EgoDrive: Egocentric Multimodal Driver Behavior Recognition using Project Aria

Michael Rice
University of Galway
School of Computer Science
Galway, Ireland
m.rice11@universityofgalway.ie

Lorenz Krause
University of Galway
School of Computer Science
Galway, Ireland

Waqar Shahid Qureshi
University of Galway
School of Computer Science
Galway, Ireland

Abstract

Egocentric sensing using wearable devices offers a unique first-person perspective for driver behavior analysis and monitoring, with the potential to accurately capture rich multimodal cues such as eye gaze, head motion, and hand activity directly from the driver's viewpoint. In this paper, we introduce a multimodal driver behavior recognition framework utilizing Meta's Project Aria smart glasses, along with a novel, synchronized egocentric driving dataset comprising high-resolution Red Green Blue (RGB) video, gaze-tracking data, Inertial Measurement Unit (IMU) signals, hand pose landmarks, and YOLO-based semantic object detections. All sensor data streams are temporally aligned and segmented into fixed-length clips, each manually annotated with one of six distinct driver behavior classes: *Driving*, *Left Mirror Check*, *Right Wing Mirror Check*, *Rear-view Mirror Check*, *Mobile Phone Usage*, and *Idle*. We design a Transformer-based recognition framework in which each modality is processed by a specialized encoder and then fused via Temporal Transformer layers to capture cross-modal temporal dependencies. To investigate the trade-off between accuracy and efficiency for real-time deployment, we introduce two model variants: EgoDriveMax, optimized for maximum accuracy, and EgoDriveRT, designed for real-time performance. These models achieve classification accuracies of 98.6% and 97.4% respectively. Notably, EgoDriveRT delivers strong performance despite operating with only 104K parameters and requiring just 2.65 ms per inference without the use of a specialized graphical processing unit—highlighting its potential for efficient, real-time in-cabin driver monitoring.

Figure 1: Project Aria Glasses. (Engel et al., 2023)

1 Introduction

Egocentric sensing offers powerful capabilities for capturing and interpreting human behavior in complex, real-world scenarios. In particular, the fusion of diverse sensor modalities can provide a rich, temporally aligned representation of user actions. However, integrating these heterogeneous data streams in a unified framework while ensuring real-time performance poses substantial technical challenges. Driver behavior analysis and action recognition provide a compelling and high-stakes application domain to explore and evaluate such systems.

In this work, we investigate the technical feasibility of such an approach to driver behavior recognition through a proof-of-concept system using Meta's Project Aria glasses (Engel et al., 2023). Our approach integrates high-resolution RGB video, eye gaze tracking, hand pose landmarks, IMU data, and semantic object detections to recognize six driver behaviors. We demonstrate

Proceedings of First International Workshop on Gaze Data and NLP associated with RANLP 2025,
pages 18–25, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-104-2-003

that effective multimodal fusion can be achieved while maintaining real-time performance, introducing two Transformer-based architectures that explore the accuracy-efficiency trade-off: EgoDriveMax and EgoDriveRT.

Our contributions are threefold: (1) we demonstrate the technical feasibility of real-time driver behaviour recognition using multimodal egocentric sensing, (2) we propose two efficient Transformer-based architectures that achieve high accuracy under strict latency and resource constraints, and (3) we introduce a proof-of-concept style, egocentric driving dataset comprising the aligned aforementioned data streams. Although our evaluation is conducted in a controlled setting with a singular participant and vehicle, the consistently strong performance across diverse driver actions indicates the potential for scalable deployment in real-world driver monitoring systems.

2 Related Work

Egocentric Vision. A rapidly growing area within computer vision, primarily driven by advances in wearable and augmented reality technologies. Meta are an established force in this domain, particularly in the open-source ecosystem, due to major contributions such as the Ego4D dataset (Grauman et al., 2021), the Project Aria initiative itself (Engel et al., 2023) and the HOT3D dataset (Banerjee et al., 2024), among others. Interest has also begun to permeate through into the automotive research space with implementations such as EgoFormer (Qazi et al., 2024), EgoSpeed-Net (Ding et al., 2022) and others paving the way for egocentric driver behavior modeling and in-cabin understanding.

Beyond Meta and the automotive sector, the academic egocentric vision landscape includes several influential datasets and methodologies. Epic-Kitchens-100 (Damen et al., 2020) provides fine-grained action recognition in kitchen environments, while EGTEA Gaze+ (Li et al., 2018) combines egocentric video with gaze data for activity understanding. Recent advances in egocentric representation learning include EgoVLM (Vinod et al., 2025) for vision-language understanding and EgoNCE (Lin et al., 2022) for self-supervised learning from temporal relationships.

Multimodal Learning. Effective multimodal learning requires architectures capable of aligning and fusing heterogeneous data streams with varying sampling rates and representational characteristics. Recent work has explored various fusion strategies, from early concatenation to attention-based approaches. Meta also possess a strong foothold in this research community, with implementations such as 'Reading in the Wild' (Yang et al., 2025) demonstrating transformer-based multimodal fusion using RGB, head pose, and eye-tracking data, for the recognition of the reading action in a variety of scenarios. Also created by Meta's researchers, Moon et al. (2023)'s IMU2CLIP work represents a significant advance in aligning IMU sensor data with textual representations through contrastive learning, displaying how motion sensors can be integrated into multimodal frameworks and providing a potential avenue for resource efficient human action recognition via motion-to-text conversion.

Many of the recent advancements in this area have been driven by either transformer-based architectures or contrastive learning focused approaches. Contrastive Language-Image Pretraining (CLIP) (Radford et al., 2021) and its variants demonstrate effective cross-modal alignment through contrastive objectives, while works like VLMo (Bao et al., 2022) explore 'Mixture-of-Modality-Experts' based approaches for vision-language tasks.

Automotive Action Recognition. Action Recognition implementations for automotive applications in the academic world have traditionally relied on exocentric cameras and/or single-modality approaches. Martin et al. (2019)'s Drive&Act dataset represents the most comprehensive effort in this space, utilizing multiple cameras types alongside pose estimation for robust driver behavior recognition in numerous lighting conditions from the third-person perspective. Furthermore, several other works such as those from Lin et al. (2021) and Li et al. (2024) explore alternative methodologies such ss RGB-D cameras and mmWave radars for driver-centric behavior identification.

Hoskeri (2023)'s proof-of-concept work comes closest to our approach, demonstrating the feasibility of using smart glasses with forward-facing cameras and IMU sensors for basic driver monitoring. Their controlled lab-based study achieved strong performance (93-99% F1) on limited steering and head movement patterns, establishing initial feasibility but leaving open questions about multimodal integration and real-world deployment.

However, the landscape of driver behavior recognition also includes a wide range of non-academic implementations. In the commercial sector, Tesla's

cabin-facing camera system and Seeing Machines' Driver Monitoring Systems (DMS) represent current industry standards, typically achieving 95%+ accuracy for basic attention detection but with limited behavioral granularity. Smart Eye's AI-powered systems demonstrate real-time gaze tracking capabilities, though primarily for attention monitoring rather than detailed action recognition.

Finally, the challenge of achieving real-time performance with such systems is an extremely pertinent one and has driven research into numerous efficiency focused architectures, with those from the commercial domain subject to much more stringent regulations than those from academia.

3 EgoDrive Dataset

To investigate the technical feasibility of multimodal egocentric driver behavior recognition, we developed a proof-of-concept style dataset. Our dataset design prioritizes technical requirements over scale and scope, with the resulting dataset potentially serving as a template for future multimodal egocentric behavioral analysis studies captured using Project Aria (Engel et al., 2023).

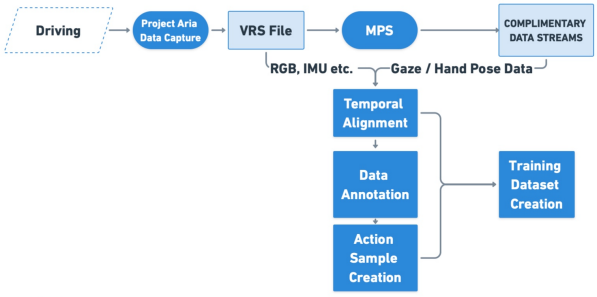

Figure 2: Dataset Creation Flowchart.

3.1 Data Gathering

As previously stated, all data was captured using Meta's Project Aria glasses as a single, integrated sensing platform. A major advantage of this approach is its inherent temporal synchronization across modalities, simplifying the reliable alignment of modality-specific timestamps by ensuring all data is referenced to a common device-time. We selected RGB camera (15fps), eye-tracking cameras (30fps), Simultaneous Localization and Mapping (SLAM) cameras (15fps), and IMUs (800Hz and 1KHz) based on their complementary roles in behavioral analysis: visual context, attention tracking, spatial awareness, and motion dynamics respectively. All recording adhered to General Data Protection Regulations (GDPR) , including informed consent where feasible, anonymization in post-processing, and secure data handling (GDPR, 2016).

A controlled, single-participant approach allows for the isolation of technical challenges in multimodal processing for a proof-of-concept based study, without confounding factors from inter-participant variability. Following the culmination of the data capture process, hand tracking and gaze estimates were obtained through Meta's Machine Perception Services (MPS).

3.2 Dataset Creation

Creating a temporally aligned multimodal dataset from asynchronous sensor streams poses several technical challenges, which our methodology had to overcome.

RGB timestamps are designated as the primary temporal reference. IMU data—comprising 6D input from both the accelerometer(3D) and gyroscope(3D) —is linearly interpolated to match the RGB frame rate, resulting in a standardized sample shape of *(Sequence Length, IMU Hz / RGB FPS, 6)*. Gaze data, sampled at twice the RGB rate, is temporally aligned using a simplified, mean nearest-neighbor matching strategy. This produces a single (x, y) pixel-coordinate gaze point per RGB frame and results in a sample shape of *(Sequence Length, 2)*. Hand landmark data, returned by Meta's Machine Perception Services (MPS), at the same sampling rate as the RGB stream, required no resampling. Each sample is thus represented with a shape of *(Sequence Length, 8)*, where each 8-dimensional vector corresponds to the x-y positions of the left and right wrists and palms.

This alignment pipeline maintains temporal coherence across modalities, while preserving each sensor's native sampling behavior. In addition, semantic context was incorporated through object detections generated by a custom-trained YOLOv11 model tailored for in-cabin environments (see Section 4 for training details). Each frame's detections are encoded as a fixed-length feature vector, with details limited to four key objects. Each of the four objects was represented using five dimensions: a binary presence indicator (0 or 1), the x and y coordinates of the top-left corner of the bounding box, followed by the bounding box's height and width.

Figure 3: Annotated Dataset Samples

Following on from the dataset's compilation, manual frame-by-frame annotation mapped frame indices to six behaviorally relevant classes: *Driving*, *Left Wing Mirror Check*, *Right Wing Mirror Check*, *Rear-view Mirror Check*, *Mobile Phone Usage*, and *Idle*. These classes were selected to represent distinct attention patterns and physical actions that create differentiable multimodal signatures. Each training sample spans **32** frames (2.13s), with longer actions segmented into multiple samples and shorter actions padded to maintain consistent temporal context. The final dataset, processed for training, consisted of 2,448 samples, with a real-world consistent bias towards the *'Driving'* action, with the exact class distribution visible in Figure 4.

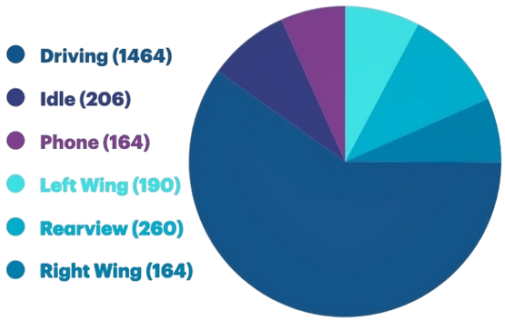

Figure 4: Dataset Class Distribution

4 Methodology

To address the challenge of fusing asynchronous and heterogeneous sensor streams, we adopt a modular processing pipeline centred around modality-specific encoders and transformer blocks. Full training methodologies for both the final models as well as the in-cabin object detection model are detailed below.

4.1 Object Detection Model

As previously stated, object detections for this implementation resulted from the training of a custom object detection model. Training frames were randomly sampled from the RGB streams of the main dataset and manually annotated with eight object classes: *Right Wing Mirror*, *Left Wing Mirror*, *Rearview Mirror*, *Gear Stick*, *Infotainment Unit*, *Speedometer*, *Steering Wheel*, and *Mobile Phone*. This process resulted in a dataset of over 4,000 annotated images. Once annotation was complete, the dataset was divided into an 80/10/10 train/validation/test split. A YOLOv11 backbone for fine-tuning was selected for its balance of efficiency and performance, achieving a precision of *96.5%* and a mAP50–95 of *88.1%* after training.

4.2 Model Architectures

We designed a Transformer-based architecture to address the core technical challenge of fusing heterogeneous sensor streams with different sampling rates and representational characteristics. Our approach processes each modality through specialized, unimodal encoders that extract meaningful features, which are then projected into a shared embedding space and passed through Temporal Transformer blocks for cross-modal reasoning.

Each sensor stream requires tailored processing to handle its unique characteristics. The RGB encoder processes visual sequences *(B, T, C, H, W)* using a pretrained *Swin-Tiny* Video Transformer (Liu et al., 2021) for spatial features, complemented by a *ResNet-18* (He et al., 2015) motion stream computing frame differences. Both streams are fused through projection networks and temporal 1D convolutions.

The gaze encoder projects normalized (0-1) x,y coordinates through linear layers, also followed by 1D convolutions, while the hand landmark encoder handles missing landmarks through learnable replacement vectors and validity masks, processing three parallel streams (coordinates, masks, missingness patterns) through feedforward networks and temporal attention.

Finally, object detection features undergo linear projection and 1D convolution for temporal modeling, while the IMU encoder processes signal through stacked 1D Convolutional Neural Networks (CNN) with pooling, followed by Gated Re-

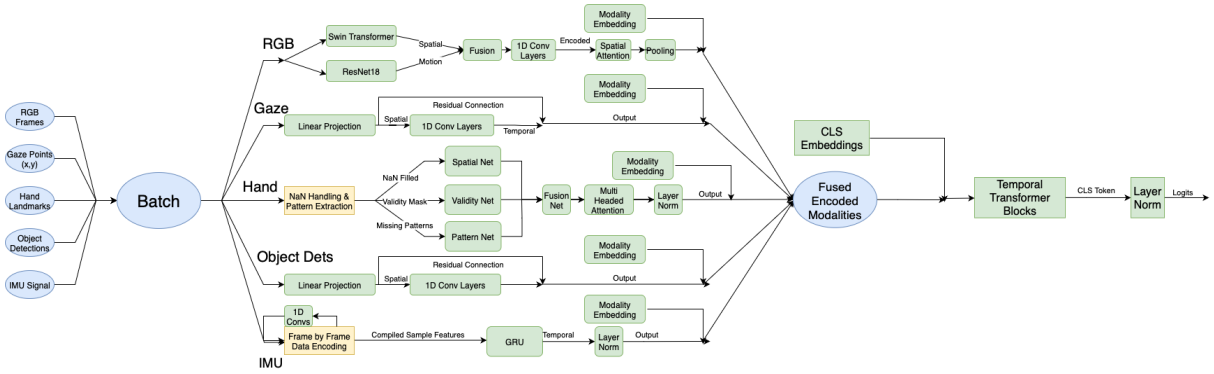

Figure 5: EgoDriveMax Architecture

current Unit (GRU) (Chung et al., 2014) layers for long-term modeling.

Projected features from all available modalities are concatenated and processed through one or more (stacked) Temporal Transformer blocks with multi-head attention, enabling the model to learn complex dependencies between behavioral cues across different sensor streams. A subsequent LayerNorm module is used to stabilize outputs.

4.3 Architectural Variants

To explore the accuracy-efficiency trade-off critical for real-time deployment, we developed two architectural variants that demonstrate different approaches to multimodal processing:

As shown in Table 1, the ***EgoDriveMax*** variant prioritizes absolute accuracy with 2 Transformer blocks, 4 attention heads, 256-dimensional features, and full RGB processing, totaling 42M parameters, while *EgoDriveRT* instead targets real-time performance with 1 block, 2 heads, 32-dimensional features, and RGB encoder removal, resulting in just 104K parameters - a 400x parameter reduction. Dropout was standardized across both models at a value of 0.1.

Model	Blocks	Heads	Feature Dim.	RGB
Max	2	4	256	✓
RT	1	2	32	✗

Table 1: Configuration details for the Max and RT model variants.

4.4 Training

All models were trained using a 60/20/20 train/validation/test split for a maximum of 20 epochs, with early stopping (patience = 5) to prevent overfitting. Optimization was performed using the Adam optimizer (Kingma and Ba, 2017) with a learning rate of 1×10^{-4}. Categorical Cross-Entropy loss was applied due to the multi-class nature of the task. To address class imbalance, loss weighting was used to increase the penalty for misclassifying underrepresented classes.

EgoDriveMax was trained on a single NVIDIA A100 GPU using Google Colab, while EgoDriveRT was trained locally on an Apple M4 chip. All training metrics and experiment logs were tracked using Weights & Biases (W&B).

Figure 6 shows the validation accuracy curve for EgoDriveRT, illustrating stable and smooth convergence. The EgoDriveMax model exhibited comparable convergence behavior during training.

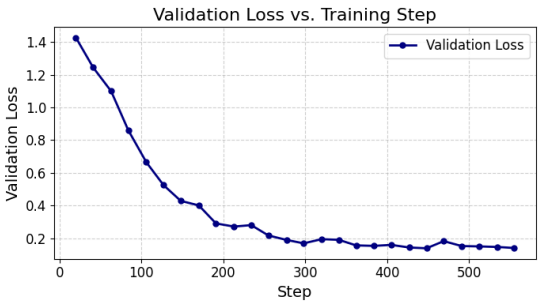

Figure 6: EgoDriveRT Validation Curve

5 Results

5.1 Proof-of-Concept Validation

We evaluate our approach to assess the technical feasibility of real-time multimodal driver behavior recognition. Our results demonstrate that effective multimodal fusion can achieve strong performance while maintaining practical inference constraints.

Table 2 shows our primary finding: the lightweight EgoDriveRT model achieves *97.4%*

accuracy with just *2.65ms* inference time on Apple's M4 chip using the Metal Performance Shaders framework, compared to EgoDriveMax's *98.6%* accuracy at *1595ms*.

Model	Acc	F1	Params	Inf Time
EgoDriveMax	**98.6%**	**98.0%**	42M	1595ms
EgoDriveRT	97.4%	96.6%	**104K**	**2.65ms**

Table 2: Model variant test results.

This *400x* parameter reduction (104K vs 42M) with minimal accuracy loss demonstrates that efficient multimodal architectures can capture essential behavioral patterns without requiring computationally expensive visual processing. Inference results displayed via annotated results from the EgoDriveMax model can be viewed below in Figure 7.

Figure 7: Example Action Detections

5.2 Per-Action Analysis

Table 3 displays the individual per-action results using both model variants, illustrating some interesting findings. The RT model's superior performance on Left Mirror Check (100% vs 96.9%) suggests that for certain actions, the simplified architecture may avoid overfitting to visual features while better leveraging complementary modalities like gaze and head motion.

The consistently strong performance across all actions using both models supports the effectiveness of the core technical approach for distinguishing behaviorally relevant driver actions. However, this performance may also be partially influenced by the controlled scope of the study; in broader, more diverse scenarios, a decline in performance would be a reasonable expectation.

Action	Acc	Prec	Rec	Model
Left Wing Mirror	96.9%	100%	96.9%	Max
Left Wing Mirror	**100%**	**100%**	**100%**	RT
Right Wing Mirror	**97.4%**	***100%***	**97.44%**	Max
Right Wing Mirror	94.9%	***100%***	94.9%	RT
Rearview Mirror	**97.9%**	97.9%	**97.9%**	Max
Rearview Mirror	91.2%	**100%**	91.2%	RT
Mobile Phone	94.1%	**94.1%**	94.1%	Max
Mobile Phone	**96.3%**	89.7%	**96.3%**	RT
Driving	**99.3%**	**98.7%**	**99.3%**	Max
Driving	98.7%	97.3%	98.7%	RT
Idle	**100%**	**100%**	**100%**	Max
Idle	97.1%	97.1%	97.1%	RT

Table 3: Model variant test results.

5.3 Ablation Study

To evaluate the contribution of each modality to overall performance, we conducted five additional training runs of the EgoDriveMax model, each time removing a singular modality. Due the complexity and robustness introduced by the multimodal setup, the model maintained strong performance across all ablations. Nonetheless, several meaningful trends emerged.

Modalities	Acc	Prec	Rec	F1
All	**98.6%**	**98.5%**	**97.6%**	**98.02%**
w/o Obj Dets	97.6%	97.4%	96.5%	96.9%
w/o Gaze	98.0%	97.6%	97.2%	97.4%
w/o RGB	97.4%	96.6%	97.3%	96.9%
w/o Hands	98.2%	98.0%	97.2%	97.6%
w/o IMU	97.4%	96.3%	97.3%	96.7%

Table 4: Ablation test results across different modality combinations.

As expected, the configuration using all available modalities achieved the highest scores across all evaluation metrics. Analyzing the F1 scores from the ablation runs, the IMU stream was found to be the most influential, providing the most discriminative features to the model. This was followed by object detections and RGB video frames, both of which contributed significantly. In contrast, the removal of gaze features and hand landmarks led to only minor drops in performance. This suggests that these modalities may be partially redundant, with their information content potentially approximated by other inputs—e.g., gaze direction could be inferred from a combination of object detection

bounding box locations and IMU-based motion patterns, reducing the utility of explicit gaze data.

6 Conclusions and Limitations

This work demonstrates the technical feasibility of real-time multimodal egocentric driver behavior recognition using wearable sensors. The most significant finding is that our lightweight *EgoDriveRT* model achieves near-optimal performance (97.4% accuracy) with 400x fewer parameters than the *EgoDriveMax* model and sub-3ms inference times. This efficiency suggests that the rich behavioral information captured through gaze tracking, hand pose, IMU data, and semantic object detection may be sufficient for accurate driver action recognition without computationally expensive visual processing.

Our single-participant controlled study validates the core technical approach, though inherently limits the generalizability of findings to broader populations and while the results certainly establish technical feasibility, real-world deployment would require validation across diverse drivers, vehicles, and environmental conditions, as well as an expanded action set, to ensure robust performance.

7 Future Work

Future research should prioritize multi-participant validation to capture inter-individual variability and explore on-device deployment strategies to preserve user privacy. The modular design also opens opportunities for personalization and continuous learning in long-term deployments. Furthermore, should future generations of the Project Aria device include onboard compute, this work presents the foundations for the development of a fully self-contained driver monitoring system. Finally, the architectural insights and dataset methodology presented here offer a strong foundation for building scalable, efficient, and context-aware egocentric driver monitoring systems.

References

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. 2024. Introducing HOT3d: An egocentric dataset for 3d hand and object tracking.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, and Furu Wei. 2022. VLMo: Unified vision-language pre-training with mixture-of-modality-experts.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2020. The EPIC-KITCHENS dataset: Collection, challenges and baselines.

Yichen Ding, Ziming Zhang, Yanhua Li, and Xun Zhou. 2022. EgoSpeed-net: forecasting speed-control in driver behavior from egocentric video data. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '22, pages 1–10. Association for Computing Machinery.

Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, Elias Mueggler, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eckenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulon, Prince Gupta, Raul Mur-Artal, Robbie Pennington, Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tassos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Balntas, Vijay Baiyya, Wilson Dreewes, Xiaqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. 2023. Project aria: A new tool for egocentric multi-modal AI research.

GDPR. 2016. Regulation - 2016/679 - EN - gdpr - EUR-lex.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar,

Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. 2021. Ego4d: Around the world in 3,000 hours of egocentric video.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Rahul Hoskeri. 2023. Poster abstract: Driving behavior monitoring with unobtrusive smart-glasses.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Guan-Hua Li, Hsin-Che Chiang, Yi-Chen Li, Shervin Shirmohammadi, and Cheng-Hsin Hsu. 2024. A driver activity dataset with multiple RGB-d cameras and mmWave radars. In *Proceedings of the 15th ACM Multimedia Systems Conference*, MMSys '24, pages 360–366. Association for Computing Machinery.

Yin Li, Miao Liu, and James M. Rehg. 2018. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Computer Vision – ECCV 2018*, volume 11207 of *Lecture Notes in Computer Science*, pages 639–655. Springer International Publishing.

Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, Chengfei Cai, Hongfa Wang, Dima Damen, Bernard Ghanem, Wei Liu, and Mike Zheng Shou. 2022. Egocentric video-language pretraining.

Zeyang Lin, Yinchuan Liu, and Xuetao Zhang. 2021. Driver-skeleton: A dataset for driver action recognition. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 1509–1514.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows.

Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon ReiB, Michael Voit, and Rainer Stiefelhagen. 2019. Drive&act: A multimodal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2801–2810. IEEE.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Aparajita Saraf, Amy Bearman, and Babak Damavandi. 2023. IMU2clip: Language-grounded motion sensor translation with multimodal contrastive learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13246–13253. Association for Computational Linguistics.

Tayeba Qazi, M. Rupesh Kumar, Prerana Mukherjee, and Brejesh Lall. 2024. EgoFormer: Ego-gesture classification in context of autonomous driving. 24(11):18133–18140.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Ashwin Vinod, Shrey Pandit, Aditya Vavre, and Linshen Liu. 2025. EgoVLM: Policy optimization for egocentric video understanding.

Charig Yang, Samiul Alam, Shakhrul Iman Siam, Michael J. Proulx, Lambert Mathias, Kiran Somasundaram, Luis Pesqueira, James Fort, Sheroze Sheriffdeen, Omkar Parkhi, Carl Ren, Mi Zhang, Yuning Chai, Richard Newcombe, and Hyo Jin Kim. 2025. Reading recognition in the wild.