Early Anticipation of Driving Maneuvers

Abdul Wasi[1], Shankar Gangisetty[1], Shyam Nandan Rai[2], and C.V. Jawahar[1]

[1] IIIT Hyderabad, India
wasilone11@gmail.com,{shankar.gangisetty@ihub-data.,jawahar@}iiit.ac.in
[2] Politecnico di Torino, Italy
shyam.rai@polito.it

Abstract. Prior works have addressed the problem of driver intention prediction (DIP) by identifying maneuvers after their onset. On the other hand, early anticipation is equally important in scenarios that demand a preemptive response before a maneuver begins. However, there is no prior work aimed at addressing the problem of driver action anticipation before the onset of the maneuver, limiting the ability of the advanced driver assistance system (ADAS) for early maneuver anticipation. In this work, we introduce Anticipating Driving Maneuvers (ADM), a new task that enables driver action anticipation before the onset of the maneuver. To initiate research in ADM task, we curate **D**riving **A**ction **A**nticipation **D**ataset, **DAAD**, that is *multi-view*: in- and out-cabin views in dense and heterogeneous scenarios, and *multimodal*: egocentric view and gaze information. The dataset captures sequences both before the initiation and during the execution of a maneuver. During dataset collection, we also ensure to capture wide diversity in traffic scenarios, weather and illumination, and driveway conditions. Next, we propose a strong baseline based on a transformer architecture to effectively model multiple views and modalities over longer video lengths. We benchmark the existing DIP methods on DAAD and related datasets. Finally, we perform an ablation study showing the effectiveness of multiple views and modalities in maneuver anticipation. Project Page: https://cvit.iiit.ac.in/research/projects/cvit-projects/daad.

Keywords: Action anticipation · Ego-centric vision · Gaze estimation · Multi-modal learning · Action recognition · Autonomous vehicles

1 Introduction

An ideal ADAS system should be capable of anticipating a potentially wrong maneuver moments before the driver intends to initiate it. That is, a system that actively learns from visual cues preceding the onset of a maneuver and alerts the driver beforehand. As an illustration, consider a scenario where a driver on a highway signals an intention to switch lanes. Despite activating the turn signal to indicate a move to the right lane, the driver, failing to check blind spots, remains unaware of a rapidly approaching vehicle in that lane. Consequently,

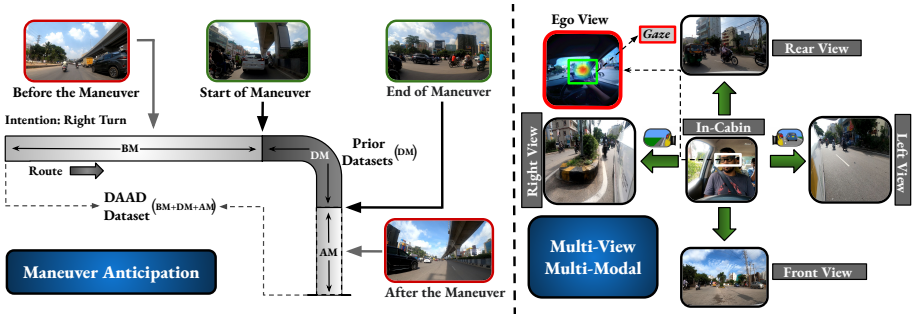

Fig. 1: Overview of DAAD dataset for ADM task. Left: Shows previous datasets containing maneuver videos from their initiation to their execution (DM), whereas our DAAD dataset features longer video sequences providing prior context (BM), which proves beneficial for early maneuver anticipation. **Right:** Illustrates the multi-view and multi-modality (Gaze through the egocentric view) in DAAD for ADM.

proceeding with the lane change without recognizing the potential collision risk could result in a hazardous situation. In this instance, relying solely on the turn signal did not encompass the surrounding traffic context, leading to the execution of the maneuver without due consideration of the potential dangers. In such cases, early anticipation enables the ADAS to prepare for the maneuver by adjusting the speed and checking for incoming traffic to avoid any safety hazards. The existence of such a system holds great significance in accident prevention, underscoring its crucial role in improving overall road safety. Existing ADAS systems detect critical maneuvers only after the driver has initiated them [14], leading to the onset of a hazardous situation. The brief time frame between identifying a potentially risky maneuver and the impending collision often proves insufficient for an effective response.

The closest attempt to address the challenge of anticipating the driver's maneuver (ADM) is found in Driver Intention Prediction (DIP) [14,21,22,24,30,37], which aims to predict the maneuver during its execution. Existing DIP methods attempt to predict the maneuver post its onset, essentially engaging in *video recognition* rather than *anticipation*[3]. This limitation arises from the nature of DIP [1, 21, 36, 50] datasets. Due to their relatively short length, these datasets provide minimal to no contextual information about the scene before the driver initiates the maneuver, constraining the potential for early anticipation as shown in Fig. 1. These datasets have maneuver clips with an average duration of less than 6 seconds, and at times, as brief as a second (see Fig. 3 for video duration statistics), encapsulating the maneuver from its initiation to execution (see DM in Fig. 1). Furthermore, their limited field of view constrains their capacity to comprehensively capture the surroundings of the vehicle, impeding effective inference. Table 1 enlists the in- and out-cabin views, modalities, and gaze information for the existing datasets.

[3] By "anticipate", we refer to the model's ability to predict a maneuver a few seconds before its actual execution.

On the other hand, most of the existing DIP methods rely on recurrent models like LSTMs that struggle to capture long-range temporal dependencies [16]. These models exhibit memory and computational limitations in processing videos that are typically longer than 5 seconds [47], limiting the scope for early anticipation. Moreover, since these methods are not compatible with multi-view[4] and multimodal data, they cannot model cues captured by different views that prove to be discriminative for early maneuver anticipation as shown in Fig. 1.

To address these shortcomings, we introduce the DAAD dataset comprising videos with sequences before the onset and during the execution of maneuvers. We refer to these sequences as *BM* (*Before the Maneuver*) and *DM* (*During the Maneuver*) as shown in Fig. 1. The existence of BM provides a scope for early maneuver anticipation. The dataset captures the egocentric view, side-mirror (i.e., blindspot) views, front and rare view, and the in-cabin view in dense and heterogeneous scenarios (see Fig. 3). Moreover, to address the shortcomings of DIP methods, we introduce a multi-view multi-modal vision transformer capable of modeling multi-view and cross-modal videos over longer durations for ADM.

In summary, our contributions are: (i) Introduce DAAD, a multi-view and multi-modal driving action anticipation dataset with longer video sequences, covering scenes before the onset of the maneuver that proves to be discriminative for early maneuver anticipation; (ii) Propose a multi-view multi-modal transformer with hybrid fusion and learnable memory that effectively utilizes the temporal information for maneuver anticipation task; (iii) Present quantitative benchmark results for the DAAD dataset across multiple baseline models, and showcase our method, demonstrating improved maneuver prediction performance.

2 Related Work

Action Anticipation. It is the task of predicting actions or movements beforehand. The field of action anticipation has seen significant progress, stimulated by the promising outcomes achieved in video recognition [10, 26, 47, 53]. It encompasses a diverse spectrum of problems in movement [25] and interaction [15, 16, 28] across short and long videos [32, 47]. Traditionally, recurrent networks [11, 12, 35] were used for action anticipation. However, with the advent of transformers [45] all the state-of-the-art action anticipation [16, 18, 51] are based on vision transformers [7]. Recently, the field of action anticipation has witnessed a drift in interest from third-person videos [13, 20, 22, 46] to first-person [5,6,12,16,31,39], and multi-modal videos [3,17,34,43,51]. However, there is no such dataset that addresses the issue of maneuver anticipation in driving.

Driver Intention Prediction. Numerous end-to-end deep learning architectures are introduced in the literature to tackle DIP challenge [14,21,22,24,30,37]. [14, 37] use a combination of 3D ResNets and LSTM. However, they perform poorly on longer video sequences [16, 47]. [30] proposed a vision transformer

[4] We use "multi-view" for more than two views. None of the aforementioned datasets other than AIDE [50] are multi-view. However, it has only 3 maneuver classes with 3 seconds long videos.

Table 1: Comparison of datasets. Our dataset is unique in containing longer multi-view multi-modal videos both in-cabin and out-cabin, along with eye gaze derived from Aria eye tracking cameras. Among others, DAAD is further diversified by capturing varying traffic densities, weather conditions, time of day, and the type of routes.

Dataset	Views		Size	Duration (hours)	Resolution	Areas	Drivers	Traffic Density	Multi-View	Multi-Modal	Eye Gaze	Day & Night	Weather Diversity	Anomalies
	In-cabin	Out-cabin												
Brain4Cars [21]	1	1	2M	10	N/A	Urban, Suburban	10	Low	✗	✗	✗	✗	✗	✗
VIENA^2 [1]	0	1	2.25M	20.83	1920x1280	N/A	15	N/A	✗	✗	✗	✓	✓	✗
HDD [36]	0	1	275K	**104**	1280x720	Urban, Suburban	N/A	Medium	✗	✗	✗	✗	✗	✗
LBW [23]	1	1	123K	7	N/A	Urban & Suburban	**28**	Low	✗	✓	✓	✗	✓	✗
AIDE [50]	1	3	561K	2.4	1920x1080	N/A	N/A	Medium	✓	✓	✗	✗	✓	✗
DAAD (Ours)	**2**	**4**	**6.6M**	85	1920x1080	Urban, Suburban, Rural	18	**High**	✓	✓	✓	✓	✓	✓

with learnable memory tokens [38] and a context-aware loss function. However, their encoder [7] struggles to leverage the temporal information and, instead, relies heavily on appearance. Our work explores a task effectively contributing towards both early anticipation and maneuver prediction.

Gaze from Ego-View. Gaze is known to provide strong indicators related to driver intent [27]. Earlier works in driving use gaze information for driver intention and attention anticipation. Many of these datasets are either captured in synthetic environments or lab settings [2, 48, 52]. Recent works [23, 33, 34, 49] show being captured in the real world, exploiting the ground-truth gaze information and videos from the front-facing camera, and drivers' face. Inspired by these, we incorporate egocentric gaze information in building a transformer for maneuver anticipation.

DIP Datasets. Brain4Cars [21] is a real-world dataset that introduces the DIP problem statement. It consists of 700 videos of lengths up to 6 seconds. VIENA^2 [1] is a large-scale synthetic DIP dataset consisting of 15,000 short videos. Another popular dataset is HDD [36], which introduces stimulus and cause in driving. Recently, the AIDE [50] dataset was introduced that facilitates contextual information from inside and outside the vehicle. All of these datasets have inherent limitations associated with their length of maneuver and have no context information of the scene before the onset of the maneuver. We address these issues in our proposed dataset that captures long video clips from multiple in- and out-cabin, and ego-centric view along with the gaze.

3 The DAAD Dataset

We introduce the DAAD dataset here. First, we outline the data capture setup and then present the annotation process for different driving intentions with causes and agents. Finally, we analyze the dataset attributes.

3.1 Data Capture and Collection

Data Capture Platform. To create the DAAD dataset, the cameras were arranged as shown in Fig. 2. The vehicle used for data capture encompassed

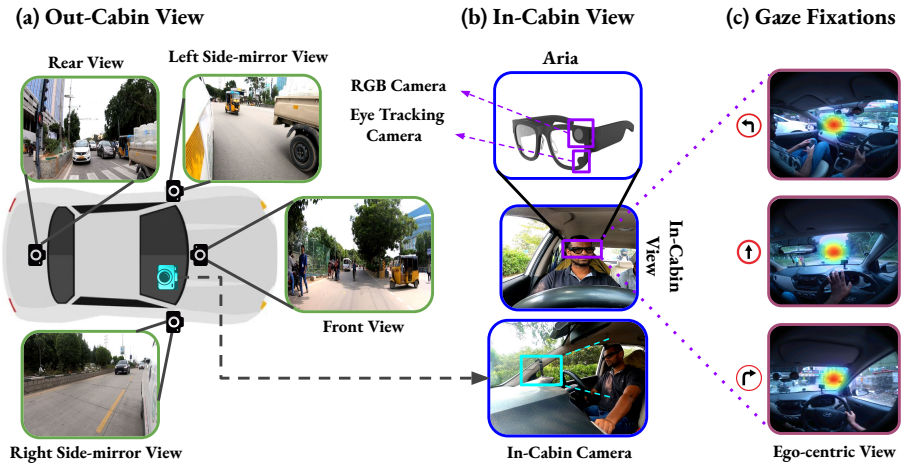

Fig. 2: Data capture setup. *(a) Out-cabin setup:* Four cameras are oriented towards frontal-view, side-mirror (left, right) views, and rear-view, *(b) In-cabin setup:* Driver-facing camera along with Aria [42] glasses for an egocentric view, *(c) Eye gaze derived from Aria eye tracking cameras:* Shown from egocentric view during a left turn, going straight, and a right turn respectively.

both the in-cabin and out-cabin views. DAAD comprises multiple data streams, synchronized and timestamped, and is captured at a rate of 30 frames per second.

Video Data. The vehicle was equipped with five monocular (GoPro 8) cameras with 1080p resolution. These cameras are positioned both outside the cabin (front, rear, left-side mirror, right-side mirror) and inside the cabin, specifically facing the driver. The out-cabin cameras offer a comprehensive view of the surrounding traffic context along with side blind spot regions that are essential for maneuver anticipation. The in-cabin camera non-intrusively captures the driver's actions during the maneuvers.

Driver Gaze. Gaze is captured using Aria [42] eye-tracking cameras. The device offers an eye-tracking resolution of 320×240, along with 8 MP video from the RBG camera with a resolution of 1408×1408.

Data Collection. The dataset was collected for 85 hours over 1400 kms on 24 different routes, spanning over 2 months. The diversity of the driving data is in terms of varying lighting conditions (morning to night), weather, drivers (18 participants, 15 male, 3 female), driver experience (from 5 to 35 years), road traffic density (from low to high), landscapes (main road, residential, highway, market, semi-urban, and rural), and the vehicles used. Furthermore, the dataset incorporates anomalies such as animals, potholes, and other objects that may pose interference during maneuvers. Table 1 compares DAAD against earlier driving datasets. Fig. 4 shows data samples in naturalistic unstructured driving conditions with respect to maneuvers, and Fig. 3 illustrates the diversity of our dataset.

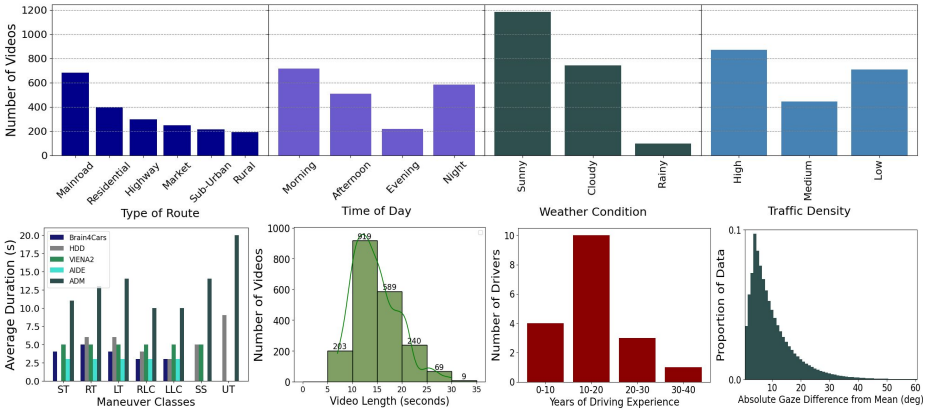

Fig. 3: Dataset statistics. **Top Row:** Number of videos for various types of routes, time of day, weather conditions, and traffic density. **Bottom Row:** Average maneuver duration for different datasets, average video lengths for our dataset, years of driving experience, and the absolute gaze difference from the mean.

3.2 Annotation Process

We annotate different driver intentions with appropriate causes. Each of the driving videos is annotated using one to four labels i.e. **intention**, **cause**, **static agent**, and **dynamic agent**. Fig. 5 shows annotated instances for each label.

The **intention** labels deal with the type of maneuver. Here, the labels used are *go straight* (ST), *right turn* (RT), *left turn* (LT), *right lane change* (RLC), *left lane change* (LLC), *slow/stop* (SS), and *U-turn* (UT). The **cause** labels are contingent on the current traffic context (existence of static and/or dynamic agent(s)). For example, a driver may be willing to take a right lane change (intention) to avoid congestion (cause) in a particular road lane, or, a driver may slow down or stop (intention) to let the other vehicle (dynamic agent) yield or cut-in (cause). Important to note is that not all of the videos in this dataset have static and dynamic agents in the scene. Therefore, the **cause** label does not need to be there in every case. Furthermore, a **dynamic agent** can be a vehicle, a pedestrian, or others (animal, bike, etc.), and a **static agent** can be a traffic light, a speed breaker, or a pothole among others. Pertinent to mention is that in this work, we stick to the usage of **intention** labels.

Sanity Check. We use the VIA video annotator [9], an open-source tool, to annotate the dataset. For data annotation, we used professional annotators with bespoke training, and an expert annotator doing quality checks to maintain consistency. For the first 100 videos, a group of five annotators individually provided annotations. Following this, an expert annotator assessed and rectified any discrepancies, ensuring accuracy and consistency in the annotations. Finally, all the videos were annotated, with each of the five annotators labeling a portion of the dataset and an expert annotator evaluated for correctness. With this process, less than 1% of the overall videos were incorrectly annotated and fixed by experts.

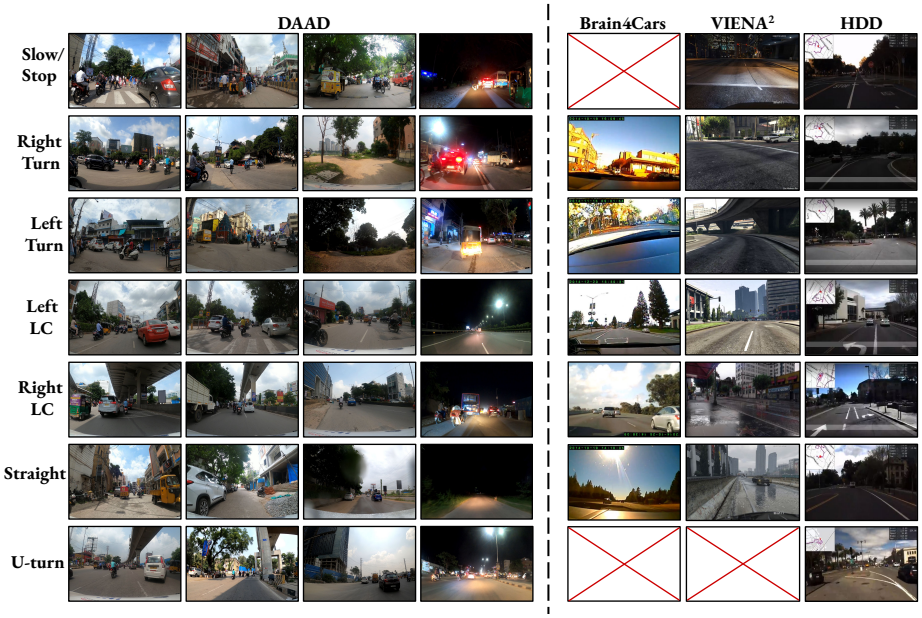

Fig. 4: Data samples. DAAD in comparison to Brain4Cars [21], VIENA^2 [1] and HDD [36] datasets. DAAD exhibits great diversity in various driving conditions (traffic density, day/night, weather, type of routes) across different driving maneuvers.

3.3 Data Statistics

Fig. 3 gives a detailed overview of the dataset statistics. The dataset has a total of 2,028 video samples, with a varying length of 5 to 35 seconds. Each sample consists of six video clips from five camera views and one Aria RGB ego-centric view. For each sample, specific label and gaze information is provided. Following the AIDE [50], a stratified sampling approach is applied and the dataset is split into training (65%), validation (15%), and testing (20%) sets. This division is performed without considering held-out subjects, acknowledging the inherent data imbalance stemming from the naturalistic nature of the dataset.

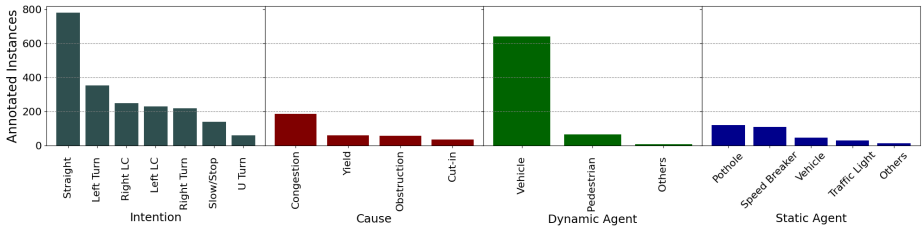

Fig. 5: Annotated instances. For driver intention (intended maneuver), cause behind a maneuver (if any), and the dynamic and static agents responsible for the cause.

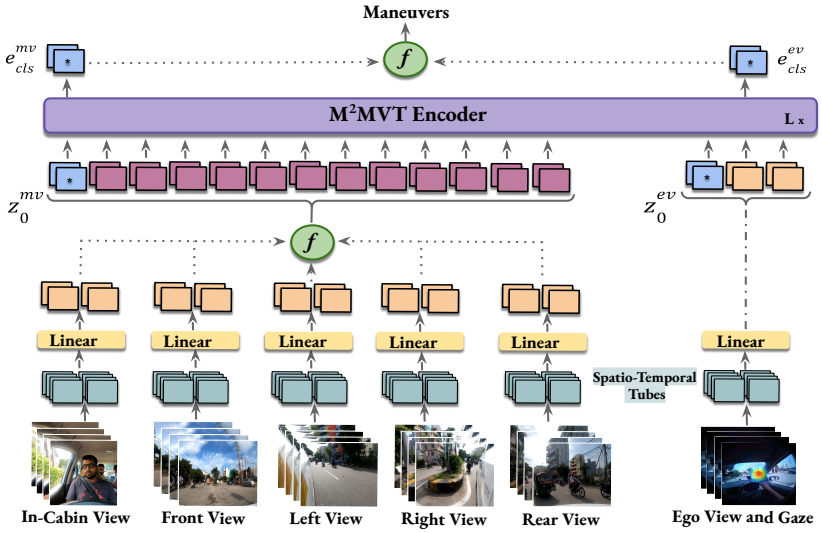

Fig. 6: ADM Framework. Separate projections are used for the in- and out-cabin views, post which the embeddings are fused (z^{mv}_n), and a classification token e^{mv}_{cls} is provided. These representations are then augmented with learnable memory embeddings, jointly given as z^{mv}_0. At the same time, the other modality (gaze from ego view) is fed separately to the encoder along with its classification token (e^{ev}_{cls}) and $E^{ev}_{eps} \in \mathbb{R}^{2\times d}$ learnable memory tokens (given jointly as z^{ev}_0). Finally, the classification tokens are concatenated (f) before anticipation.

Ethical Statement. Each participant was informed of the risks involved in data collection and signed a consent form reviewed by the Institutional Research Board (IRB) which allows the dataset to be publicly available for research purposes. All drivers were older than 18 years and held a valid driver's license. During driving, an instructor was on board in the passenger seat to provide safety instructions.

4 Anticipating Driving Maneuvers Framework

4.1 Our Approach

Fig. 6 gives an overview of the proposed framework. The network takes as input a set of multi-view and multi-modal videos that capture: (i) in-cabin driver's view, (ii) out-cabin traffic context from front-view, side-mirrors view, and rear-view, (iii) driving scene ego-centric view and gaze. These streams are processed independently to obtain two distinct embeddings: multi-view embedding z^{mv}_0 and ego-view embedding z^{ev}_0. Our goal is to fuse multi-modal multi-view long video representations for maneuver anticipation. To accomplish this, we introduce M^{2}MVT encoder augmented with episodic memory to retain prior context. Now, we present a detailed description of the modules involved.

Video Representations. We represent the input from the K in-cabin, out-cabin, and ego views as a 4D tensor $\mathbf{V} = \left\{\mathbf{V}_k \in \mathbb{R}^{T\times H\times W\times 3}\right\}_{k=1}^{K}$ where T, H, W and C are the temporal, spatial, and channel dimensions respectively.

M^{2}MVT Encoder. We split each input view into a set of non-overlapping spatio-temporal tubes (4D sub-tensors) $\mathbf{v}$ of dimensions $t \times h \times w \times c$ that are projected to get embeddings of size d. Here, $\mathbf{v}_k^{vit} := \left[v_k^1 X_k, ..., v_k^n X_k\right]$ for $k = 1, 2, ..., K$, and X_k is the projection matrix. Embeddings from multiple views (except for the ego view) are then fused into a combined sequence z_n^{mv}. After prepending a learnable class token to z_n^{mv}, it becomes $z_0^{mv} = \left[e_{cls}^{mv}, v_1^{vit}; ...; v_{K-1}^{vit}\right]$ where e_{cls}^{mv} is the joint classification token for the five views. Similarly, for the ego view, $z_0^{ev} = \left[e_{cls}^{ev}, v_K^{vit}\right]$. Based on learnings from [30] and [38], we prepend $N = 12$ **episodic memory tokens** $E_{eps} \in \mathbb{R}^{N\times d}$ to the input tokens, where d is the embedding dimension. Out of these, the multi-view input is augmented with $N-2$ tokens ($E_{eps}^{mv} \in \mathbb{R}^{(N-2)\times d}$), while for the ego-view, $E_{eps}^{ev} \in \mathbb{R}^{2\times d}$. Similar to [26], the input tensors are then pooled, post which the attention is computed on reduced sequence lengths. Encoders at each subsequent stage progressively down-sample the resolution. It is important to note that the fused episodic memory tokens undergo the same pooling operation, post which they are flattened again, however, with a reduced sequence length. We retain the decomposed relative position embedding [26].

Inspired by [44], we found out that early fusion of the embeddings of five views followed by a late fusion with the ego view improves the prediction performance of M^{2}MVT on our dataset. For early fusion, separate projections are used for all five views, post which their embeddings are projected into a shared representation and fused with the episodic memory tokens to learn a single joint `<CLS>` token e_{cls}^{mv}. For the ego view, a separate `<CLS>` token e_{cls}^{ev} is learned for the memory-augmented embeddings, which is then fused (concatenated) with the joint-classification token of the five views (e_{cls}^{mv}) before anticipation.

4.2 Implementation Details

M^{2}MVT is pre-trained on the Kinetics-600 [4] dataset. The videos are sampled at 30 FPS, with all the streams having a spatial resolution of 224×224. Before the input clips are projected into space-time patches, sampling over the temporal domain follows the [10] procedure. Specifically, we sample clips from the full-length video, and the input to the network is $T^{'}$ frames with a temporal stride of τ denoted as $T^{'} \times \tau$. We further evaluate it with two different frame samplings, 16×4 (sampling at a temporal stride of 4 from 16 frame input clips) and 32×3. The pooling operation (over the input and episodic memory embeddings) and the decomposed relative position embeddings are computed at spatio-temporal levels. The model was trained on 4 RTX 2080 GPUs over 80 epochs with a batch size of 4 clips. We retain the loss function of [30] and train our network using the AdamW optimizer [29], with a base learning rate of $1e^{-4}$.

5 Experiments

5.1 Baselines

We accommodate multiple views and multi-modality in the baseline methods for a fair comparison. Gebert *et. al* [14] take the in-cabin frames and pass them through a FlowNet [8], a 3D ResNet [19] and then through an LSTM for temporal modeling of the frames. Rong *et. al* [37] pass the out-cabin data stream through FlowNet before being passed through a ConvLSTM encoder [40] followed by a decoder. Parallely, in-cabin frames pass through a 3D ResNet-50 [19] and are fused with the output of the decoder for maneuver prediction. During baselining, we rely on the late fusion of all the views for these two methods. For [14], a multi-stream network [41] is used for a joint representation of all five views and the ego view similar to how they do it for in- and out-cabin views, followed by a late fusion. Pertinent to mention is that since our videos are relatively longer, we sampled 40 frames from each view for the experiment. In [37], a similar procedure is followed. However, all of the out-cabin views go through the same network as the out-cabin stream in the original work. In CEMFormer, Ma *et. al* [30] fuse the linear embeddings from the in-cabin and out-cabin patches to learnable memory tokens [38] from previous iterations to maintain a context of the past features.

During baselining, we follow the early fusion recipe wherein post-dividing the input views into patches, the embeddings are fused with episodic memory tokens, all of which is an input to the ViT encoder. For MViT and MViTv2, we take input from all six data streams and the memory tokens in a way similar to CEMFormer. However, instead of 2D patches, the input is now projected into space-time tubes. For all of these methods, we retain the pre-training, fine-tuning, and patchification strategies. For M^{2}MVT, we pass different modalities through the same encoder due to its strong performance on related tasks [17].

Table 2: Performance comparison of baseline methods. Accuracy and F_1 score (%) over different data sources on the Brain4Cars [21], HDD [36], VIENA^2 [1], AIDE [50], and our datasets. For DIP, In- and Out-Cabin-based M^{2}MVT gives the **best** performance on Brain4Cars and the lowest on our dataset.

Data Source	Method	Brain4Cars		AIDE		DAAD (Ours)	
		Acc.	F1	Acc.	F1	Acc.	F1
In-Cabin	Gebert *et. al* [14]	74.12 ± 0.52	70.70 ± 0.24	69.35 ± 0.3	67.88 ± 0.75	40.06 ± 0.05	42.70 ± 0.20
	Rong *et. al* [37]	77.24 ± 0.03	74.92 ± 0.02	69.11 ± 0.06	70.34 ± 0.25	41.10 ± 0.08	42.43 ± 0.34
	CEMFormer [30]	81.59 ± 0.90	80.49 ± 0.40	72.38 ± 0.29	71.59 ± 0.01	46.74 ± 0.43	44.18 ± 0.02
	M^{2}MVT	**81.87 ± 0.24**	**80.90 ± 0.03**	-	-	50.43 ± 0.04	48.11 ± 0.15
Out-Cabin	Gebert *et. al* [14]	72.89 ± 0.56	69.59 ± 0.04	72.89 ± 0.56	69.59 ± 0.04	52.65 ± 0.04	48.15 ± 0.04
	Rong *et. al* [37]	58.71 ± 0.04	62.75 ± 0.05	73.45 ± 0.03	70.17 ± 0.46	50.31 ± 0.04	54.05 ± 0.03
	CEMFormer [30]	63.27 ± 0.26	65.19 ± 0.21	75.90 ± 0.24	73.25 ± 0.15	58.87 ± 0.03	59.31 ± 0.05
	M^{2}MVT	**64.07 ± 0.02**	**65.35 ± 0.55**	-	-	58.78 ± 0.05	59.91 ± 0.35
In and Out-Cabin	Gebert *et. al* [14]	77.18 ± 0.24	78.20 ± 0.52	70.94 ± 0.35	71.86 ± 0.05	52.79 ± 0.50	52.36 ± 0.22
	Rong *et. al* [37]	81.87 ± 0.03	80.42 ± 0.10	73.89 ± 0.40	72.19 ± 0.04	53.47 ± 0.9	54.37 ± 0.22
	CEMFormer [30]	83.37 ± 0.03	82.73 ± 0.05	75.90 ± 0.24	73.25 ± 0.15	58.33 ± 0.74	61.68 ± 0.03
	M^{2}MVT	**83.18 ± 0.05**	**84.11 ± 0.35**	-	-	61.74 ± 0.20	63.82 ± 0.15

Data Source	Method	HDD		VIENA^2	
		Acc.	F1	Acc.	F1
Out-Cabin	Gebert *et. al* [14]	62.74 ± 0.00	64.43 ± 0.04	67.21 ± 0.76	66.39 ± 0.45
	Rong *et. al* [37]	63.89 ± 0.10	63.77 ± 0.33	71.92 ± 0.26	70.21 ± 0.01
	CEMFormer [30]	68.40 ± 0.03	66.16 ± 0.35	73.52 ± 0.35	72.95 ± 0.25
	M^{2}MVT	**72.51 ± 0.03**	**71.89 ± 0.40**	**75.63 ± 0.24**	**73.47 ± 0.01**

Table 3: Evaluation of M^{2}MVT on varying (a) Traffic Density, (b) Weather Condition, (c) Time of Day, and (d) Type of routes. Lowest in red.

Traffic Density	Acc. (↑)	F1 (↑)
Low	68.21	69.38
Medium	65.91	66.47
High	63.14	63.89

(a)

Weather Condition	Acc. (↑)	F1 (↑)
Sunny	64.92	65.28
Cloudy	68.81	69.47
Rainy	59.18	60.79

(b)

Time of Day	Acc. (↑)	F1 (↑)
Morning	69.63	69.82
Afternoon	68.76	69.04
Evening	69.12	70.77
Night	56.50	54.20

(c)

Type of Route	Acc. (↑)	F1 (↑)
Mainroad	65.15	66.56
Residential	64.39	66.13
Highway	72.14	71.41
Market	62.50	63.26
Sub-Urban	70.09	70.50
Rural	57.59	55.48

(d)

5.2 DAAD Benchmarking and Analysis

As shown in Table 2, we report the comparison results of different baseline models over standard driving datasets and our dataset across three data sources. The following are some key observations: (i) We observe that CEMFormer and M^{2}MVT, which use episodic memory achieve higher accuracy for in-cabin, and in- and out-cabin views on the Brain4Cars dataset, and for out-cabin views on the AIDE dataset. Unlike Brain4Cars, which uses the frontal view for learning traffic context, AIDE uses a multi-view out-cabin setup, highlighting the importance of multi-view in the DIP task. The higher out-cabin accuracy on DAAD validates our inference. (ii) We observe that there is a significant drop in the performance score of baselines on our dataset compared to other standard datasets. It can be attributed to two reasons. Firstly, DAAD consists of longer videos (see Fig. 1), making it challenging for the existing methods to perform on them [47]. Secondly, a considerable portion (Fig. 3) of our dataset consists of scenarios captured in dense and heterogeneous scenarios (Table 3), adding to its complexities. (iii) The ViT-based CEMFormer performs better than the LSTM-based models. It does so by the incorporation of episodic memory embeddings and early fusion of in- and out-cabin views. Moreover, the context-consistency loss function [30] augments its performance.

Effect of time-to-maneuver. Fig. 7a compares the accuracy of M^{2}MVT over time for different DIP datasets. For the DAAD dataset (DAAD-Full), we demonstrate that by having accuracy above the random chance (approximately 15%) before the culmination of *BM* (28.74%), the model can correctly anticipate the maneuver (ADM) before its onset. For DAAD-BM, we observe that when M^{2}MVT is trained on sequences before the onset of maneuver (BM) and tested on the whole video, it is still able to predict the upcoming maneuver above the random chance (27.76%). This supports our assertion that these sequences ($BM{:}t-4$ *to* t) contain indicative cues about the potential type of maneuver. For DAAD-DM, which is essentially a DIP task, we get an accuracy of 69.81%, suggesting that DIP is convoluted in dense and heterogeneous scenarios. The prediction accuracy over time for all maneuver classes in the comparative datasets is provided in the supplementary material.

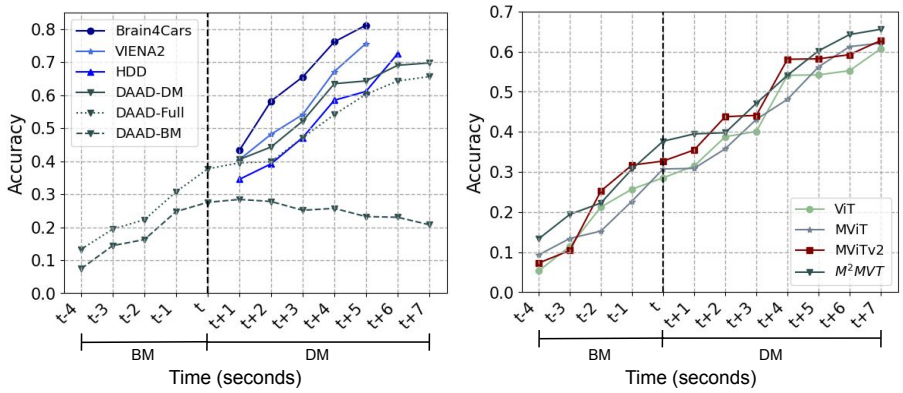

Fig. 7: Effect of time-to-maneuver. (a) Accuracy over time for different driving datasets on M^{2}MVT (with MViTv2 encoder). We conducted three separate experiments for DAAD dataset. (i) *DAAD-DM:* Training and testing only on maneuver sequences (DM). (ii) *DAAD-Full:* Training and testing on whole video. (iii) *DAAD-BM:* Training on a portion of the video captured before the onset of maneuver (BM) and testing on the whole video; (b) Accuracy over time for our dataset on ViT, MViT, MViTv2 encoders, and the proposed method (M^{2}MVT). Here, t is the time of onset of maneuver.

Importance of Dense, Diverse, and Heterogeneous Attributes. Table 3 compares the Accuracy and F_1 score over diverse scenarios. The following observations are made: (i) The performance dips in high-traffic scenarios, owing to the erratic motion of other traffic agents influencing our motion trajectory. (ii) For weather conditions, accuracy in sunny conditions is less than cloudy due to occasional glare. Also, due to rain, the cameras and window panes show decreased visibility, affecting both out-cabin and ego-view. (iii) Decreased visibility during nighttime contributes to a drop in accuracy compared to other times of the day. (iv) While driving in rural settings, sub-optimal visual cues are learned due to the presence of unstructured roads and surroundings.

5.3 Comparison to Proposed Approach

As shown in Table 4, our approach outperforms the ViT-based CEMFormer by 4.85% accuracy and 3.63% F1 score on all views and gaze with 61.22% less parameters. It can be attributed to several reasons. Firstly, the hybrid fusion network augments the performance by efficiently learning the cross-modal interactions between the multiple camera views and gaze. Its early fusion of multiple views, for which the episodic memory tokens jointly model the view, further improves learning. Secondly, the presence of implicit temporal bias in the encoder. M^{2}MVT exhibits strong modeling of temporal information, a phenomenon unfound in traditional ViTs that predominantly rely on appearance. Lastly, the spatio-temporal patchification and pooling along with learnable memory tokens, and the decomposed relative positional embeddings, further account for an improved performance on maneuver anticipation. M^{2}MVT, 16×4 gives an accuracy

Table 4: Encoder comparison of proposed framework accuracy and F_1 score (%) on various transformer encoders like ViT [30], MViT [10], MViTv2 [26] and ours (32×3 variants for all) over different data sources. For the task of anticipation, all views and gaze-based M^{2}MVT gives the **best** performance.

Encoder	Metric	Data Source						
		Aria RGB	Aria Gaze	In-Cabin & Aria RGB	In-Cabin & Gaze	Out-Cabin & Aria RGB	Out-Cabin & Gaze	All Views & Gaze
ViT [30]	**Acc.(↑)**	37.17	42.02	47.78	51.33	53.80	59.12	60.74
	F1 (↑)	38.88	40.10	49.50	51.01	54.92	60.95	63.09
	Param (M) (↓)	87.30	87.30	88.10	88.10	90.50	90.50	91.30
MViT [10]	**Acc.(↑)**	37.90	44.54	48.99	52.64	53.59	60.89	62.13
	F1 (↑)	38.47	41.29	50.08	52.83	54.11	59.63	63.65
	Param (M) (↓)	36.80	36.80	37.60	37.60	40.00	40.00	**40.80**
MViTv2 [26]	**Acc.(↑)**	39.17	45.22	51.48	53.11	54.28	61.94	62.78
	F1 (↑)	39.89	46.04	50.74	52.25	55.49	60.85	64.08
	Param (M) (↓)	51.40	51.40	52.20	52.20	54.60	54.60	55.20
M^{2}MVT (Ours)	**Acc.(↑)**	39.17	45.22	53.19	53.84	57.44	62.87	**65.59**
	F1 (↑)	39.89	46.04	52.77	54.02	58.50	63.33	**66.72**
	Param (M) (↓)	51.40	51.40	52.70	52.70	55.10	55.10	55.90

of 64.19% on all views and gaze. Further experimental details on the 16×4 variant can be found in the supplementary material.

Effect of time-to-maneuver. In Fig. 7b, M^{2}MVT encoder gives an accuracy of 37.66% at the onset of maneuver (t=0), which is 9.12% more than the ViT and 4.95% more than the MViTv2 at the same time, demonstrating a significant performance in early maneuver anticipation.

5.4 Ablation Study

Analysis of Confusion Matrices. In Fig. 8, we observe that DAAD achieves a lower performance compared to other datasets, highlighting that the task of ADM is challenging on longer videos in dense and heterogeneous environments. It specifically confuses between *U turn* and *Right turn*, and *go straight* and *slow/stop*. For the former, the dataset was captured in a left-hand side driving country, with *U turns* always taken towards the *right*. Furthermore, the presence of roads with no lane marks leads to wrongly classifying *lane changes* as *go straight* or sometimes, as a *turn*.

Effectiveness of Multiple Views and Modalities. Table 5a gives a comparison of accuracy and F1 score from different views and highlights the importance of multi-view and gaze in DIP. Using M^{2}MVT, we get an accuracy of 65.59% on all views with gaze. However, with the removal of gaze, the accuracy plummets by 2.41%, highlighting the importance of multi-modality in maneuver anticipation. Next, the accuracy drops by more than 5% without the front view and by less than 2% without the driver facing in-cabin view. This result contrasts with the accuracy of Brain4Cars on [37], [14] and [30], where the in-cabin accuracy is more than the out-cabin (front-facing). It can be attributed to the fact that Brain4Cars has short videos of maneuvers in which the driver shows explicit head movements in the direction of the maneuver. This, however, is not necessarily true for longer videos where the driver exhibits complex movements to evaluate the scene context before executing a maneuver. The existence of sequences

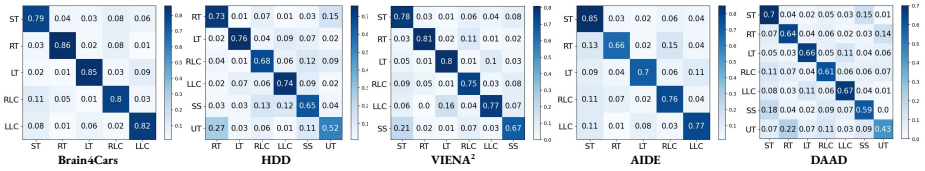

Fig. 8: Confusion matrices for Brain4Cars, HDD, VIENA^2, AIDE and DAAD on the proposed M^{2}MVT method.

Table 5: Ablation study. (a) Experiments on multiple views and modalities for our method on DAAD. (b) Accuracy and F_1 score (%) for varying episodic memory tokens on M^{2}MVT method.

Front	Rear	In-Cabin	Left	Right	Gaze	Acc. (↑)	F1 (↑)
✗	✓	✓	✓	✓	✓	60.53	56.92
✓	✗	✓	✓	✓	✓	62.85	56.66
✓	✓	✗	✓	✓	✓	63.82	59.31
✓	✓	✓	✗	✓	✓	63.02	62.74
✓	✓	✓	✓	✗	✓	61.99	59.74
✓	✓	✓	✓	✓	✗	63.28	65.31
✓	✓	✓	✓	✓	✓	**65.59**	**66.72**

(a)

N	Acc. (↑)	F1 (↑)
0	63.46	63.97
4	63.79	64.07
8	64.92	65.04
12	**65.59**	**66.72**
16	65.35	66.24

(b)

captured in dense and haphazard traffic and on roads with blind curves further convolutes this modeling.

Influence of Learnable Memory Tokens. Table 5b examines the impact of varying the number of learnable memory tokens on the performance of M^{2}MVT. [38] show that by using more than 5 memory tokens, the model shows no considerable improvement. Rather, in some cases [30], the accuracy plummets by using more tokens. We, however, note that in M^{2}MVT, the accuracy increases up to $N = 12$ tokens, post which it decreases. It suggests that with the increase in the number of views, the episodic memory tokens need to be increased as well.

6 Conclusion

In this paper, we present DAAD, a multi-view and multi-modal dataset to aid the next-generation ADAS to improve road safety by early maneuver anticipation. Next, we propose M^{2}MVT that acts as a strong baseline for DAAD dataset. Our proposed model is a hybrid-fusion multiscale vision transformer with learnable memory embeddings that efficiently models cross-modal spatiotemporal interactions. We provide an extensive experimentation to demonstrate the importance of multi-view and multi-modal data streams across diverse scenarios for maneuver prediction on the DAAD. We hope that our new proposed task of ADM and DAAD will pave the way for development of robust road safety systems.

7 Acknowledgements

This work is supported by iHub-Data and Mobility at IIIT Hyderabad and Project Aria from Meta.

References

1. Aliakbarian, M.S., Saleh, F.S., Salzmann, M., Fernando, B., Petersson, L., Andersson, L.: Viena2: A driving anticipation dataset (2018)
2. Amadori, P.V., Fischer, T., Wang, R., Demiris, Y.: Decision anticipation for driving assistance systems. In: ITSC. pp. 1–7. IEEE (2020)
3. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: ICCV. pp. 609–617 (2017)
4. Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., Zisserman, A.: A short note about kinetics-600. CoRR (2018)
5. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: ECCV. pp. 720–736 (2018)
6. Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. IJCV pp. 1–23 (2022)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
8. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: ICCV. pp. 2758–2766 (2015)
9. Dutta, A., Zisserman, A.: The VIA annotation software for images, audio and video. In: ACM Multimedia. pp. 2276–2279 (2019)
10. Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C.: Multiscale vision transformers. In: ICCV. pp. 6824–6835 (2021)
11. Furnari, A., Farinella, G.M.: What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In: ICCV. pp. 6252–6261 (2019)
12. Furnari, A., Farinella, G.M.: Rolling-unrolling lstms for action anticipation from first-person video. IEEE TPAMI **43**(11), 4021–4036 (2020)
13. Gao, J., Yang, Z., Nevatia, R.: RED: reinforced encoder-decoder networks for action anticipation. In: BMVC. BMVA Press (2017)
14. Gebert, P., Roitberg, A., Haurilet, M., Stiefelhagen, R.: End-to-end prediction of driver intention using 3d convolutional neural networks. In: IEEE Intelligent vehicles symposium (IV). pp. 969–974 (2019)
15. Girase, H., Agarwal, N., Choi, C., Mangalam, K.: Latency matters: Real-time action forecasting transformer. In: CVPR. pp. 18759–18769 (2023)
16. Girdhar, R., Grauman, K.: Anticipative video transformer. In: ICCV. pp. 13505–13515 (2021)
17. Girdhar, R., Singh, M., Ravi, N., van der Maaten, L., Joulin, A., Misra, I.: Omnivore: A single model for many visual modalities. In: CVPR. pp. 16102–16112 (2022)
18. Gong, D., Lee, J., Kim, M., Ha, S.J., Cho, M.: Future transformer for long-term action anticipation. In: CVPR. pp. 3052–3061 (2022)
19. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: CVPR. pp. 6546–6555 (2018)
20. Huang, D.A., Kitani, K.M.: Action-reaction: Forecasting the dynamics of human interaction. In: ECCV. pp. 489–504. Springer (2014)

21. Jain, A., Koppula, H.S., Soh, S., Raghavan, B., Saxena, A.: Car that knows before you do: Anticipating maneuvers via learning temporal driving models. ICCV (2015)
22. Jain, A., Singh, A., Koppula, H.S., Soh, S., Saxena, A.: Recurrent neural networks for driver activity anticipation via sensory-fusion architecture. In: ICRA. pp. 3118–3125. IEEE (2016)
23. Kasahara, I., Stent, S., Park, H.S.: Look both ways: Self-supervising driver gaze estimation and road scene saliency. In: ECCV. pp. 126–142. Springer (2022)
24. Khairdoost, N., Shirpour, M., Bauer, M.A., Beauchemin, S.S.: Real-time driver maneuver prediction using lstm. IEEE Transactions on Intelligent Vehicles **5**(4), 714–724 (2020)
25. Koppula, H.S., Saxena, A.: Anticipating human activities using object affordances for reactive robotic response. IEEE TPAMI pp. 14–29 (2015)
26. Li, Y., Wu, C.Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., Feichtenhofer, C.: Mvitv2: Improved multiscale vision transformers for classification and detection. In: CVPR. pp. 4804–4814 (2022)
27. Liu, C., Chen, Y., Tai, L., Ye, H., Liu, M., Shi, B.E.: A gaze model improves autonomous driving. In: ACM symposium on eye tracking research & applications. pp. 1–5 (2019)
28. Liu, M., Tang, S., Li, Y., Rehg, J.M.: Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In: ECCV. pp. 704–721. Springer (2020)
29. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (Poster) (2019)
30. Ma, Y., Ye, W., Cao, X., Abdelraouf, A., Han, K., Gupta, R., Wang, Z.: Cemformer: Learning to predict driver intentions from in-cabin and external cameras via spatial-temporal transformers. In: ITSC. pp. 4960–4966. IEEE (2023)
31. Nagarajan, T., Li, Y., Feichtenhofer, C., Grauman, K.: Ego-topo: Environment affordances from egocentric video. In: CVPR. pp. 163–172 (2020)
32. Nawhal, M., Jyothi, A.A., Mori, G.: Rethinking learning approaches for long-term action anticipation. In: ECCV. pp. 558–576 (2022)
33. Pal, A., Mondal, S., Christensen, H.I.: Looking at the right stuff-guided semantic-gaze for autonomous driving. In: CVPR. pp. 11883–11892 (2020)
34. Palazzi, A., Abati, D., Solera, F., Cucchiara, R., et al.: Predicting the driver's focus of attention: the dr (eye) ve project. IEEE TPAMI **41**(7), 1720–1733 (2018)
35. Pang, B., Zha, K., Cao, H., Shi, C., Lu, C.: Deep rnn framework for visual sequential applications. In: CVPR. pp. 423–432 (2019)
36. Ramanishka, V., Chen, Y.T., Misu, T., Saenko, K.: Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In: CVPR (2018)
37. Rong, Y., Akata, Z., Kasneci, E.: Driver intention anticipation based on in-cabin and driving scene monitoring. In: IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). pp. 1–8 (2020)
38. Sandler, M., Zhmoginov, A., Vladymyrov, M., Jackson, A.: Fine-tuning image transformers using learnable memory. In: CVPR. pp. 12155–12164 (2022)
39. Sener, F., Singhania, D., Yao, A.: Temporal aggregate representations for long-range video understanding. In: ECCV. pp. 154–171. Springer (2020)
40. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. NeurIPS **28** (2015)
41. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. NeurIPS **27** (2014)

42. Somasundaram, K., Dong, J., Tang, H., Straub, J., Yan, M., Goesele, M., Engel, J.J., Nardi, R.D., Newcombe, R.A.: Project aria: A new tool for egocentric multi-modal AI research. CoRR (2023)
43. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: Videobert: A joint model for video and language representation learning. In: ICCV. pp. 7464–7473 (2019)
44. Tziafas, G., Kasaei, H.: Early or late fusion matters: Efficient rgb-d fusion in vision transformers for 3d object recognition. In: IROS. pp. 9558–9565. IEEE (2023)
45. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017)
46. Vondrick, C., Pirsiavash, H., Torralba, A.: Anticipating visual representations from unlabeled video. In: CVPR. pp. 98–106 (2016)
47. Wu, C.Y., Li, Y., Mangalam, K., Fan, H., Xiong, B., Malik, J., Feichtenhofer, C.: Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In: CVPR. pp. 13587–13597 (2022)
48. Wu, M., Louw, T., Lahijanian, M., Ruan, W., Huang, X., Merat, N., Kwiatkowska, M.: Gaze-based intention anticipation over driving manoeuvres in semi-autonomous vehicles. In: IROS. pp. 6210–6216. IEEE (2019)
49. Xia, Y., Zhang, D., Kim, J., Nakayama, K., Zipser, K., Whitney, D.: Predicting driver attention in critical situations. In: ACCV. pp. 658–674. Springer (2019)
50. Yang, D., Huang, S., Xu, Z., Li, Z., Wang, S., Li, M., Wang, Y., Liu, Y., Yang, K., Chen, Z., et al.: Aide: A vision-driven multi-view, multi-modal, multi-tasking dataset for assistive driving perception. In: ICCV. pp. 20459–20470 (2023)
51. Zhong, Z., Schneider, D., Voit, M., Stiefelhagen, R., Beyerer, J.: Anticipative feature fusion transformer for multi-modal action anticipation. In: CVPR. pp. 6068–6077 (2023)
52. Zhou, F., Yang, X.J., De Winter, J.C.: Using eye-tracking data to predict situation awareness in real time during takeover transitions in conditionally automated driving. IEEE TITS **23**(3), 2284–2295 (2021)
53. Zhu, X., Xiong, Y., Dai, J., Yuan, L., Wei, Y.: Deep feature flow for video recognition. In: CVPR. pp. 2349–2358 (2017)