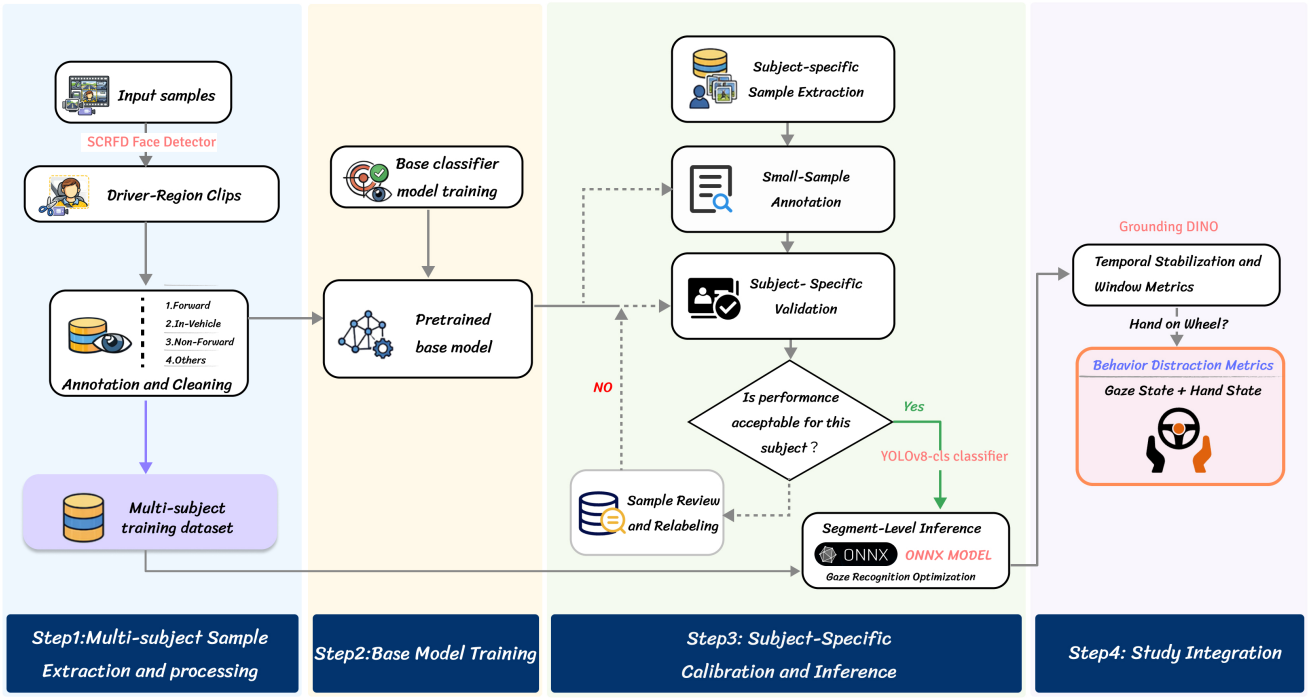


Figure 1: Overview of the annotation workflow. The gaze detection module included SCRFD face detection and YOLOv8-cls AOI classification, followed by subject-specific calibration with a manual performance review loop. The hand-on-wheel module used GroundingDINO to detect contact evidence, followed by temporal state stabilization. The output of both modules was used to generate window-level behavioral or eye-tracking metrics.

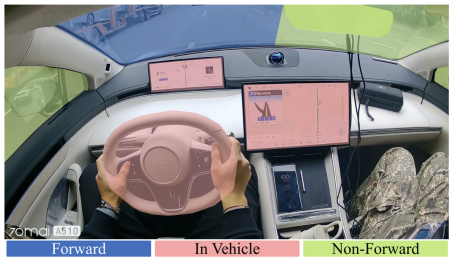


Figure 2: Illustration of the Areas of Interest (AOIs). Blue overlay denotes Forward, pink overlay denotes In-Vehicle, and green overlay denotes Non-Forward AOIs.

developed models and checkpoint metadata adopted the applicable license AGPL-3.0.

Although YOLOv8-cls is convenient for deployment because of its compact architecture and mature export pipeline, these engineering advantages do not imply superior predictive accuracy. We therefore compared YOLOv8s-cls, ResNet50 [5], EfficientNet-B0 [20], and a DeiT-Tiny vision transformer [23] under a common participant holdout protocol, testing whether the deployment choice holds an accuracy advantage. All models received 224 × 224-pixel inputs.

2.2 Hand-on-Wheel Module

In addition to the modules that operated on the driver's face/head region for gaze detection, our workflow also had a module targeting the steering-wheel region for hand position detection. Specifically, the hand branch estimated steering-wheel engagement as ON, OFF, or UNCERTAIN, derived from hand/wheel detections and spatial contact.

The module targeting hand location was based on GroundingDINO [13] to detect hand and steering wheel contacts via text prompts, and the contact evidence was computed based on bounding-box overlap at 5 Hz. A frame-level state machine mapped this evidence to states using an intersection-over-union (IoU) threshold of 0.05 for both ON and OFF. When a frame provided no reliable evidence, the previous state was retained for up to 0.8 s (four samples at the 5-Hz rate); if unreliable evidence persisted beyond this duration, the state became UNCERTAIN.

In addition, to improve the model efficiency and allow repeated large-scale processing, we trained a lightweight state-level classifier (Figure 3), with GroundingDINO as the teacher model to generate ON/OFF/UNCERTAIN states. We referred to this transfer as *hard-label state distillation*: unlike conventional knowledge distillation based on softened teacher logits [6], our implementation used only the teacher model's discrete state output labels and standard hard-label cross-entropy. Specifically, for crop x_i, the teacher pipeline supplied one discrete state $\hat{y}_i^T \in \{\text{ON}, \text{OFF}, \text{UNCERTAIN}\}$, and the

student was optimized with the standard hard-label cross-entropy objective

$$\mathcal{L}_{\text{hard}} = -\frac{1}{N}\sum_{i=1}^{N} \log p_s\left(\hat{y}_i^T \mid x_i\right), \tag{1}$$

where p_s was the student class probability. No teacher logits, temperature parameter, or soft-target loss was used. The hand-state outputs were reviewed by human coders at the final-state level.

2.3 Temporal Stabilization

Note that the modules above would only generate frame-level output, which will further be aggregated into event-level output. To ensure smooth stabilized output, we applied a temporal stabilization strategy to both the gaze and hand streams. For the gaze stream, the output switched to Others after eight consecutive frames marked as absent (i.e., no AOI was detected) and returned to AOI output after three consecutive detected frames. Further, to suppress label flicker, we required a new AOI label to persist for three consecutive frames before replacing the original AOI label. For the hand stream, frames without reliable contact evidence retained the previous state for up to 0.8 s before the state became UNCERTAIN.

3 Validation of the Workflow

3.1 Dataset

To validate the proposed workflow, a dataset was collected, which included video data from a naturalistic driving study conducted in South China (Guangzhou City and its surrounding areas). Data collection was carried out between February 2025 and March 2026, and 15 drivers participated in this study. All videos were recorded at a resolution of 1920 × 1080 pixels and a frame rate of 30 Hz. Ethical approval for this study was granted by the Human Research Ethics Committee of the Hong Kong University of Science and Technology (Guangzhou) (Protocol No. HKUST(GZ)-HSP-2024-0076), and written informed consent was provided by all participants. For analysis, the video dataset was divided into fixed, non-overlapping 20-s *windows*, which served as the downstream analysis unit rather than a model parameter. This window length followed the SHRP2-derived coding convention used by Risteska et al. [19], in which 20-s units served as the unit of analysis for the same gaze measures examined here. All 7,327 windows had complete gaze and hand-state outputs, and the final participant-level summaries included the 7,306 windows that passed a 0.98 gaze-coverage threshold (i.e., at least 98% of a window's frames carried a valid AOI label). The remaining 21 windows were excluded from the summaries due to low quality, e.g., face-detection failures, occlusion, or extreme lighting.

To facilitate driving behavior analyses, we calculated several metrics that are commonly used in the driving domain based on the stabilized frame-level outputs. First, gaze time percentages, a metric commonly used to quantify drivers' visual attention allocation [9, 22], were computed using stabilized AOI labels within each 20-s window, with samples labeled as Others excluded from denominators. Further, long off-path glances, which are highly related to crash risks [18, 19], were defined as consecutive off-path gazes (In-Vehicle or Non-Forward), longer than 1.6 s [8] and 2.0 s [14] thresholds.

Table 1: Mean held-out performance in the five-seed 15-participant LOPO gaze detection experiment. Each model has 75 held-out runs.

Model	Macro-F1	Bal. acc.	Window acc.
ResNet50	0.590	0.621	0.618
EfficientNet-B0	0.584	0.612	0.627
YOLOv8s-cls	0.560	0.599	0.604
DeiT-Tiny	0.308	0.347	0.422

Table 2: Mean macro-F1 for participant-specific adaptation by label budget (three participants, five seeds per budget).

Model	25	50	100	200
ResNet50	0.441	0.517	0.644	0.662
YOLOv8s-cls	0.326	0.425	0.633	0.712

3.2 Evaluation of Gaze Detection Module

3.2.1 Module Generalization and Adaptation. To evaluate cross-participant generalization of AOI classification, we ran a leave-one-participant-out (LOPO) experiment over 15 participants with five seeds and four backbones, using three-class macro-F1 over Forward, In-Vehicle, and Non-Forward as the primary metric. For each class, the F1 score is the harmonic mean of precision (the proportion of frames predicted as that class that truly belong to it) and recall (the proportion of frames of that class that were correctly identified); macro-F1 averages the three per-class F1 scores with equal weight. The label "Others" was excluded due to the small sample size. Two secondary metrics were also reported: balanced accuracy (Bal. acc.) and window accuracy (Window acc.). For the balanced accuracy, we averaged the per-class recall over the three classes with equal weight. For the window accuracy, the majority ground-truth label and the majority predicted label were computed per window, and a window was counted as correct when these two labels matched.

Table 1 reports mean held-out performance across the 75 runs per model. It can be observed that ResNet50 achieved the highest mean metrics, while YOLOv8s-cls remained useful for deployment but was not superior to other models. Across splits, the best non-YOLO mean exceeded YOLOv8s-cls on 11 of 15 participants.

We also examined how many manually labeled samples were needed to calibrate the model to a new participant. For three participants, we fine-tuned on 25, 50, 100, or 200 calibration images per participant—referred to as the label budget—while keeping each participant's test set fixed, and compared ResNet50 with YOLOv8s-cls using three-class macro-F1 (five seeds per budget). Table 2 shows that ResNet50 outperformed YOLOv8s-cls at 25–100 labels, whereas YOLOv8s-cls overtook ResNet50 at 200 labels. The ranking of backbones thus depended on the label budget, so backbone comparisons in future work should control and report the label budget.

3.2.2 Deployment Evaluation. Deployment evaluation assessed the engineering feasibility of the Gaze Detection Module. In our LOPO experiment, all trained AOI models were exported to ONNX with no top-1 parity (ONNX-vs-PyTorch top-1 agreement) below 0.99.

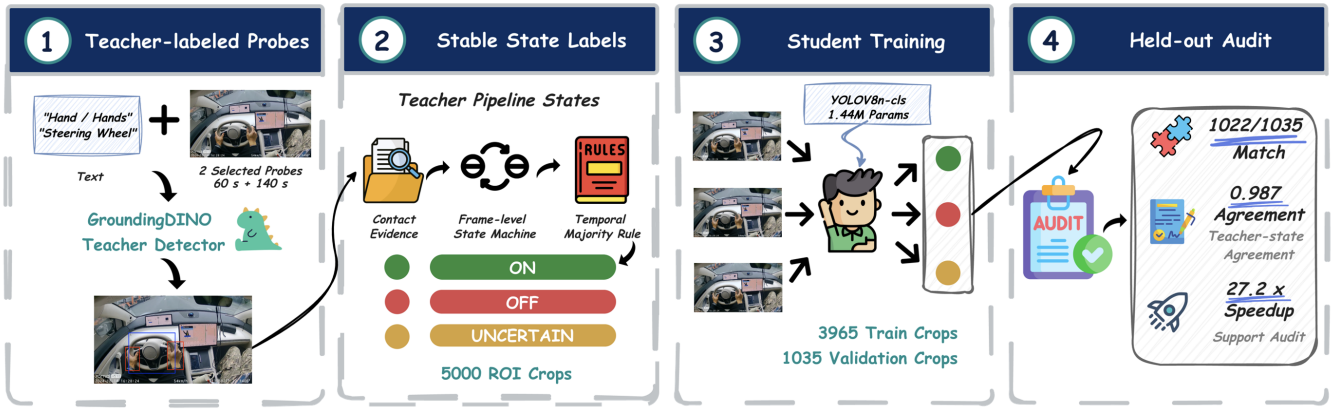


Figure 3: Hard-label state-distillation pipeline for hand-on-wheel. GroundingDINO served as the teacher model, and its discrete ON/OFF/UNCERTAIN states supervised a lightweight student model without soft targets. The held-out results were based on the dataset that will be introduced later.

Table 3: ONNX CPU deployment benchmark for the four AOI backbones

Model	MB	B1 p50 ms	B1 img/s	B32 img/s
YOLOv8s-cls	19.4	5.26	183.7	903.6
ResNet50	89.6	7.20	139.1	525.7
EfficientNet-B0	15.3	18.63	53.4	288.3
DeiT-Tiny	21.2	17.19	56.7	291.1

Table 3 reports a compact CPU benchmark for the four backbones. We found that YOLOv8s-cls was smaller and faster than the non-YOLO backbones, supporting YOLOv8-cls as a convenient but not uniquely accurate choice within the workflow.

3.2.3 Temporal Stabilization. To validate the effectiveness of temporal stabilization, we compared raw per-frame AOI predictions with stabilized AOI labels on two random participants, totaling 643,603 frames over 436.95 minutes. We found that stabilization reduced label switches from 279.82 to 42.22 per minute, an 84.91% decrease. Raw and stabilized labels diverged on 106,418 frames, accounting for 16.54% of the subset. The temporal stabilization also changed the resulting behavioral measures: the number of off-path episodes exceeding 1.6 s increased from 923 to 1,289, and those exceeding 2.0 s increased from 696 to 1,032. This demonstrates that temporal stabilization affects the measured behaviors beyond merely smoothing the frame-level predictions.

3.3 Evaluation of Hand-State Detection Module

The hand-state detection module assigned an ON, OFF, or UNCERTAIN state to 3,914,412 frames across all 15 participants. For several participants and camera placements, UNCERTAIN was the most frequent state, indicating sensitivity to camera layout, occlusion, and steering wheel visibility. Unlike the gaze detection, the hand state detection was not evaluated under a participant-holdout protocol due to large variations in camera angles and cabin layout.

Specifically, the student model was trained on class-balanced samples generated by the teacher pipeline, with 2,500 training and 300 held-out crops per hand state (7,500 training and 900 validation crops in total), drawn from multiple participants and camera layouts and from non-adjacent time blocks. Selected crops were required to have matching raw and stabilized states, to lie at least 0.5 s away from a state transition, and, for ON, to have reliable contact evidence. In the held-out test, the student model reached 0.969 precision, 0.930 recall, and 0.949 F1 for ON, 0.894 precision, 0.897 recall, and 0.895 F1 for OFF, and 0.910 precision, 0.943 recall, and 0.926 F1 for UNCERTAIN, with an overall agreement of 0.923 with the teacher. OFF was the weakest of the three states, suggesting that further review is needed for difficult cases when the hands are near the hand–wheel contact boundary.

As for the computational cost, the student processed 5,000 crops in 6.3 s versus 172.8 s for the GroundingDINO teacher, a 27.2× speedup. We therefore present the student as an acceleration checkpoint and an audit aid, not as a full replacement for the teacher module in application.

4 Discussion

In this study, we presented a human-in-the-loop workflow that converts raw in-cabin naturalistic driving video into two dimensions of driver behavior: visual attention allocation, captured as AOI-level gaze labels and long off-path glance events, and manual engagement, captured as hand-on-wheel state. Neither can be read directly from the video. A fully automated CV pipeline cannot handle this task reliably on its own due to differences in camera placement, occlusion, and lighting [15, 26]. Thus, our workflow kept a human in the loop and made each step transparent. The workflow therefore keeps region selection, gaze labeling, participant-specific calibration, temporal stabilization, and window-level metric extraction as separate, reviewable stages rather than collapsing them into a single end-to-end score, so that a measurement error can be localized to a stage instead of being absorbed into an aggregate metric.

The workflow also changes what a naturalistic video corpus can support. Traditional video reduction requires trained coders to label long recordings frame by frame, which is costly and scales poorly [10, 12], so annotation is typically confined to a sampled subset and much of the recorded material is not analyzed [16]. In our workflow, calibrating the model to each participant with a small set of labeled images made it feasible to run inference over millions of frames, shifting the human effort from exhaustive coding to small-sample calibration and staged review. Because each output remains traceable to the stage that produced it, a later behavioral analysis can trace the workflow and validate the outcome - this is vital for behavior research.

The temporal stabilization was also found to be valuable. This stage requires a gaze label to persist across several consecutive frames before it takes effect. We found this is more than removing frame-to-frame jitter; instead, it changed the resulting behavioral measures, for example, by substantially increasing the counts of long off-path glances (Section 3.2.3). This also showed the benefit of the traceable workflow: if such stabilization rules are hidden inside an end-to-end pipeline, an analyst cannot tell that the reported counts were shaped by them.

5 Summary

5.1 Limitations and Future Work

Several limitations should be noted. First, all evidence comes from 15 drivers recruited in a single region, so the transferability of our workflow to other driver populations, vehicles, cockpit layouts, cameras, and road contexts remains untested. Second, the workflow requires participant-specific calibration labels rather than zero-label generalization, and because cockpit-layout differences can dominate results, each new layout requires repeated human review. Third, our model comparison is conditioned on the number of calibration images: YOLOv8-cls is compact and simple to export but may not be the most accurate, and the performance of models varied between calibration sizes, so the evaluation of the models should be interpreted with caution. Fourth, the evidence on temporal stabilization comes from a two-participant subset and a single set of temporal parameters; further validation is needed on a larger sample size. Next, we followed previous studies to decide certain thresholds in the workflow (e.g. 0.98 as the gaze-coverage threshold). Future studies should conduct thorough sensitivity analyses to determine the optimal thresholds for metrics extraction. Finally, the hand-state detection module was not evaluated by holding out participants; its states remain sensitive to camera layout and steering-wheel visibility, UNCERTAIN dominated the frames of several participants, and the student model has so far been compared only against its teacher rather than against human labels. Thus, more validation of the hand-state detection model is needed.

5.2 Conclusions

We presented a human-in-the-loop workflow that turns naturalistic in-cabin video into two reviewable supervisory features, AOI gaze labels and hand-on-wheel state, through a small set of manual labels, participant-specific calibration, model-assisted inference, and staged human review. We provide preliminary evidence that model-assisted annotation can be a reliable and traceable tool for analyses of naturalistic driving data. Applied to millions of frames from 15 drivers, it made these features available at a scale that exhaustive manual coding cannot reach, while keeping each feature traceable to the stage that produced it. The workflow also showed that temporal stabilization that grouped per-frame output into behavioral events changed the outcome of the features at the window level, indicating that such stabilization should be reported alongside the features to ensure the transparency of the analyses.

Acknowledgments

This work was supported by the Nansha District Key Area S&T Scheme, Science and Technology Bureau of Nansha District (No. 2023ZD006).

References

[1] Oliver Carsten. 2016. UDRIVE: European naturalistic driving and riding for infrastructure and vehicle safety and environment. *Transportation Research Procedia* 14 (2016), 3149–3158. doi:10.1016/j.trpro.2016.05.256

[2] Lex Fridman, Daniel E Brown, Michael Glazer, William Angell, Spencer Dodd, Benedikt Jenik, Jack Terwilliger, Aleksandr Patsekin, Julia Kindelsberger, Li Ding, et al. 2019. MIT advanced vehicle technology study: Large-scale naturalistic driving study of driver behavior and interaction with automation. *IEEE Access* 7 (2019), 102021–102038.

[3] Lex Fridman, Philipp Langhans, Joonbum Lee, and Bryan Reimer. 2016. Driver gaze region estimation without use of eye movement. *IEEE Intelligent Systems* 31, 3 (2016), 49–56.

[4] Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. 2021. Sample and Computation Redistribution for Efficient Face Detection. arXiv:2105.04714 [cs.CV] https://arxiv.org/abs/2105.04714

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Piscataway, NJ, USA, 770–778. doi:10.1109/CVPR.2016.90

[6] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. arXiv:1503.02531 [stat.ML] https://arxiv.org/abs/1503.02531

[7] T Hoang Ngan Le, Yutong Zheng, Chenchen Zhu, Khoa Luu, and Marios Savvides. 2016. Multiple scale faster-rcnn approach to driver's cell-phone usage and hands on steering wheel detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. IEEE Computer Society, Las Vegas, NV, USA, 46–53.

[8] William J Horrey and Christopher D Wickens. 2007. In-vehicle glance duration: distributions, tails, and model of crash risk. *Transportation research record* 2018, 1 (2007), 22–28.

[9] International Organization for Standardization. 2020. *Road Vehicles — Measurement and Analysis of Driver Visual Behaviour with Respect to Transport Information and Control Systems*. International Standard ISO 15007:2020. International Organization for Standardization, Geneva, Switzerland. https://www.iso.org/standard/63220.html

[10] Reinier J Jansen, Sander T van der Kint, and Frouke Hermens. 2021. Does agreement mean accuracy? Evaluating glance annotation in naturalistic driving data. *Behavior research methods* 53, 1 (2021), 430–446.

[11] Sheila G. Klauer, Thomas A. Dingus, Vicki L. Neale, Jeremy D. Sudweeks, and David J. Ramsey. 2006. *The Impact of Driver Inattention on Near-Crash/Crash Risk: An Analysis Using the 100-Car Naturalistic Driving Study Data*. Technical Report DOT HS 810 594. National Highway Traffic Safety Administration, Washington, DC.

[12] Sheila G Klauer, Miguel Perez, and Julie McClafferty. 2011. Naturalistic driving studies and data coding and analysis techniques. In *Handbook of traffic psychology*. Elsevier, San Diego, CA, USA, 73–85.

[13] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2024. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. In *Computer Vision – ECCV 2024*. Springer Nature Switzerland, Cham, Switzerland, 38–55. doi:10.1007/978-3-031-72970-6_3 arXiv:2303.05499 [cs.CV]

[14] National Highway Traffic Safety Administration. 2013. *Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices*. Notice of Federal Guidelines, Docket No. NHTSA–2010–0053. U.S. Department of Transportation, Washington, DC. Federal Register, Vol. 78, No. 81, pp. 24817–24890.

[15] Juan Diego Ortega, Neslihan Kose, Paola Cañas, Min-An Chao, Alexander Unnervik, Marcos Nieto, Oihana Otaegui, and Luis Salgado. 2020. DMD: A Large-Scale Multi-Modal Driver Monitoring Dataset for Attention and Alertness

Analysis. In *Computer Vision – ECCV 2020 Workshops*. Springer International Publishing, Cham, Switzerland, 387–405. doi:10.1007/978-3-030-66823-5_23 arXiv:2008.12085 [cs.CV]

[16] Linda Pipkorn, T Zach Noonan, Bruce Mehler, Bryan Reimer, and Pnina Gershon. 2025. Changes in glance and hands-on-wheel behaviors when engaging with non-driving-related tasks while using Tesla Autopilot. *Transportation Research Part F: Traffic Psychology and Behaviour* 109 (2025), 232–241.

[17] Shrinivas Pundlik, Seonggyu Choe, Patrick Baker, Chen-Yuan Lee, Naser Al-Madi, Alex R Bowers, and Gang Luo. 2025. Data extraction and processing methods to aid the study of driving behaviors at intersections in naturalistic driving. arXiv:2505.05487 [cs.CV] https://arxiv.org/abs/2505.05487

[18] Bryan Reimer, Bruce Mehler, and Birsen Donmez. 2014. A study of young adults examining phone dialing while driving using a touchscreen vs. a button style flip-phone. *Transportation research part F: traffic psychology and behaviour* 23 (2014), 57–68.

[19] Martina Risteska, Dina Kanaan, Birsen Donmez, and Huei-Yen Winnie Chen. 2021. The effect of driving demands on distraction engagement and glance behaviors: Results from naturalistic data. *Safety science* 136 (2021), 105123.

[20] Mingxing Tan and Quoc V. Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*. PMLR, Long Beach, CA, USA, 6105–6114. arXiv:1905.11946 [cs.LG] https://arxiv.org/abs/1905.11946

[21] Ashish Tawari, Kuo Hao Chen, and Mohan M Trivedi. 2014. Where is the driver looking: Analysis of head, eye and iris for robust gaze zone estimation. In *17th International IEEE conference on intelligent transportation systems (ITSC)*. IEEE, Qingdao, China, 988–994.

[22] Emma Tivesten and Marco Dozza. 2015. Driving context influences drivers' decision to engage in visual–manual phone tasks: Evidence from a naturalistic driving study. *Journal of safety research* 53 (2015), 87–96.

[23] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training Data-Efficient Image Transformers and Distillation through Attention. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, Virtual Event, 10347–10357. arXiv:2012.12877 [cs.CV] https://arxiv.org/abs/2012.12877

[24] Ultralytics. 2026. Ultralytics YOLO. Computer software. https://github.com/ultralytics/ultralytics Version 8.4.9; AGPL-3.0 license.

[25] Trent Victor, Marco Dozza, Jonas Bärgman, Christian-Nils Boda, Johan Engström, Carol Flannagan, John D. Lee, and Gustav Markkula. 2015. *Analysis of Naturalistic Driving Study Data: Safer Glances, Driver Inattention, and Crash Risk*. Technical Report SHRP 2 Report S2-S08B-RW-1. Transportation Research Board, Washington, DC.

[26] Sourabh Vora, Akshay Rangesh, and Mohan Manubhai Trivedi. 2018. Driver Gaze Zone Estimation Using Convolutional Neural Networks: A General Framework and Ablative Analysis. *IEEE Transactions on Intelligent Vehicles* 3, 3 (2018), 254–265. doi:10.1109/TIV.2018.2843120 arXiv:1802.02690 [cs.CV]

[27] Jiyang Wang, Weiheng Chai, Archana Venkatachalapathy, Kai Liang Tan, Arya Haghighat, Senem Velipasalar, Yaw Adu-Gyamfi, and Anuj Sharma. 2021. A survey on driver behavior analysis from in-vehicle cameras. *IEEE Transactions on Intelligent Transportation Systems* 23, 8 (2021), 10186–10209.